Comput Stat
DOI 10.1007/s00180-011-0276-y

ORIGINAL PAPER

Generalized canonical correlation analysis with missing values

Michel van de Velden · Yoshio Takane

Received: 10 November 2010 / Accepted: 21 July 2011

Abstract Generalized canonical correlation analysis is a versatile technique that allows the joint analysis of several sets of data matrices. The generalized canonical correlation analysis solution can be obtained through an eigenequation and distributional assumptions are not required. When dealing with multiple set data, the situation frequently occurs that some values are missing. In this paper, two new methods for dealing with missing values in generalized canonical correlation analysis are introduced. The first approach, which does not require iterations, is a generalization of the Test Equating method available for principal component analysis. In the second approach, missing values are imputed in such a way that the generalized canonical correlation analysis objective function does not increase in subsequent steps. Convergence is achieved when the value of the objective function remains constant. By means of a simulation study, we assess the performance of the new methods. We compare the results with those of two available methods; the missing-data passive method, introduced in Gifi's homogeneity analysis framework, and the GENCOM algorithm developed by Green and Carroll. An application using world bank data is used to illustrate the proposed methods.

Keywords Generalized canonical correlation analysis · Missing values

M. van de Velden (✉)
Econometric Institute, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands
e-mail: vandevelden@ese.eur.nl

Y. Takane
Department of Psychology, McGill University, 1205 Dr. Penfield Ave.,
Montreal, QC H3A 1B1, Canada
e-mail: takane@psych.mcgill.ca

Published online: 02 August 2011

1 Introduction

In canonical correlation analysis (Hotelling 1936) linear combinations of two sets of variables are obtained in such a way that the correlation between the linear combinations is a maximum. Generalizations to a similar approach for more sets of variables have been the topic of several studies (Horst 1961; Carroll 1968; Kettenring 1971). Consequently, several different approaches have been proposed. Kettenring (1971) provides an overview of four different generalizations. In the framework of homogeneity analysis Van der Burg (1988) and Gifi (1990) introduced nonlinear canonical correlation analysis, also referred to by the algorithm name OVERALS, which takes Carroll (1968) generalized canonical correlation analysis as a special case. Here, we will also use the generalization proposed by Carroll (1968). An excellent description of his method, in a similar notation as we employ in this paper, can be found in Steenkamp et al. (1994) who consider the method for marketing applications. An important advantage of Carroll's approach is its computational ease and the fact that the method takes ordinary canonical correlation analysis as a special case.

In generalized canonical correlation analysis several sets of variables are analyzed simultaneously. This makes the method suited for the analysis of various types of data. For example, in marketing research, subjects may be asked to rate a set of objects on a set of attributes. For each individual, a data matrix can then be constructed where the objects are represented row-wise and the attributes column-wise. Then, using generalized canonical correlation analysis a graphical representation, sometimes referred to as a perceptual map, can be made on the basis of the individuals' observation matrices. Note that, the observation matrices do not necessarily contain the same attributes. Steenkamp et al. (1994) focussed on this flexibility in their analysis of idiosyncratic sets of attributes.

Another type of application, considered by Green and Carroll (1988), concerns the derivation of a composite configuration from a set of configurations. For example, multidimensional scaling solutions (perceptual maps) for the same objects from different countries can be used as input data. Generalized canonical correlation analysis can then be applied to the coordinate matrices to obtain a composite configuration. Finally, generalized canonical correlation analysis can be used when, for the same set of subjects, we have data on sets of variables. For example, in their analysis of socio-economic determinants of HIV pandemic and nations efficiencies, Zanakis et al. (2007) used a set of 50 explanatory variables which could be divided into different sets (e.g. economic indicators, education related variables, etc.). For such multiple set data, generalized canonical correlation analysis can be used to obtain a configuration depicting the cases.

Since generalized canonical correlation analysis deals with possibly large sets of data, the possibility of the occurrence of missing values is significant. Some procedures to deal with missings in generalized canonical correlation analysis have been proposed, however, no attempt has been made to compare and evaluate the alternatives. In this paper, we review two existing procedures and propose two alternative methods that are conjectured to offer important advantages over the existing methods. We shall only concern ourselves with methods specifically aimed at dealing with missing values in generalized canonical correlation analysis. General methods

(e.g. multiple imputation, Rubin 1987) that require distributional assumptions, are beyond the scope of this paper. The performance of the proposed methods under various conditions will be assessed by means of a simulation study. The results of this simulation study clearly indicate the validity and, in some cases, superiority of the new methods.

The paper is organized as follows. In the next section, we briefly introduce generalized canonical correlation analysis. In Sect. 3, we consider methods that do not require data imputation and rely on non-iterative solutions. Two methods involving data imputation are described in Sect. 4. A simulation study in which all methods are compared is presented in Sect. 5 and followed by an application of the approaches to World Bank data in Sect. 6. We conclude the paper with a brief summary of our results.

2 Generalized canonical correlation analysis

In generalized canonical correlation analysis linear combinations are obtained in such a way that the sum of squared correlations of the linear combinations of the variables with a so-called group configuration is a maximum. Let $\mathbf{Y}$ denote the unknown group configuration. The order of $\mathbf{Y}$ is $m \times k$, where m is the number of rows for each observation matrix $\mathbf{X}_i$ (i.e. the ith data set) and k is the dimensionality of the solution. The data matrices $\mathbf{X}_i$ are first centered. Sometimes, if the variables are for example measured on different scales, they are also standardized. Note that the sizes of the observation matrices $\mathbf{X}_i$ are $m \times p_i$ (with $p_i \leq m - 1$)for $i = 1, \ldots, n$. The dimensionality of the solution, k, must be chosen by the researcher.

We can formulate as objective

$$\min \phi\left(\mathbf{Y}, \mathbf{A}_i\right) = \min \operatorname{trace} \sum_{i=1}^{n} \left(\mathbf{Y} - \mathbf{X}_i \mathbf{A}_i\right)' \left(\mathbf{Y} - \mathbf{X}_i \mathbf{A}_i\right) \tag{1}$$

subject to the restriction

$$\mathbf{Y}'\mathbf{Y} = \mathbf{I}_k. \tag{2}$$

It is known, e.g. Carroll (1968), that for observed $\mathbf{X}_i$ matrices, the group configuration $\mathbf{Y}$ can be obtained from the eigenequation

$$\left(\sum_{i=1}^{n} \mathbf{X}_i \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} \mathbf{X}_i'\right)\mathbf{Y} = \mathbf{Y}\boldsymbol{\Lambda}, \tag{3}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with as elements the k largest eigenvalues of $\sum_{i=1}^{n} \mathbf{X}_i \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} \mathbf{X}_i'$ (where we have assumed that the $\mathbf{X}_i's$ are of full column rank) and the matrices $\mathbf{A}_i$ can be calculated as

$$\mathbf{A}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} \mathbf{X}_i'\mathbf{Y}. \tag{4}$$

An interesting feature of the method is the fact that the sets of variables $\mathbf{X}_i$ may contain different variables. Hence, the number of variables in each set does not need to

be the same. Steenkamp et al. (1994) used this freedom to analyze object evaluations where each individual used there own set of attributes to evaluate objects.

3 Noniterative methods for dealing with missing values in generalized canonical correlation analysis

Van der Burg (1988) and Gifi (1990) suggested a method for dealing with missing values in nonlinear generalized canonical correlation analysis in which selection matrices are used to discard complete rows containing at least one missing value. Hence, if one element is missing in a row, the complete row is discarded. This method is often applied in the homogeneity analysis framework as set forth in Gifi (1990). It is referred to as the missing-data-passive approach to missing values. Van de Velden and Bijmolt (2006) used an equivalent method for the case where complete rows were missing. An advantage of the missing-data-passive approach is its computational ease. The solution can be obtained directly by means of an eigenequation. However, discarding complete rows if only one value is missing, clearly implies a considerable loss of information. Moreover, as the data are centered with respect to the fully observed rows, bias may be introduced. In particular, when values are not missing completely at random, this bias may be severe. To account for this problem, we propose an alternative approach in which a constant term is estimated separately. The proposed new method is a generalization of a method first proposed by Shibayama (1995).

For the sake of completeness and to facilitate an easy way of comparing the methods, we first summarize the missing-data-passive approach.

3.1 The missing-data-passive approach

In the missing-data-passive approach proposed in the context of nonlinear canonical correlation analysis, rows of the data matrices are removed if they contain one or more missing elements. The generalized canonical correlation approach is then applied by only using the observed rows. This method can easily be implemented by introducing a so-called selection matrix. Let $\mathbf{K}_i$ denote a diagonal matrix with its diagonals either ones or zeros. The ones correspond to rows for which there are no missings in the ith observation matrix and the zeros correspond to rows of $\mathbf{X}_i$, for which at least one value is missing. Obviously, the resulting selection matrices $\mathbf{K}_i$ are symmetric idempotent, that is, $\mathbf{K}_i = \mathbf{K}_i' = \mathbf{K}_i\mathbf{K}_i$. In the missing-data-passive approach, the data are first centered with respected to the fully observed rows. This centering can be achieved by defining:

$$\mathbf{Q}_i = \mathbf{I} - \left(\mathbf{1}'\mathbf{K}_i\mathbf{1}\right)^{-1}\mathbf{1}\mathbf{1}'\mathbf{K}_i. \tag{5}$$

Inserting the centering and selection matrices into Eq. (1), we get

$$\min \phi\left(\mathbf{Y}, \mathbf{A}_i\right) = \min \operatorname{trace} \sum_{i=1}^{n} \left(\mathbf{Y} - \mathbf{Q}_i\mathbf{X}_i\mathbf{A}_i\right)' \mathbf{K}_i \left(\mathbf{Y} - \mathbf{Q}_i\mathbf{X}_i\mathbf{A}_i\right) \tag{6}$$

which we minimize subject to the restriction

$$\mathbf{Y}'\mathbf{K}\mathbf{Y} = \mathbf{I}_k, \tag{7}$$

where

$$\mathbf{K} = \sum \mathbf{K}_i.$$

It is not difficult to see that the resulting group configuration can be obtained from the eigenequation:

$$\mathbf{K}^{-1/2}\left(\sum_{i=1}^{n} \mathbf{K}_i\mathbf{Q}_i\mathbf{X}_i\left(\mathbf{X}_i'\mathbf{Q}_i'\mathbf{K}_i\mathbf{Q}_i\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{Q}_i'\mathbf{K}_i\right)\mathbf{K}^{-1/2}\mathbf{Y}_s = \mathbf{Y}_s\mathbf{\Lambda},$$

where $\mathbf{\Lambda}$ is the diagonal matrix with as elements the k largest eigenvalues and we have assumed that the $\mathbf{X}_i's$ are of full column rank, and $\mathbf{Y}_s$ is an $m \times k$ matrix of corresponding orthonormal eigenvectors. Hence, the appropriately standardized group configuration can be obtained as

$$\mathbf{Y} = \mathbf{K}^{-\frac{1}{2}}\mathbf{Y}_s.$$

The matrices $\mathbf{A}_i$ can be calculated as

$$\mathbf{A}_i = \left(\mathbf{X}_i'\mathbf{Q}_i'\mathbf{K}_i\mathbf{Q}_i\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{Q}_i'\mathbf{K}_i\mathbf{Y}.$$

3.2 The Test Equating method

The Test Equating method, was proposed by Shibayama (1995) in a one-dimensional setting. However, Takane (1995) showed that the method could easily be extended to a k-dimensional solution similar to principal component analysis. Moreover, Takane and Oshima-Takane (2003) showed that the Test Equating method is closely related to the missing-data-passive approach in homogeneity analysis (e.g. Meulman 1982; Gifi 1990). The difference between the two methods lies in the estimation of a mean term in the Test Equating method. Here, we further generalize the Test Equating method to the generalized canonical correlation analysis case. This new method is conjectured to outperform the missing-data-passive method especially when the missing elements are related to the values. That is, if missingness is related to the values (i.e. high values are more likely to be missings), the test equating method should yield better results than the missing-passive-approach that assumes the missings to occur completely at random.

To apply the Test Equating method in generalized canonical correlation analysis, we must employ row-wise deletion similar as was the case in the missing-data-passive approach described in Sect. 3.1. Hence, if a row contains at least one missing value, the complete row will be removed. However, instead of the centering step employed in the missing-data-passive approach, the Test Equating method requires the estimation of a constant term. Thus, in the Test Equating method, the group configuration is approximated by a constant term plus k linear combinations of the columns of $\mathbf{X}_i$. We can formulate this as follows:

$$\min \phi\left(\mathbf{Y}, \mathbf{A}_i, \mathbf{a}_{i0}\right) = \min \operatorname{trace} \sum_{i=1}^{n} \left(\mathbf{Y} - \mathbf{X}_i\mathbf{A}_i - \mathbf{1}\mathbf{a}_{i0}'\right)' \mathbf{K}_i \left(\mathbf{Y} - \mathbf{X}_i\mathbf{A}_i - \mathbf{1}\mathbf{a}_{i0}'\right). \tag{8}$$

We can solve this minimization problem sequentially. First, differentiation with respect to $\mathbf{a}_{i0}$ yields as first order condition,

$$\mathbf{a}_{i0} = \left(\mathbf{1}'\mathbf{K}_i\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{K}_i\left(\mathbf{Y} - \mathbf{X}_i\mathbf{A}_i\right).$$

Subsituting this into (8), yields, after some manipulations,

$$\min \phi\left(\mathbf{Y}, \mathbf{A}_i\right) = \min \operatorname{trace} \sum_{i=1}^{n} \left(\mathbf{Y} - \mathbf{X}_i\mathbf{A}_i\right)' \mathbf{Q}_i'\mathbf{K}_i\mathbf{Q}_i \left(\mathbf{Y} - \mathbf{X}_i\mathbf{A}_i\right), \tag{9}$$

where $\mathbf{Q}_i$ is as defined in (5). Let

$$\mathbf{P}_i = \mathbf{Q}_i'\mathbf{K}_i\mathbf{Q}_i,$$

it is easily verified that $\mathbf{P}_i$ is symmetric idempotent, i.e. $\mathbf{P}_i = \mathbf{P}_i' = \mathbf{P}_i\mathbf{P}_i$. Solving (9) subject to the constraint

$$\mathbf{Y}'\mathbf{P}\mathbf{Y} = \mathbf{I} \tag{10}$$

where

$$\mathbf{P} = \sum_{i=1}^{n} \mathbf{P}_i,$$

yields, assuming for the moment that all inverses exist,

$$\mathbf{P}^{-1/2}\left(\sum_{i=1}^{n} \mathbf{P}_i\mathbf{X}_i \left(\mathbf{X}_i'\mathbf{P}_i\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{P}_i\right)\mathbf{P}^{-\frac{1}{2}}\mathbf{Y}_s = \mathbf{Y}_s\mathbf{\Lambda},$$

where $\mathbf{Y}_s$ is an orthonormal matrix of eigenvectors and $\mathbf{\Lambda}$ is the corresponding diagonal matrix containing the k-largest eigenvalues in decreasing order. The solution thus becomes:

$$\begin{aligned}\mathbf{Y} &= \mathbf{P}^{1/2}\mathbf{Y}_s \\ \mathbf{A}_i &= \left(\mathbf{X}_i'\mathbf{P}_i\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{P}_i\mathbf{Y},\end{aligned}$$

and

$$\mathbf{a}_{i0} = \left(\mathbf{1}'\mathbf{K}_i\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{K}_i\left(\mathbf{Y} - \mathbf{X}_i\mathbf{A}_i\right).$$

Comparison with the results in Sect. 3.1 immediately shows the similarity between the two methods. The only difference concerns the standardization with respect to $\mathbf{P}$ rather than $\mathbf{K}$. This computationally minor difference has considerable consequences. In the missing-data-passive approach, the data are taken into deviation from the mean over observed rows. As mentioned before, such a procedure may not be appropriate if the missing values are related to certain aspects of a variable, certain values will be under– or over represented. Hence, the mean over the observed values is a biased estimator of the constant term. The Test Equating method does account for this problem. Finally, an additional advantage of the Test Equating method is that the constraint (10), ensures that the group configuration is centered. That is, $\mathbf{Y}'\mathbf{1} = \mathbf{0}$. In the missing-data-passive approach, this is not the case. Moreover, if missings indeed do occur completely at random, the estimated mean in the Test Equating method will be similar to the mean in the missin-passive-approach. Hence, in that case, the methods yield similar results.

4 Iterative imputation based methods for dealing with missing values in generalized canonical correlation analysis

An important advantage of the two methods described in the previous section is its computational simplicity. On the other hand, an important disadvantage of both methods is that by removing entire rows, useful and valid information may be discarded. This may especially be the case when data matrices consists of many columns. To overcome this problem, one may choose an imputation method. Some general imputation methods, i.e. multiple imputation (Little and Rubin 1987; Rubin 1987) can be used. However, these schemes require some distributional assumptions. If one does not want to make such assumptions, methods specifically designed for generalized canonical correlation analysis may be employed. Green and Carroll (1988) proposed an imputation based method that uses regression analysis to impute the missings. However, convergence is not guaranteed for their algorithm. Furthermore, there is no mechanism that ensures that the final solution is, with respect to the carried out minimization problem, better than the starting solution. We propose a new algorithm that specifically addresses these two issues. For the sake of clarity we briefly reiterate the Green and Carroll algorithm before introducing our new method.

4.1 Green and Carroll's GENCOM algorithm

Green and Carroll (1988) proposed an iterative procedure for dealing with missing elements in generalized canonical correlation analysis. For the sake of clarity we will briefly reiterate their method here. The basic principle in their approach, which they call GENCOM, is to estimate the missing values using linear regressions of the variables on the group space.

The GENCOM algorithm can be summarized as follows:

1. For each $\mathbf{X}_i$ calculate $\hat{\mathbf{X}}_i^{(t)}$ by replacing the missing values by the column averages. Thus, for each column, the average is calculated by summing the observed values and dividing this through the total number of observations in a column.

2. Calculate $\mathbf{Y}^{(t)}$ by applying generalized canonical correlation analysis to the $\hat{\mathbf{X}}_i^{(t)}$ matrices, and by then adding a column of ones to the configuration matrix. This column of ones serves to estimate the constant in the linear regression model carried out in the next step.
3. For each column of $\mathbf{X}_i$, use ordinary least squares to fit: $\mathbf{x}_{i(j)}^{*} = \mathbf{Y}^{(t)*}\mathbf{b}_i^{(t)}$, where $\mathbf{x}_{i(j)}^{*}$ is the jth column of the original data matrix $\mathbf{X}_i$ after removing the rows corresponding to missing values for that column, and $\mathbf{Y}^{(t)*}$ is the matrix of corresponding rows of $\mathbf{Y}^{(t)}$. Hence, the regression is based on observed values only and $\mathbf{b}_i^{(t)} = \left(\mathbf{Y}^{(t)*\prime}\mathbf{Y}^{(t)*}\right)^{-1}\mathbf{Y}^{(t)*\prime}\mathbf{x}_{i(j)}^{*}$.
4. Construct $\mathbf{B}_i^{(t)} = \left[\mathbf{b}_1^{(t)}\ \mathbf{b}_2^{(t)}\ \ \mathbf{b}_{p_i}^{(t)}\right]$ and let $\mathbf{X}_i^{(t)*} = \mathbf{Y}^{(t)}\mathbf{B}_i^{(t)}$.
5. Calculate $\hat{\mathbf{X}}_i^{(t+1)}$ by replacing the missing values of the original $\mathbf{X}_i^{*}$ matrix with the corresponding elements of $\mathbf{X}_i^{(t+1)*}$, whilst keeping the observed values unaltered.
6. Insert $\hat{\mathbf{X}}_i^{(t+1)}$ in step 2, and repeat until the differences between two subsequent $\mathbf{Y}^{(t)}$ matrices becomes smaller than a certain convergence criterion.
   Note that, like before, index i indicates different observation matrices, whereas index t was used to indicate different iterations.

4.2 Minimized contribution approach

Green and Carroll (1988) do not give details on numerical properties of their algorithm. There are, however, two important issues concerning the GENCOM algorithm. First of all, although in each step $\mathbf{Y}$ and $\mathbf{B}_i$ are optimal with respect to the imputed $\mathbf{X}_i$ matrices, there is no mechanism ensuring that subsequent $\mathbf{Y}'s$ become more similar. That is, convergence is not guaranteed. Secondly, the value ϕ will always be at a minimum for a given set of (imputed) $\mathbf{X}_i$ matrices. However, there is no mechanism that ensures that this value will go down. Consequently, it may occur that the sum of differences between the group configuration $\mathbf{Y}$ and the linear combinations of the imputed $\mathbf{X}_i$ matrices is smaller in the first iteration than in the last iteration. (Obviously, if in subsequent steps the change in the $\mathbf{X}_i$ matrices is small it is plausible that the change in $\mathbf{Y}$ is also small. Hence, when the imputed values do not change, i.e. when the regression estimates are "stable" the group configuration is likely to be stable as well).

To resolve these issues we propose a new algorithm that imputes the missing values of the $\mathbf{X}_i$ matrices in such a way that the value of the objective function does not increase. Based on these imputed $\mathbf{X}_i$ matrices a new configuration is calculated in the usual way. Thus, the value of the objective function cannot increase in subsequent steps of the iteration process. The algorithm terminates when the value of the objective function remains constant. Recently, Albers and Gower (2010) developed a general approach to handling missing values in Procrustes analysis. Our algorithm fits in their framework and resembles an algorithm proposed by Ten Berge et al. (1993) for the treatment of missing values in generalized Procrustes analysis with orthogonal rotations.

The new algorithm that we propose is an alternating least-squares algorithm. The imputed values will be chosen in such a way that their contribution to the objective is

minimized. To achieve this, we first solve the usual generalized canonical correlation analysis problem with respect to $\mathbf{Y}$ and $\mathbf{A}_i$ whilst considering the $\mathbf{X}_i$ matrices, in which the missing elements are replaced by some initial values, constant. Next, we will use the same objective function but this time we minimize with respect to the missing values for $\mathbf{X}_i$ whilst considering $\mathbf{Y}$ and $\mathbf{A}_i$ constant. This process is then repeated until the value of the objective function remains constant. As the value of the objective function cannot increase in subsequent steps, convergence is guaranteed.

Recall objective function (1), where the $\mathbf{X}_i$ matrices may contain missing elements. We will impute values for the missings in such a way that the value of the objective function decreases in each step. Hence, while keeping $\mathbf{Y}$ and $\mathbf{A}_i$ fixed we must minimize ϕ with respect to the missing (to be imputed) elements of $\mathbf{X}_i$. This problem has not been solved before.

We can formulate the problem in the following way. Let

$$\mathbf{X}_i = \mathbf{X}_i^o + \mathbf{X}_i^m, \tag{11}$$

where $\mathbf{X}_i^o$ is the $m \times p_i$ matrix with the observed values and zeros for the non-observed values. The values of $\mathbf{X}_i^o$ are constant whereas the entries in $\mathbf{X}_i^m$ that correspond to missing values are the variables with respect to which we carry out the minimization. The other entries, corresponding to observed values, of $\mathbf{X}_i^m$ will be ignored. Using (11) we get

$$\mathbf{Y} - \mathbf{X}_i\mathbf{A}_i = \mathbf{Y} - \mathbf{X}_i^0\mathbf{A}_i - \mathbf{X}_i^m\mathbf{A}_i = \mathbf{Y}_i^* - \mathbf{X}_i^m\mathbf{A}_i$$

so that we can express the objective as

$$\min \phi = \min \sum_{i=1}^{n} \operatorname{trace} \left(\mathbf{Y}_i^* - \mathbf{X}_i^m\mathbf{A}_i\right)' \left(\mathbf{Y}_i^* - \mathbf{X}_i^m\mathbf{A}_i\right).$$

We want to minimize this function with respect to the variable elements of $\mathbf{X}_i^m$. Clearly

$$\operatorname{trace} \left(\mathbf{Y}_i^* - \mathbf{X}_i^m\mathbf{A}_i\right)' \left(\mathbf{Y}_i^* - \mathbf{X}_i^m\mathbf{A}_i\right) = \operatorname{vec} \left(\mathbf{Y}_i^* - \mathbf{X}_i^m\mathbf{A}_i\right)' \operatorname{vec} \left(\mathbf{Y}_i^* - \mathbf{X}_i^m\mathbf{A}_i\right),$$

where the vec operator transforms a matrix to a vector by stacking the columns. Using a well known relationship between the vec operator and the Kronecker product (see e.g. Magnus and Neudecker 1998) we get

$$\operatorname{vec} \left(\mathbf{Y}_i^* - \mathbf{X}_i^m\mathbf{A}_i\right) = \operatorname{vec} \left(\mathbf{Y}_i^*\right) - \left(\mathbf{A}_i' \otimes \mathbf{I}_m\right) \operatorname{vec} \mathbf{X}_i^m.$$

The matrix $\mathbf{X}_i^m$, and hence its vectorization, contains several elements which correspond to observed values. These elements are of no importance and should be kept constant. By employing a selection matrix $\mathbf{L}_i$ we select only those elements in $\mathbf{X}_i^m$ which correspond to the missing values. Let the number of missing values in the

original $\mathbf{X}_i$ matrix be q_i. A $q_i \times mp$ matrix $\mathbf{L}_i$, whose elements are zero or one, is constructed in such a way that

$$\mathbf{L}_i \operatorname{vec} \mathbf{X}_i^m = \mathbf{x}_i.$$

Hence, $\mathbf{x}_i$ is a $q_i \times 1$ vector whose elements we want to determine in such a way that ϕ is minimized. It is not difficult to see that $\mathbf{L}_i^{'}\mathbf{L}_i \operatorname{vec} \mathbf{X}_i^m = \operatorname{vec} \mathbf{X}_i^m$, so that

$$\operatorname{vec}\left(\mathbf{Y}_i^* - \mathbf{X}_i^m \mathbf{A}_i\right) = \mathbf{y}_i - \mathbf{C}_i \mathbf{x}_i,$$

where $\mathbf{y}_i = \operatorname{vec}\left(\mathbf{Y}_i^*\right)$ and $\mathbf{C}_i = \left(\mathbf{A}_i^{'} \otimes \mathbf{I}_m\right) \mathbf{L}_i^{'}$ and we can express the objective function as

$$\min_{\mathbf{x}_i} \phi = \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{C}_i \mathbf{x}_i)^{'} (\mathbf{y}_i - \mathbf{C}_i \mathbf{x}_i).$$

This problem can be solved using matrix differentiation. As first-order condition for $\mathbf{x}_i$ we get

$$\mathbf{C}_i^{'} \mathbf{y}_i = \mathbf{C}_i^{'} \mathbf{C}_i \mathbf{x}_i. \tag{12}$$

Hence, if $\left|\mathbf{C}_i^{'} \mathbf{C}_i\right| \neq 0$,

$$\mathbf{x}_i = \left(\mathbf{C}_i^{'} \mathbf{C}_i\right)^{-1} \mathbf{C}_i^{'} \mathbf{y}_i. \tag{13}$$

Moreover, if $\left|\mathbf{C}_i^{'} \mathbf{C}_i\right| = 0$, a vector $\mathbf{x}_i$ satisfying the first-order condition (12) may be obtained by replacing the inverse of $\mathbf{C}_i^{'} \mathbf{C}_i$ by its Moore–Penrose inverse.

The updated $\mathbf{X}_i$ matrices can be obtained by inserting the q_i elements of $\mathbf{x}_i$ in the appropriate places.

The algorithm can be summarized as follows:

1. Replace the non-observed values in the original $\mathbf{X}_i$ matrices by some initial values, for example, the column averages or zeros.
2. Center the imputed $\mathbf{X}_i$ matrices.
3. Calculate the generalized canonical correlation analysis solution, i.e. the group configuration $\mathbf{Y}$ and the value of the objective function ψ, in the usual way using the imputed $\mathbf{X}_i$ matrices.
4. Use (13) to calculate the vector with missing values $\mathbf{x}_i$, and update the $\mathbf{X}_i$ matrices accordingly.
5. Go back to step 2 and repeat until the difference between two subsequent values for the objective function ψ is negligible.

The new algorithm will always converge as the value of the objective function (1) decreases monotonically. It may be possible that the convergence point is a mere

accumulation point. To avoid this, random starts may be considered. Another problem that may occur involves degenerate solution in the sense that imputed values become extremely large, thus dominating the solution. In practical situations this can easily be avoided by imposing a restriction on the values to be imputed. For example, if the original data are ratings, the maximum rating is an obvious restriction. For continuous data, natural extremes also exist. Alternatively, the values can be bounded by taking the mean value plus, for example, four standard deviations.

5 Simulation study

To investigate the properties of all four approaches, we conduct a simulation study. In the simulation study, synthetic data are generated for several parameter settings so that the methods can be evaluated under various conditions. To assess the performance of the methods, we consider the measures "variance accounted for" (VAF) and the alienation coefficient. In Sect. 5.2 we describe these measures and their functions.

5.1 Data generation process

The data generation process can be summarized as follows:

1. For fixed m and k, an $m \times k$ group configuration $\mathbf{Y}_{true}$ is constructed by drawing from a standard normal distribution and then calculating an orthogonal base.
2. For each observation matrix we draw an $m \times k$ (standard normal) random matrix multiplied by a factor $r = 0.125$, and add this matrix to $\mathbf{Y}_{true}$. The resulting matrix is then post-multiplied by a $k \times p_i$ (uniform) random matrix to obtain the ith observation matrix $\mathbf{X}_i$.
3. For each $m \times p_i$ observation matrix $\mathbf{X}_i$, we draw a matrix indicating which elements are observed and which are missing.
4. We repeat this process n times leading to n "observation" matrices $\mathbf{X}_i$.

5.2 Evaluation criteria and analysis

After generation of the data sets, we apply the four methods described in this paper. To assess the performance of the methods we consider how well the obtained group configuration is able to describe the original data. Steenkamp et al. (1994), proposed the following measure, which they called variance accounted for (VAF). Select the jth column of $\mathbf{X}_i$, say $\mathbf{x}_{i(j)}$ and calculate the multiple squared correlation coefficient, R^2, from the linear regression $\mathbf{x}^*_{i(j)} = \mathbf{Y}^*\mathbf{b}_j + e_{ij}$, where the superscripted * indicates that the rows of $\mathbf{Y}$ and $\mathbf{x}_{i(j)}$ corresponding to missing rows (i.e. elements) of $\mathbf{x}_{i(j)}$ have been removed. Repeat this for all columns of $\mathbf{X}_i$ and for all data matrices. The VAF is defined as the average of all calculated squared multiple correlation coefficients.

In addition, as the true configuration is known, we can also assess how well the solutions "recover" the true configuration. To do this we compare the Euclidean distances between the rows of the true configuration, with the Euclidean distances between the rows of the retrieved configuration. Let $\mathbf{T}$ denote the matrix with as

elements the Euclidean distances between the rows of the true configuration. The ijth element of $\mathbf{T}$ is: $t_{ij} = \sqrt{\left(\mathbf{y}_i^{true} - \mathbf{y}_j^{true}\right)' \left(\mathbf{y}_i^{true} - \mathbf{y}_j^{true}\right)}$, where $\mathbf{y}_i^{true}$ is the ith row of $\mathbf{Y}_{true}$ written as a $k \times 1$ column vector. Similarly, let $\mathbf{O}$ denote the matrix with as elements the Euclidean distances between the rows of the derived configuration: $o_{ij} = \sqrt{\left(\mathbf{y}_i - \mathbf{y}_j\right)' \left(\mathbf{y}_i - \mathbf{y}_j\right)}$, where $\mathbf{y}_i$ is the ith row of the obtained group configuration $\mathbf{Y}$, written as a $k \times 1$ column vector. A congruence coefficient, which measures to which degree the Euclidean distances in the two configurations are similar, may be defined as:

$$c = \frac{\text{trace}(\mathbf{T}'\mathbf{O})}{\sqrt{\text{trace}(\mathbf{T}'\mathbf{T})\,\text{trace}(\mathbf{O}'\mathbf{O})}}.$$

The congruence coefficient lies between zero and one, where the maximum is attained when the distances in the two configurations are equal. To allow easier discriminations, Borg and Leutner (1985) introduce the following alienation coefficient:

$$a = \sqrt{\left(1 - c^2\right)}.$$

Similar to Bijmolt and Wedel (1999), we will use this alienation coefficient, which may be interpreted as the squared root of unexplained variance, to assess how well the true configuration is recovered.

5.3 Experimental design

In generating the synthetic data sets, we fix the dimensionality of both the true and approximated group configuration at two. We then vary a number of factors that might affect the performance of the methods. Concerning the number of objects (rows) per matrix we consider two cases: Relatively few rows for each set and relatively many rows for each set. We will treat these two cases separately.

5.3.1 Relatively few rows for each set: $m = 14$

This corresponds, for example, to applications in which a set of objects (the rows) are evaluated using a set of attributes (the columns). The number of objects that individuals are able to assess will differ from application to application. We chose 14 objects to mimic situations in marketing research where the rows correspond to brands. In such cases, the number of objects typically lies between 10 and 20. The number of attributes are varied in such a way that:

1. The number of attributes for each set is obtained by drawing from a normal distribution with mean 4 and standard deviation 2, and rounding the number to the nearest integer, with a minimum value of 2. Hence, the expected value for the number of columns is slightly higher than 4: $E[p] > 4$.

Table 1 Average variation accounted for VAF

$E[p] >$ Missings	Missing passive		Test equating		Gencom		Min. contribution	
	4	8	4	8	4	8	4	8
CAR 5%	0.71	0.51	0.74	0.66	0.75	0.76	0.79	0.77
CAR 10%	0.62	0.30	0.72	0.41	0.75	0.76	0.76	0.73
CAR 20%	0.50	0.26	0.67	0.29	0.75	0.76	0.71	0.71
CAR 40%	0.47	0.29	0.57	0.28	0.75	0.75	0.70	0.71
VDM, highest 1	0.60	0.35	0.69	0.55	0.72	0.73	0.74	0.70
VDM, highest 2	0.45	0.20	0.63	0.33	0.70	0.71	0.69	0.64
VDM, highest 3	0.35	0.20	0.58	0.29	0.70	0.71	0.67	0.57

Number of rows 14. Number of observation matrices $n = 10$

2. The number of attributes for each set is obtained by drawing from a normal distribution with mean 8 and standard deviation 2, and rounding the number to the nearest integer, with a minimum value of 2. Hence, the expected value for the number of columns is slightly higher than 8: $E[p] > 8$.

For the number of sets we also consider two cases:

1. Few (10) sets. This corresponds to the situation in which, for example, different multidimensional scaling configurations are compared
2. Many (100) sets. This corresponds to the situation in which each data matrix represents an observation matrix for an individual.

To ensure that we always obtain a solution we always generate one full data matrix (this is essential for the noniterative methods where complete data matrices are discarded when there are relatively many missing values). For the remaining matrices we generate the missings according to the following scenarios:

(a) Completely missing at random (CAR).
(b) Value dependent missings (VDM).

In scenario (a), missings occur completely at random, we consider four cases with probabilities for missing values in each observation matrix, equal to 5, 10, 20 and 40% respectively. Under scenario (b), we consider the situation in which the probability of values to be missing is directly related to the simulated values. This could, for example, occur when certain true values are less desirable and hence reluctantly reported.We consider three cases. For each column, the elements corresponding to (1) the highest three, (2) the highest two and (3) the highest value, are missing.

Results The results of the simulation study with few rows are presented in Tables 1, 2, 3 and 4. It should be noted that for the minimized contribution approach we restricted the imputed values to be less than 4 standard deviations from the observed values. Without such a restriction the results deteriorate considerably. Especially when the number of missing values increases. Furthermore, as in the noniterative approaches entire rows are removed when one or more values in that

Table 2 Average alienation coefficients

$E[p] >$ Missings	Missing passive		Test equating		Gencom		Min. contribution	
	4	8	4	8	4	8	4	8
CAR 5%	0.14	0.30	0.14	0.24	0.11	0.09	0.12	0.13
CAR 10%	0.18	0.47	0.18	0.45	0.11	0.09	0.14	0.16
CAR 20%	0.30	0.54	0.28	0.54	0.13	0.10	0.19	0.19
CAR 40%	0.42	0.55	0.40	0.56	0.17	0.16	0.27	0.22
VDM, highest 1	0.18	0.38	0.18	0.33	0.12	0.11	0.16	0.16
VDM, highest 2	0.29	0.50	0.27	0.47	0.14	0.13	0.20	0.22
VDM, highest 3	0.40	0.50	0.33	0.51	0.15	0.14	0.23	0.27

Number of rows 14. Number of observation matrices $n = 10$

Table 3 Average variation accounted for VAF

$E[p] >$ Missings	Missing passive		Test equating		Gencom		Min. contribution	
	4	8	4	8	4	8	4	8
CAR 5%	0.70	0.59	0.77	0.76	0.77	0.76	0.79	0.77
CAR 10%	0.61	0.40	0.77	0.75	0.77	0.77	0.76	0.75
CAR 20%	0.48	0.27	0.76	0.52	0.77	0.77	0.73	0.74
CAR 40%	0.43	0.33	0.72	0.31	0.78	0.78	0.70	0.73
VDM, highest 1	0.58	0.46	0.73	0.72	0.74	0.73	0.74	0.72
VDM, highest 2	0.43	0.26	0.71	0.61	0.72	0.72	0.69	0.68
VDM, highest 3	0.32	0.24	0.69	0.42	0.73	0.72	0.69	0.63

Number of rows 14. Number of observation matrices $n = 100$

Table 4 Average alienation coeficient

$E[p] >$ Missings	Missing passive		Test equating		Gencom		Min. contribution	
	4	8	4	8	4	8	4	8
CAR 5%	0.05	0.08	0.05	0.06	0.04	0.04	0.05	0.05
CAR 10%	0.06	0.25	0.06	0.11	0.04	0.04	0.05	0.06
CAR 20%	0.11	0.51	0.10	0.39	0.05	0.04	0.06	0.09
CAR 40%	0.29	0.55	0.23	0.55	0.07	0.07	0.10	0.14
VDM, highest 1	0.10	0.22	0.09	0.12	0.06	0.05	0.08	0.07
VDM, highest 2	0.17	0.60	0.12	0.28	0.06	0.06	0.09	0.09
VDM, highest 3	0.32	0.56	0.18	0.43	0.07	0.06	0.11	0.14

Number of rows 14. Number of observation matrices $n = 100$

row are missing, the situation may occur that for certain rows no observations are available. This will especially be the case when we have relatively few sets. In our simulation study, we avoided this situation by always generating one full data matrix.

We see that increasing the number of missings generally leads to a decrease in fit. This decrease, however, appears to be much stronger for the non-iterative approaches. In particular, with many missing values there is a strong decrease. For the situation in which there are few (10) observation matrices (see Tables 1, 2) we also see that the performance of the non-iterative approaches suffers when the number of columns is increased. One reason for this is that for the non-iterative approaches, entire rows are removed when one or more values in that row are missing. With many columns, this may lead to a considerable loss in information.

In Tables 1 and 3 we see that, with respect to the variance accounted for, the Test Equating method outperforms the missing-data-passive method in all cases. Furthermore, as conjectured in Sect. 3.2, the Test Equating method clearly outperforms the missing-data-passive approach when the missings are not random. Note also that, although the iterative approaches outperform both non-iterative approaches in nearly all cases, the differences, especially when there are many sets and few columns, are small.

Comparison of Tables 1 and 3 reveals that, for all methods, the influence of having more data matrices on the VAF is limited. On the other hand, Tables 2 and 4 show that recovery of the true configuration slightly improves (i.e. the alienation coefficients decrease) when more sets are available.

5.3.2 Relatively many rows for each set: $m = 100$

This setting corresponds to applications where the rows correspond to cases. The columns represent variables. Each matrix has observations on sets of variables. In a sense, this could be considered the conventional generalized canonical correlation analysis case. For the number of sets we again consider two cases: 4 sets and 8 sets. Furthermore, the number of columns (i.e. variables per set) is varied in the same manner as before. That is, the expected number of variables per set is either slightly larger than 4 or slightly larger than 8. For the missing values in this setting we again consider the scenarios in which missings occur completely at random (CAR) or are related to the underlying values (VDM).

For the missing completly at random scenario we consider the same 4 cases as before, i.e. missings occur with probabilities 5, 10, 20 and 40%. In the value dependent missings we again consider three scenarios; the highest 5, 10, and 20% of the values are missing. Again, we restricted the imputed values in the minimized contribution approach to be less than 4 standard deviations from the observed values and we always generated one complete matrix in each simulation.

Results The results of the simulation study with sets of $m = 100$ rows, are presented in Tables 5, 6, 7 and 8. We see that the variance accounted for in this scenario is generally lower than in the setting with few observations whereas the alienation coefficients are quite a bit larger. Again we see that the Test Equating method outperforms the missing-data-passive approach. Especially when the missings do not occur at random. Also, although the iterative approaches yield better results, the differences with the Test Equating method are, especially when there are relatively few missings, not that large. If we compare Tables 5 with 7 and 6 with 8, we see that the variance

Table 5 Average variation accounted for VAF

$E[p]$ > Missings	Missing passive		Test equating		Gencom		Min. contribution	
	4	8	4	8	4	8	4	8
CAR 5%	0.46	0.38	0.49	0.46	0.51	0.48	0.50	0.47
CAR 10%	0.42	0.30	0.48	0.43	0.51	0.48	0.49	0.46
CAR 20%	0.35	0.18	0.43	0.32	0.52	0.48	0.48	0.44
CAR 40%	0.27	0.18	0.36	0.20	0.53	0.49	0.46	0.41
VDM, highest 5%	0.53	0.27	0.74	0.42	0.75	0.43	0.74	0.43
VDM, highest 10%	0.35	0.18	0.43	0.39	0.47	0.42	0.46	0.41
VDM, highest 20%	0.28	0.14	0.38	0.29	0.46	0.43	0.43	0.38

Number of rows 100. Number of observation matrices $n = 4$

Table 6 Average alienation coeffcients

$E[p]$ > Missings	Missing passive		Test equating		Gencom		Min. contribution	
	4	8	4	8	4	8	4	8
CAR 5%	0.30	0.24	0.30	0.24	0.27	0.19	0.30	0.22
CAR 10%	0.34	0.29	0.34	0.29	0.28	0.20	0.34	0.25
CAR 20%	0.41	0.44	0.41	0.43	0.30	0.21	0.37	0.29
CAR 40%	0.49	0.56	0.49	0.56	0.33	0.23	0.42	0.34
VDM, highest 5%	0.32	0.19	0.32	0.19	0.28	0.14	0.30	0.16
VDM, highest 10%	0.53	0.26	0.74	0.25	0.75	0.15	0.74	0.18
VDM, highest 20%	0.44	0.46	0.43	0.42	0.37	0.23	0.32	0.29

Number of rows 100. Number of observation matrices $n = 4$

Table 7 Average variation accounted for VAF

$E[p]$ > Missings	Missing passive		Test equating		Gencom		Min. contribution	
	4	8	4	8	4	8	4	8
CAR 5%	0.43	0.35	0.47	0.46	0.48	0.47	0.48	0.46
CAR 10%	0.39	0.27	0.47	0.44	0.48	0.47	0.47	0.46
CAR 20%	0.31	0.16	0.43	0.37	0.49	0.47	0.46	0.44
CAR 40%	0.21	0.14	0.34	0.18	0.50	0.47	0.44	0.40
VDM, highest 5%	0.37	0.27	0.44	0.42	0.45	0.43	0.45	0.43
VDM, highest 10%	0.30	0.18	0.41	0.39	0.44	0.42	0.43	0.41
VDM, highest 20%	0.22	0.10	0.35	0.31	0.43	0.41	0.40	0.37

Number of rows 100. Number of observation matrices $n = 8$

accounted for appears to decrease when more sets are used whereas the recovery of the true configuration improves upon having more sets. If we have more sets, it becomes easier to fit noise, leading to the increase of variance accounted for, whilst more data also leads to a better recovery of the true configuration.

Table 8 Average alienation coefficients

$E[p] >$ Missings	Missing passive		Test equating		Gencom		Min. contribution	
	4	8	4	8	4	8	4	8
CAR 5%	0.21	0.17	0.21	0.17	0.19	0.14	0.21	0.16
CAR 10%	0.25	0.22	0.25	0.22	0.20	0.14	0.24	0.18
CAR 20%	0.32	0.36	0.32	0.35	0.21	0.15	0.29	0.23
CAR 40%	0.46	0.55	0.46	0.55	0.25	0.17	0.36	0.31
VDM, highest 5%	0.24	0.19	0.24	0.19	0.20	0.14	0.21	0.16
VDM, highest 10%	0.29	0.26	0.29	0.25	0.21	0.15	0.24	0.18
VDM, highest 20%	0.40	0.41	0.38	0.36	0.24	0.17	0.28	0.23

Number of rows 100. Number of observation matrices $n = 8$

6 Illustration

The data catalog of the world bank (http://data.worldbank.org/) provides download access to over 2,000 indicators from world bank resources. From the data catalog we collected country-wise indicators from the year 2007. The world bank data catalog organizes the variables into different sets. For each set, a selection of the available indicators was chosen based on their appropriateness, uniqueness (some variables are highly correlated as they are essentially the same variable measured in a different way) and availability (measurements on some variables are only available for a small sub-group of countries). Note that our purpose here is to illustrate the method. We do not claim to give a comprehensive analysis of the countries' positions with respect to the world bank data.

The world bank distinguishes 10 sets of indicators. Due to data scarcity in several sets we selected indicators from 6 of these sets: Economic policy and debt (10 indicators), education (19 indicators), environment (16 indicators), health (17 indicators), infrastructure (7 indicators), labor (5 indicators). The original data consisted of over 200 countries and regions. The regional observations are aggregated results for regions constructed on the grounds of geographical classifications (i.e. Sub-Saharan Africa) or economic classifications (i.e. High Income). As in the non-iterative approaches complete rows (i.e. countries or regions) are removed if one variable is missing in a set, we removed all countries/regions that did not have at least two complete sets of observations. This makes it possible to compare all approaches. The final data set used in our analysis consisted of 131 countries and 17 regions. On average, approximately 10% of the values were missing. The sets corresponding to education and labor indicators had 23 and 13% missings respectively. Country-wise, the number of missings (over all included indicators) ranged between no missings to 39% missings. Note that, for our purpose of illustrating the described methods, we decided to keep the data cleansing to a minimum. The obtained results may possibly be improved upon by using a more thorough and expert driven data cleansing procedure.

We applied all four approaches to the final data set. To allow comparisons as well as graphical representations of the solutions, we chose a 2 dimensional solution in

Table 9 Alienation coefficients between different group configurations

	Gencom	Min. contribution	Missing passive
Test equating	0.25	0.25	0.08
Missing passive	0.28	0.25	
Min. contribution	0.25		

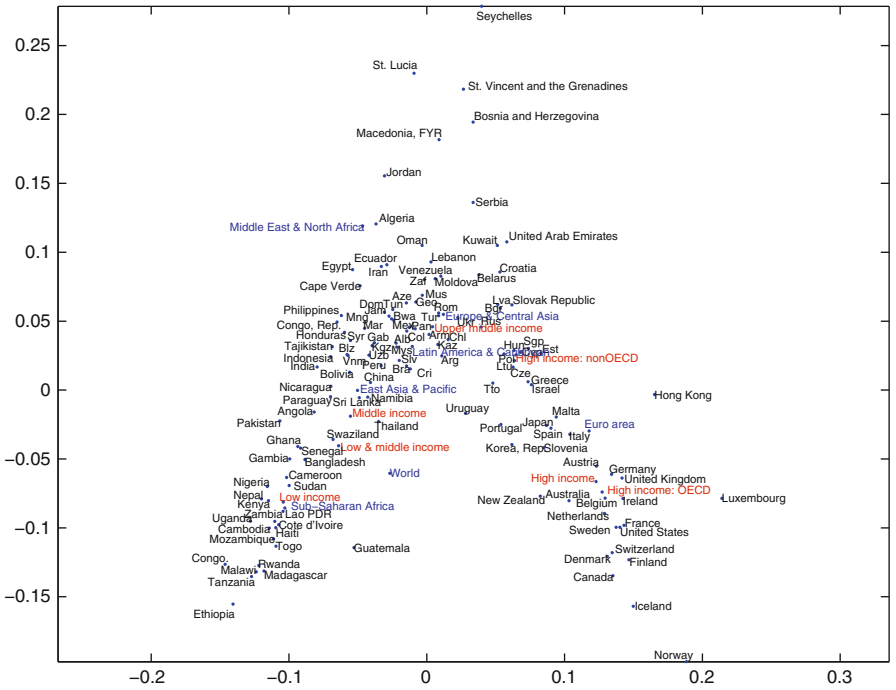

Fig. 1 Generalized canonical correlation analysis. Two dimensional group configuration of countries based on 2007 World Bank Data. See Table 10 for country labels

all analyses. Increasing the number of dimensions had a relatively small effect on the variance accounted measures in all analyses indicating the appropriateness of this choice. The minimized contribution approach yields the largest variance accounted for (0.42) but the difference with the gencom approach is small (0.41). The non-iterative approaches, on the other hand both yield a considerably lower amounts of variance accounted for 0.26. To see whether the methods yielded similar group configurations we calculated the alienation coefficients between all pairs. The results are summarized in Table 9. We see that the missing-data-passive and Test Equating method yield nearly equivalent results, whereas the differences with and among the non-iterative approaches are much larger.

Figure 1 presents the two dimensional group configurations obtained using the minimized contribution approach. For a quick interpretation we focus on the positions of the aggregated observation points. The countries appear to be ordered from poor (Low Income, on the lower left side) to rich (High Income, on the lower right). Geograph-

Table 10 Country labels for selected countries in Fig. 1

Abbreviation	Country	Abbreviation	Country
Alb	Albania	Kgz	Kyrgyz Republic
Arg	Argentina	Lva	Latvia
Arm	Armenia	Ltu	Lithuania
Aze	Azerbedjan	Mys	Malaysia
Blz	Belize	Mus	Mauritius
Bwa	Botswana	Mex	Mexico
Brz	Brazil	Mng	Mongolia
Bgr	Bulgaria	Mar	Morocco
Chl	Chili	Pan	Panama
Col	Columbia	Pol	Poland
Cri	Costa Rica	Rom	Romania
Cyp	Cyprus	Rus	Russian Federation
Cze	Czech Republic	Sgp	Signapore
Dom	Dominican Republic	Zaf	South Africa
Slv	El Salvador	Syr	Syrian Arab Republic
Est	Estonia	Tto	Trinidad and Tobago
Gab	Gabon	Tun	Tunesia
Geo	Georgia	Tur	Turkey
Hun	Hungaria	Ukr	Ukraine
Jam	Jamaica	Uzb	Uzbekistan
Kaz	Kazakhstan	Vnm	Vietnam

ically, this corresponds to a trajectory from Sub-Saharan Africa through East-Asian & Pacific, Middle-East and North Africa, Latin America and Caribbean, Europe and Central Asia to the Euro Area. The second dimension is more difficult to interpret. On the top of the plot we do see relatively small countries however the arch shape of the cloud suggests that the solution is one-dimensional.

7 Summary and conclusions

Generalized canonical correlation analysis is a mathematically simple, yet versatile technique with potential applications in many fields of research. In generalized canonical correlation analysis, linear combinations of sets of variables are obtained in such a way that the sum of squared distances between the linear combintation and an overall group configuration becomes minimal. When the data sets contain missing values, two procedures exist: Missing-data-passive, in which rows for which a missing value exists, are removed from the data, and GENCOM, an iterative approach proposed by Green and Carroll (1988), where missing vaues are imputed based on linear regression estimates. In this paper, we introduced two new methods for dealing with missing values in generalized canonical correlation analysis. The first approach,

the Test Equating method, does not require iterations. Like the missing-data-passive method, it removes rows that contain missing values. When missings do not occur completely at random, the existing missing-data-passive procedure yields biased results as the data are considered in deviation from the mean of the observed rows. We conjectured that the new Test Equating method, in which a constant term is separately estimated, would perform better when missings do not occur at random. Our simulation study clearly confirmed this conjecture. The Test Equating method consistenly outperformed the missing-data passive approach and the difference in performance increased with an increase in missing values. Given the computational similarity and simplicity of these two non-iterative approaches, the Test Equating method is therefore to be preferred over the missing-data passive approach.

The second new approach derived in this paper is the minimized contribution approach. In this approach, missing values are imputed in such a way that the generalized canonical correlation analysis objective function is minimized. Unlike the missing-data-passive and Test Equating method, no data are discarded in this method. Instead, an iterative procedure is employed to obtain the optimal values. This method can be seen specifically as an alternative to the GENCOM algorithm. An important theoretical advantage of the minimized contribution approach is that it, unlike the GENCOM algorithm, always converges. In our simulation experiment, we indeed found that the GENCOM algorithm in some cases fails to converge. However, the results of GENCOM appeared hardly affected by this. Moreover, the overall results of the simulation indicate that GENCOM performs slightly better than the minimized contribution approach.

Comparing the non-iterative and iterative approaches we see that the iterative approaches generally outperform the non-iterative ones. However, especially when the number of missings is small and there are relatively few columns (e.g. variables) per set, the differences with the Test Equating method are small. In such cases we therefore suggest to use this non-iterative approach as a direct and fast method. Alternatively, and especially when the number of variables per set is relatively large and there are many missings, either GENCOM or the minimized contribution approach can be used. The simulation study showed that differences are small but on average slightly in favor of GENCOM. As neither of these methods is computationally very demanding, it may be worthwhile considering both approaches and choosing the one attaining best fit and interpretational properties. In our illustration, the minimized contribution approach offered a satisfactory solution in terms of variance accounted for and interpretability.

References

Albers CJ, Gower JC (2010) A general approach to handling missing values in procrustes analysis. Adv Data Anal Classif 4:223–237

Bijmolt TH, Wedel M (1999) A comparison of multidimensional scaling methods for perceptual mapping. J Mark Res 36:277–285

Borg I, Leutner D (1985) Measuring the similarity between MDS configurations. Multivar Behav Res 20:325–334
Carroll JD (1968) Generalization of canonical correlation analysis to three or more sets of variables. In: Proceedings of the American psychological association, pp 227–228
Gifi A (1990) Nonlinear multivariate analysis. Wiley, Chichester
Green PE, Carroll JD (1988) A simple procedure for finding a composite of several multidimensional scaling solutions. J Acad Mark Sci 16:25–35
Horst P (1961) Generalized canonical correlation and their applications to experimental data. J Clin Psychol 17:331–347
Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377
Kettenring JR (1971) Canonical analysis of several sets of variables. Biometrika 58:433–451
Little R, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York
Magnus J, Neudecker H (1999) Matrix differential calculus with applications in statistics and econometrics. Wiley, Chichester
Meulman JJ (1982) Homogeneity analysis of incomplete data. DSWO Press, Leiden
Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
Shibayama T (1988) Kessokuchi o fukumu testo sukoa no tahenryoukaiseki (multivariate analysis of test scores with missing data). Unpublished Doctoral Dissertation, Faculty of Education, University (in Japanese)
Shibayama T (1995) A linear composite method for test scores with missing values. Niigata daigaku kyouikugakubu kiyou (Memoirs of the Faculty of Education, Niigata University) 36:445–455
Steenkamp J-BEM, Van Trijp HCM, Ten Berge JMF (1994) Perceptual mapping based on idiosyncratic sets of attributes. J Mark Res 31:15–27
Takane Y (1995) Seiyakutsuki Shuseibunbunsekihou (Constrained principal component analysis). Asakurashoten, Tokyo
Takane Y, Oshima-Takane Y (2003) Relationships between two methods for dealing with missing data in principal component analysis. Behaviormetrika 30:145–154
Ten Berge JMF, Kiers HAL, Commandeur JJF (1993) Orthogonal procrustes rotation for matrices with missing values. British J Math Stat Psychol 46:119–134
Van de Velden M, Bijmolt TH (2006) Generalized canonical correlation analysis of matrices with missing rows: a simulation study. Psychometrika 71:323–331
Van der Burg E (1988) Nonlinear canonical correlation and some related techniques. DSWO Press, Leiden
Zanakis SH, Alvarez C, Li V (2007) Socio-economic determinants of HIV/AIDS pandemic and nations efficiencies. Eur J Oper Res 176:1811–1838