University of Groningen

Clustering and dimension reduction for mixed variables

Vichi, Maurizio; Vicari, Donatella; Kiers, Henk A. L.

Published in:
Behaviormetrika

DOI:
[10.1007/s41237-018-0068-6](https://doi.org/10.1007/s41237-018-0068-6)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Vichi, M., Vicari, D., & Kiers, H. A. L. (2019). Clustering and dimension reduction for mixed variables. *Behaviormetrika*, 46, 243-269. <https://doi.org/10.1007/s41237-018-0068-6>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Clustering and dimension reduction for mixed variables

Maurizio Vichi¹ · Donatella Vicari¹ · Henk A. L. Kiers²

Received: 8 March 2018 / Accepted: 21 September 2018 / Published online: 11 March 2019
© The Behaviormetric Society 2018

Abstract

Clustering (partitioning) and simultaneous dimension reduction of objects and variables of a two-way two-mode data matrix is proposed here. The methodology is based on a general model that includes *K*-means clustering, factorial *K*-means, projection pursuit clustering (also known as reduced *K*-means), principal component analysis and intermediate cases of object clustering and variable reduction. Since we often have sets consisting of both qualitative and quantitative variables, the general model is now extended to deal with the general relevant case of mixed variables, analogous to variants of PCA handling qualitative (nominal and ordinal) variables in addition to quantitative variables. The model, called clustering and dimension reduction (CDR), is fully discussed in all the special cases cited above. For least-squares estimation of the model, an efficient coordinate descent algorithm is presented. Finally, a simulation study and two analyses on real data illustrate the features of CDR and study the performance of the proposed algorithm.

Keywords PCA · CATPCA · *K*-means · Factorial *K*-means · Reduced *K*-means · Dimension reduction

Communicated by Michel van de Velden.

✉ Donatella Vicari
donatella.vicari@uniroma1.it

Henk A. L. Kiers
h.a.l.kiers@rug.nl

¹ Department of Statistical Sciences, Sapienza University of Rome, P.le A. Moro 5, I-00185 Rome, Italy

² Heymans Institute, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

1 Introduction

The process of nontrivial extraction of potentially useful and unknown information via data reduction is increasingly being used by analysts and researchers. The idea is to identify key attributes and provide explicit information with a readable and understandable form for the users. The dimension reduction task is commonly achieved by advanced cluster analysis (CA) methodologies for the objects and by factor analysis (FA) techniques, including principal component analysis (PCA) for the variables.

Often, two-step tandem approaches are implemented where two different methods are applied sequentially, that is, the second is applied to results of the first. This can be achieved in two alternative ways which produce generally different reduced (mean profiles $\times$ factors) data matrices containing the relevant information from the observed data.

In the first sequential procedure, PCA or FA is applied first and then CA is applied on the reduced set of factors from PCA or FA. The second sequential procedure reverses the order in the application of the two methods, by applying at first CA on the original data, to obtain a reduced set of centroids, and then a reduced set of factors is extracted via PCA or FA computed on such centroids.

However, these two-step sequential procedures optimize two different objective functions with the result that dimensions defined by PCA or FA are not the best for CA or, vice versa, CA accounts for a part of the variance which is not necessarily the most relevant in determining the factors.

Several methodologies have been proposed to overcome this problem via simultaneous approaches which perform clustering of the objects and dimension reduction of the variables by optimizing one objective criterion.

Bock (1987), Van Buuren and Heiser (1989), and later De Soete and Carroll (1994) proposed a modified K -means procedure, named projection pursuit clustering, GROUPALS, and reduced K -means (PPC/GROUPALS/RKM), respectively, where each cluster is represented by a centroid in a low-dimensional space, chosen as the closest point to the objects in each cluster. In matrix algebra, the method can be described as follows. Let $\mathbf{X}$ denote an $I \times J$ columnwise-centered data matrix, $\mathbf{U}$ an $I \times K$ binary matrix indicating the memberships of the objects to K clusters, $\mathbf{Y}$ a $K \times Q$ cluster centroids matrix, $\mathbf{A}$ a $J \times Q$ loading matrix for Q dimensions with $\mathbf{A}'\mathbf{A} = \mathbf{I}$, and $\mathbf{E}$ a residual matrix. Then PPC/GROUPALS/RKM can be considered as fitting the model:

$$\mathbf{X} = \mathbf{UYA}' + \mathbf{E}, \quad (1)$$

by minimizing the loss function:

$$J_{PPC/GROUPALS/RKM}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{UYA}'\|^2. \quad (2)$$

This method can be seen as the simultaneous version of the second sequential procedure described above (i.e., K -means, modelling $\mathbf{X}$ as $\mathbf{UC}$, for a centroids matrix $\mathbf{C}$, followed by PCA on $\mathbf{UC}$). Because $\mathbf{UC}$ estimates the data by the corresponding cluster centroids, a PCA of this data set actually maximizes the between

variance of the clusters in the reduced space. See also Uno et al. (2016) for a factor analysis version of PPC/GROUPALS/RKM, though factor analysis is beyond the scope of this paper.

Vichi and Kiers (2001) proposed a different methodology named factorial *K*-means (FKM) that allows to select the most relevant components for the “best” classification. Factorial *K*-means consists of fitting the model:

$$\mathbf{XAA}' = \mathbf{UYA}' + \mathbf{E}. \quad (3)$$

The model is fitted in the least-squares sense by minimizing:

$$f_{FKM}(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = ||\mathbf{XAA}' - \mathbf{UYA}'||^2 \quad (4)$$

This can be seen as the simultaneous version of the first sequential procedure (i.e., PCA estimating $\mathbf{X}$ by $\mathbf{XAA}'$ followed by *K*-means on $\mathbf{XAA}'$). As the loss function actually consists of squared deviations of (projected) scores from their centroids, this method minimizes the within variance of the clusters in the reduced space. See also Timmerman et al. (2010) for a comparison of the two methods above.

Now, since the total variance accounted for in the reduced space is not constant, depending on the number of components retained and on the pattern of the component loadings, the PPC/GROUPALS/RKM maximizes the between variance, but unconditionally on the within variance, that in real applications could be quite large. On the other hand, FKM minimizes the within variance of the reduced space but unconditionally on the amount of the between variance that could be low. Therefore, it is useful to combine the two methodologies to get solutions optimizing both the between and the within variance. This idea is not new and has been discussed in Vichi and Kiers (2001), Hwang et al. (2006) and, for three-way data, by Vichi et al. (2007), while Yamamoto and Hwang (2014) developed a variant of this, for adequately dealing with situations where a set of strongly interrelated variables irrelevant to the clustering might adversely affect the solution. Such combinations of the two methodologies can be done in various ways, that are essentially equivalent. The present paper starts by describing the very general model that includes *K*-means clustering, factorial *K*-means, PPC/GROUPALS/RKM, principal component analysis and intermediate cases of object clustering and variable reduction. This very general model is here called the clustering and dimension reduction (CDR) model.

The main purpose of the present paper is to enhance the applicability of CDR to other than quantitative data. The first procedure for clustering and dimension reduction of mixtures of qualitative and quantitative data is GROUPALS (Van Buuren and Heiser 1989), which is equivalent to Van de Velden et al.’s (2017) cluster correspondence analysis in case of only categorical variables. GROUPALS combines reduced *K*-means clustering and optimal scaling (see Gifi 1990) of variables of different measurement levels. More recently, Hwang et al. (2006) have proposed a variant of multiple correspondence analysis that takes into account cluster-level heterogeneity in respondents’ preferences/choices. The method involves combining multiple correspondence analysis and *K*-means in a unified framework. The former is used for uncovering a low-dimensional space of multivariate categorical variables, while the latter is used for identifying relatively homogeneous clusters of respondents. In the present paper, we propose a variant of our general CDR method that

can handle qualitative and quantitative variables. We will show that GROUPALS is a special instance of this method. So in essence, we generalize the GROUPALS method to a more general and more flexible approach of combining clustering and dimension reduction.

The overview of the paper is as follows. In Sect. 2, we discuss the general model of CDR. In Sect. 3, it is described how, in the context of PCA, nominal and ordinal variables can be handled using optimal scaling (cf. Gifi 1990). In Sect. 4, it is explained how these approaches of handling nominal and ordinal variables are incorporated in our general model CDR. In Sect. 5, it is shown in what respect this method is a generalization of GROUPALS. Finally, a simulation study together with the analysis of two data sets illustrates the methodology in Sect. 6.

2 Clustering and dimension reduction of the variables (CDR)

Following Vichi et al. (2007) for the three-way case, we could propose a method that combines the loss functions of FKM and PPC/GROUPALS/RKM directly. However, here we propose to combine the PCA loss function and the FKM loss function, and it will be seen that PPC/GROUPALS/RKM then shows up ‘in the middle’. Thus, we propose CDR as the method that minimizes the loss function:

$$\begin{aligned} f(\mathbf{F}, \mathbf{A}, \mathbf{U}, \mathbf{Y}) &= \alpha \|\mathbf{X} - \mathbf{F}\mathbf{A}'\|^2 + (1 - \alpha) \|\mathbf{X}\mathbf{A}\mathbf{A}' - \mathbf{U}\mathbf{Y}\mathbf{A}'\|^2 \\ &= \alpha \|\mathbf{X} - \mathbf{F}\mathbf{A}'\|^2 + (1 - \alpha) \|\mathbf{X}\mathbf{A} - \mathbf{U}\mathbf{Y}\|^2, \end{aligned} \quad (5)$$

subject to $\mathbf{A}'\mathbf{A} = \mathbf{I}$, where $\mathbf{F}$ denotes an $I \times Q$ matrix of component scores, and we used that $\|\mathbf{X}\mathbf{A}\mathbf{A}' - \mathbf{U}\mathbf{Y}\mathbf{A}'\|^2 = \|\mathbf{X}\mathbf{A} - \mathbf{U}\mathbf{Y}\|^2$. The first term of (5) is the loss function minimized by PCA, and the second term by FKM. Thus, minimizing this loss function, we obtain a compromise between PCA and FKM. Indeed, setting $\alpha = 1$, we obtain PCA, setting $\alpha = 0$ we obtain FKM, and for values of α between 0 and 1 we compromise between PCA and FKM with varying emphasis on PCA or FKM. Interestingly, setting $\alpha = 1/2$ boils down to PPC/GROUPALS/RKM. This can be seen as follows. First, it should be remarked that the optimal $\mathbf{F}$ in (5) equals $\mathbf{X}\mathbf{A}$, as is readily verified. Then, using $\alpha = 1/2$ and substituting $\mathbf{X}\mathbf{A}$ for $\mathbf{F}$, we obtain:

$$\begin{aligned} f(\mathbf{A}, \mathbf{U}, \mathbf{Y}) &= 1/2 \left(\|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}'\|^2 + \|\mathbf{X}\mathbf{A}\mathbf{A}' - \mathbf{U}\mathbf{Y}\mathbf{A}'\|^2 \right) \\ &= 1/2 \left(\|\mathbf{X}\|^2 - 2\text{tr}\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} + \text{tr}\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} + \text{tr}\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} - 2\text{tr}\mathbf{X}'\mathbf{U}\mathbf{Y}\mathbf{A}' + \|\mathbf{U}\mathbf{Y}\mathbf{A}'\|^2 \right) \\ &= 1/2 \left(\|\mathbf{X} - \mathbf{U}\mathbf{Y}\mathbf{A}'\|^2 \right) = 1/2 f_{\text{PPC/GROUPALS/RKM}}. \end{aligned} \quad (6)$$

Thus, CDR can be seen to cover a whole range of methods from PCA via PPC/GROUPALS/RKM to FKM, by varying α from 1 to 0. Summarized, we see that CDR for:

$\alpha = 0$ equals factorial K -means

$\alpha = 1/2$ equals PPC/GROUPALS/RKM

$\alpha = 1$ equals principal component analysis.

As mentioned above, this approach differs from that by Vichi et al. (2007), as in the latter a compromise is defined between FKM and PPC/GROUPALS/RKM, whereas we define a compromise between FKM and PCA.

To minimize the loss function (5), we can actually resort to a minor modification of the algorithm for factorial K -means (Vichi and Kiers 2001). Again, we use that the optimal $\mathbf{F}$ that is given by $\mathbf{XA}$, and substituting $\mathbf{XA}$ for $\mathbf{F}$, we obtain:

$$f(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \alpha \|\mathbf{X} - \mathbf{XAA}\|^2 + (1 - \alpha) \|\mathbf{XA} - \mathbf{UY}\|^2. \quad (7)$$

So an alternating least-squares algorithm for minimizing (5) is obtained by alternatingly updating $\mathbf{A}$, $\mathbf{U}$ and $\mathbf{Y}$. To update $\mathbf{U}$, we have to minimize the second term, that is the factorial K -means loss function, hence we can use the factorial K -means step for this (see Vichi and Kiers 2001, p. 57). Next, we update $\mathbf{Y}$ and $\mathbf{A}$ jointly, using that, the optimal $\mathbf{Y}$ is $\mathbf{Y} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}$ (as in factorial K -means, using that also, $\mathbf{Y}$ occurs only in the second term). Then to update $\mathbf{A}$, we have to minimize:

$$\begin{aligned} f(\mathbf{A}, \mathbf{U}) &= \alpha \|\mathbf{X}\|^2 - \alpha \text{tr} \mathbf{A}' \mathbf{X}' \mathbf{XA} + (1 - \alpha) \|\mathbf{XA} - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1} \mathbf{U}' \mathbf{XA}\|^2 \\ &= \alpha \|\mathbf{X}\|^2 - \alpha \text{tr} \mathbf{A}' \mathbf{X}' \mathbf{XA} + (1 - \alpha) \text{tr} \mathbf{A}' \mathbf{X}' \mathbf{XA} - (1 - \alpha) \text{tr} \mathbf{A}' \mathbf{X}' \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1} \mathbf{U}' \mathbf{XA} \\ &= \alpha \|\mathbf{X}\|^2 + (1 - 2\alpha) \text{tr} \mathbf{A}' \mathbf{X}' \mathbf{XA} - (1 - \alpha) \text{tr} \mathbf{A}' \mathbf{X}' \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1} \mathbf{U}' \mathbf{XA} \\ &= \alpha \|\mathbf{X}\|^2 + \text{tr} \mathbf{A}' [(1 - 2\alpha) \mathbf{X}' \mathbf{X} - (1 - \alpha) \mathbf{X}' \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1} \mathbf{U}' \mathbf{X}] \mathbf{A} \\ &= \alpha \|\mathbf{X}\|^2 + \text{tr} \mathbf{A}' \mathbf{X}' [(1 - 2\alpha) \mathbf{I}_J - (1 - \alpha) \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1} \mathbf{U}'] \mathbf{XA}, \end{aligned} \quad (8)$$

where $\mathbf{I}_J$ is the identity matrix of size J . Thus, to minimize (8), we must take $\mathbf{A}$ equal to the matrix with the first Q eigenvectors¹ of the matrix $\mathbf{X}'((1 - \alpha)\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' - (1 - 2\alpha)\mathbf{I}_J)\mathbf{X}$, which is a simple variant of the updating step in factorial K -means. Repeating this process will monotonically decrease the loss function until convergence. The solution cannot be guaranteed to be a global optimum, so a multistart procedure is recommended to be used, and the best solution will be retained.

3 Handling nominal and ordinal variables along with quantitative variables in the context of PCA

3.1 Handling nominal variables along with quantitative variables in PCA and PCAMIX

CATPCA (Linting et al. 2007; Meulman et al. 2009, see also, but by the name of PRINCALS, Gifi 1990) is a variant of PCA that allows for variables of nominal or ordinal measurement level in addition to quantitative variables. There are two ways to handle

¹ Note that here and elsewhere in the paper, it is assumed that eigenvalues are ordered from high to low, and the eigenvectors are given in the order of the eigenvalues associated with them.

nominal variables: using multiple quantifications and using single quantifications. The latter can be seen as a special case of handling ordinal variables, where the order constraint is dropped. Here we focus on the multiple quantifications. The use of multiple quantifications seems more popular, and actually, when used in CATPCA gives a method equivalent to multiple correspondence analysis. In case there are nominal and quantitative variables, we end up with the method dubbed PCAMIX by Kiers (1991), originally proposed independently by several authors (de Leeuw 1973; de Leeuw and van Rijckeversel 1980; Escofier 1979; Nishisato 1980, pp. 103–107) and although the suggestions differ slightly in the way in which quantitative variables are transformed, essentially the same approach is used to handle qualitative variables. PCAMIX can most easily be described as follows. Without loss of generality, the first P variables are supposed to be categorical and coded by indicator matrices $\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_P$. The remaining $J - P$ variables are supposed to be quantitative, standardized (with zero mean and unit variance) and collected in the $I \times (J - P)$ data matrix $\mathbf{Z}$ with centered and normalized scores (such that $\mathbf{Z}$ has unit sums of squares columnwise). Now PCAMIX boils down to applying PCA to the following ‘data set’:

$$\mathbf{X} = [\mathbf{JG}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{JG}_P\mathbf{D}_P^{-1/2}, \mathbf{Z}], \quad (9)$$

where $\mathbf{J} = (\mathbf{I} - \mathbf{1}\mathbf{1}'/I)$ denotes the centering operator, the symbol $\mathbf{1}$ denotes a vector of unit elements of appropriate size, and $\mathbf{D}_j^{1/2} = \mathbf{G}_j' \mathbf{G}_j$ is the diagonal matrix with category frequencies on its diagonal. Specifically, PCAMIX minimizes:

$$f(\mathbf{F}, \mathbf{A}) = ||[\mathbf{JG}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{JG}_P\mathbf{D}_P^{-1/2}, \mathbf{Z}] - \mathbf{FA}'||^2, \quad (10)$$

subject to $\mathbf{A}'\mathbf{A} = \mathbf{I}$. As in ordinary PCA, the solution for $\mathbf{F}$ is given by the first Q eigenvectors of the matrix $\mathbf{XX}'$, normalized to sums of squares equal to the associated eigenvalues, which in this case can be verified to equal:

$$\mathbf{XX}' = \sum_j \mathbf{JG}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{J} + \mathbf{ZZ}'. \quad (11)$$

From (9) it becomes clear that, in case of only quantitative variables, the solution boils down to the first Q eigenvectors of $\mathbf{ZZ}'$, as in PCA of $\mathbf{Z}$. In case of only categorical variables, the solution boils down to the eigenvectors of $\sum_j \mathbf{JG}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{J}$, and hence gives the object scores in a MCA of the categorical variables (see Gifi 1990, p. 177).

3.2 Handling ordinal variables along with quantitative variables in PCA; CATPCA

In CATPCA, handling ordinal variables is easily arranged by “optimal scaling” (see Gifi 1990, p. 58 ff). That is, the PCA loss function is minimized, but rather than specifying scores for the ordinal variables, the scores for the ordinal variables themselves are estimated as well, under the restrictions that they should be monotonic transformations of the actual data. That is, here we propose to handle ordinal variables

as quantitative variables for which scale values are to be estimated optimally, given the model. In other words, we propose to use optimal scaling for ordinal variables, which in the present situation comes down to minimizing the PCA loss function over optimal scalings for ordinal variables. Specifically, let $\mathbf{q}_j = \mathbf{G}_j \mathbf{s}_j$, ($j = 1, \dots, T$) denote the scaled scores of T ordinal variables, then we minimize:

$$f(\mathbf{A}, \mathbf{F}, \mathbf{s}_1, \dots, \mathbf{s}_T) = \left\| [\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{Z}] - \mathbf{F}\mathbf{A}' \right\|^2, \quad (12)$$

over $\mathbf{A}$ subject to the constraint $\mathbf{A}'\mathbf{A} = \mathbf{I}$, $\mathbf{F}$, and $\mathbf{s}_1, \dots, \mathbf{s}_T$, and in addition subject to the constraints that the elements of $\mathbf{s}_j$ are weakly ordered from low to high, that $\|\mathbf{G}_j \mathbf{s}_j\|^2 = \mathbf{s}_j' \mathbf{D}_j \mathbf{s}_j = 1$ and that $\mathbf{1}' \mathbf{G}_j \mathbf{s}_j = 0$ ($j = 1, \dots, T$), where the first constraint ensures that scale values respect the original rank order in the scores, and the second and third constraints are common in optimal scaling, to avoid trivial solutions. The loss function in (12) is the PRINCIPALS loss function, see Gifi (p. 162).

4 The general model: handling nominal, ordinal, and quantitative variables in CDR

4.1 General model and algorithm

In Sect. 2, CDR has been described as a method for quantitative variables. Using the approaches described above for handling nominal variables (as in PCAMIX) and ordinal variables (as in CATPCA), we are now in a position to describe how CDR can handle all these types of variables together. Specifically, we propose here to minimize:

$$f(\mathbf{A}, \mathbf{U}, \mathbf{Y}, \mathbf{F}, \mathbf{s}_1, \dots, \mathbf{s}_T) = (1 - \alpha) \left\| [\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{J}\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_p \mathbf{D}_p^{-1/2}, \mathbf{Z}] \mathbf{A} - \mathbf{U}\mathbf{Y} \right\|^2 + \alpha \left\| [\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{J}\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_p \mathbf{D}_p^{-1/2}, \mathbf{Z}] - \mathbf{F}\mathbf{A}' \right\|^2, \quad (13)$$

over $\mathbf{A}$, $\mathbf{U}$, $\mathbf{Y}$, $\mathbf{F}$, and $\mathbf{s}_1, \dots, \mathbf{s}_T$ subject to the same constraints as mentioned for (12). Substituting $[\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{J}\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_p \mathbf{D}_p^{-1/2}, \mathbf{Z}] \mathbf{A}$ for the optimal value of $\mathbf{F}$, we get,

$$f(\mathbf{A}, \mathbf{U}, \mathbf{Y}, \mathbf{s}_1, \dots, \mathbf{s}_T) = (1 - \alpha) \left\| [\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{J}\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_p \mathbf{D}_p^{-1/2}, \mathbf{Z}] \mathbf{A} - \mathbf{U}\mathbf{Y} \right\|^2 + \alpha \left\| [\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{J}\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_p \mathbf{D}_p^{-1/2}, \mathbf{Z}] - [\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{J}\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_p \mathbf{D}_p^{-1/2}, \mathbf{Z}] \mathbf{A} \mathbf{A}' \right\|^2. \quad (14)$$

Now we extend the CDR ALS algorithm, consisting of alternately updating $\mathbf{U}$, $\mathbf{Y}$ and $\mathbf{A}$ (see Sect. 2), by adding, after these steps, a fourth step of successively updating the columns $\mathbf{q}_j = \mathbf{G}_j \mathbf{s}_j$, $j = 1, \dots, T$.

To describe how to update $\mathbf{q}_j = \mathbf{G}_j \mathbf{s}_j$, given $\mathbf{U}, \mathbf{Y}, \mathbf{A}$, and $\mathbf{q}_l = \mathbf{G}_l \mathbf{s}_l, l = 1, \dots, T, l \neq j$, it is convenient to write $\mathbf{Z}_{-j}$ for the matrix $[\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{JG}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{JG}_T \mathbf{D}_T^{-1/2}, \mathbf{Z}]$ from which the j^{th} column is left out, and $\mathbf{A}_{-j}$ for the matrix $\mathbf{A}$ from which the j^{th} row ($\mathbf{a}_j'$) is left out. Then, the problem is to minimize:

$$\begin{aligned}
 g(\mathbf{s}_j) &= (1 - \alpha) \|\mathbf{q}_j \mathbf{a}_j' + \mathbf{Z}_{-j} \mathbf{A}_{-j} - \mathbf{U} \mathbf{Y}\|^2 \\
 &\quad + \alpha \|[\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{JG}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{JG}_T \mathbf{D}_T^{-1/2} \mathbf{Z}] - (\mathbf{q}_j \mathbf{a}_j' + \mathbf{Z}_{-j} \mathbf{A}_{-j}) \mathbf{A}'\|^2 \\
 &= (1 - \alpha) \|\mathbf{G}_j \mathbf{s}_j \mathbf{a}_j' + \mathbf{Z}_{-j} \mathbf{A}_{-j} - \mathbf{U} \mathbf{Y}\|^2 + \alpha \|\mathbf{G}_j \mathbf{s}_j - (\mathbf{G}_j \mathbf{s}_j \mathbf{a}_j' + \mathbf{Z}_{-j} \mathbf{A}_{-j}) \mathbf{a}_j\|^2 \\
 &\quad + \alpha \|\mathbf{Z}_{-j} - (\mathbf{G}_j \mathbf{s}_j \mathbf{a}_j' + \mathbf{Z}_{-j} \mathbf{A}_{-j}) \mathbf{A}_{-j}'\|^2 \\
 &= (1 - \alpha) \|\mathbf{G}_j \mathbf{s}_j \mathbf{a}_j' + \mathbf{Z}_{-j} \mathbf{A}_{-j} - \mathbf{U} \mathbf{Y}\|^2 + \alpha \|(1 - \mathbf{a}_j' \mathbf{a}_j) \mathbf{G}_j \mathbf{s}_j - \mathbf{Z}_{-j} \mathbf{A}_{-j} \mathbf{a}_j\|^2 \\
 &\quad + \alpha \|\mathbf{Z}_{-j} - \mathbf{Z}_{-j} \mathbf{A}_{-j} \mathbf{A}_{-j}' - \mathbf{G}_j \mathbf{s}_j \mathbf{a}_j' \mathbf{A}_{-j}'\|^2 \\
 &= (1 - \alpha) (\mathbf{a}_j' \mathbf{a}_j) \mathbf{s}_j' \mathbf{D}_j \mathbf{s}_j - 2(1 - \alpha) \mathbf{s}_j' \mathbf{G}_j' (\mathbf{U} \mathbf{Y} - \mathbf{Z}_{-j} \mathbf{A}_{-j}) \mathbf{a}_j \\
 &\quad + \alpha (1 - \mathbf{a}_j' \mathbf{a}_j)^2 \mathbf{s}_j' \mathbf{D}_j \mathbf{s}_j - 2\alpha (1 - \mathbf{a}_j' \mathbf{a}_j) \mathbf{s}_j' \mathbf{G}_j' \mathbf{Z}_{-j} \mathbf{A}_{-j} \mathbf{a}_j \\
 &\quad + \alpha (\mathbf{a}_j' \mathbf{A}_{-j}' \mathbf{A}_{-j} \mathbf{a}_j) \mathbf{s}_j' \mathbf{D}_j \mathbf{s}_j - 2\alpha \mathbf{s}_j' \mathbf{G}_j' (\mathbf{Z}_{-j} - \mathbf{Z}_{-j} \mathbf{A}_{-j} \mathbf{A}_{-j}') \mathbf{A}_{-j} \mathbf{a}_j + c^* \\
 &= ((1 - \alpha) (\mathbf{a}_j' \mathbf{a}_j) + \alpha (1 - \mathbf{a}_j' \mathbf{a}_j)^2 + \alpha (\mathbf{a}_j' \mathbf{A}_{-j}' \mathbf{A}_{-j} \mathbf{a}_j)) \mathbf{s}_j' \mathbf{D}_j \mathbf{s}_j \\
 &\quad - 2\mathbf{s}_j' \mathbf{G}_j' ((1 - \alpha) (\mathbf{U} \mathbf{Y} - \mathbf{Z}_{-j} \mathbf{A}_{-j}) \mathbf{a}_j + \alpha (1 - \mathbf{a}_j' \mathbf{a}_j) \mathbf{Z}_{-j} \mathbf{A}_{-j} \mathbf{a}_j \\
 &\quad + \alpha (\mathbf{Z}_{-j} - \mathbf{Z}_{-j} \mathbf{A}_{-j} \mathbf{A}_{-j}') \mathbf{A}_{-j} \mathbf{a}_j) + c^*,
 \end{aligned} \tag{15}$$

where c^* denotes a constant with respect to $\mathbf{s}_j$. Now upon defining:

$$\alpha^* \equiv (1 - \alpha) (\mathbf{a}_j' \mathbf{a}_j) + \alpha (1 - \mathbf{a}_j' \mathbf{a}_j)^2 + \alpha (\mathbf{a}_j' \mathbf{A}_{-j}' \mathbf{A}_{-j} \mathbf{a}_j), \tag{16}$$

and

$$\mathbf{s}^* \equiv (1 - \alpha) (\mathbf{U} \mathbf{Y} - \mathbf{Z}_{-j} \mathbf{A}_{-j}) \mathbf{a}_j + \alpha (1 - \mathbf{a}_j' \mathbf{a}_j) \mathbf{Z}_{-j} \mathbf{A}_{-j} \mathbf{a}_j + \alpha (\mathbf{Z}_{-j} - \mathbf{Z}_{-j} \mathbf{A}_{-j} \mathbf{A}_{-j}') \mathbf{A}_{-j} \mathbf{a}_j, \tag{17}$$

we get

$$\begin{aligned}
 g(\mathbf{s}_j) &= \alpha^* (\mathbf{s}_j' \mathbf{D}_j \mathbf{s}_j - 2(\alpha^*)^{-1} \mathbf{s}_j' \mathbf{G}_j' \mathbf{s}^*) + c^* \\
 &= \alpha^* \|\mathbf{D}_j^{-1/2} (\mathbf{s}_j - (\alpha^*)^{-1} \mathbf{D}_j^{-1} \mathbf{G}_j' \mathbf{s}^*)\|^2 + c^{**},
 \end{aligned} \tag{18}$$

where c^{**} denotes a constant with respect to $\mathbf{s}_j$. It follows that, to minimize $g(\mathbf{s}_j)$ subject to the order constraint and the constraints $\mathbf{s}_j' \mathbf{D}_j \mathbf{s}_j = 1$ and $\mathbf{1}' \mathbf{G}_j \mathbf{s}_j = 0$, it suffices to minimize:

$$(\mathbf{s}_j - (\alpha^*)^{-1} \mathbf{D}_j^{-1/2} \mathbf{G}_j' \mathbf{s}^*) \mathbf{D}_j (\mathbf{s}_j - (\alpha^*)^{-1} \mathbf{D}_j^{-1/2} \mathbf{G}_j' \mathbf{s}^*), \tag{19}$$

subject to these constraints. This is a standard weighted monotone regression problem, see Gifi (1990, Appendix C, p. 529).

The extended ALS algorithm for CDR including ordinal variables thus consists of, after initializing all parameter matrices, alternately updating $\mathbf{A}$, $\mathbf{U}$, $\mathbf{Y}$, and $\mathbf{s}_1 \dots \mathbf{s}_T$. As each update decreases the loss function, which is bounded below by zero, this process will converge to a stable function value, and when function values after successive full iterations differ less than a predefined value, we consider the algorithm converged. In principle, this does not guarantee that the parameters have then converged as well, but in practice this seems to be the case nevertheless.

4.2 Problematic cases: $\alpha = 0$

When dealing with at least some categorical variables, the method with $\alpha = 0$ (i.e., factorial K -means), will lead to a degenerate solution when the number of dimensions equals the number of categories of a variable ($Q = c_j$ for a variable j). To see this, first consider the case where the number of clusters (K), equals Q and c_j , hence $K = c_j = Q$. Then, the minimum of the loss function can always be found as follows. To minimize:

$$f(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = ||[\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{JG}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{JG}_p\mathbf{D}_p^{-1/2}, \mathbf{Z}]\mathbf{A} - \mathbf{UY}||^2, \quad (20)$$

we can take all submatrices of $\mathbf{A}$ equal to $\mathbf{0}$, except the block associated with categorical variable j , which is set equal to $\mathbf{I}$. Then taking $\mathbf{U}$ equal to $\mathbf{G}_j$, and $\mathbf{Y}$ equal to $(\mathbf{I} - \mathbf{11}'\mathbf{D}_j/I)\mathbf{D}_j^{-1/2}$, and using that $\mathbf{G}_j(\mathbf{I} - \mathbf{11}'\mathbf{D}_j/I) = (\mathbf{G}_j - \mathbf{G}_j\mathbf{11}'\mathbf{D}_j/I) = (\mathbf{G}_j - \mathbf{11}'\mathbf{D}_j/I) = (\mathbf{G}_j - \mathbf{11}'\mathbf{G}_j/I) = \mathbf{JG}_j$ we get:

$$f(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = ||\mathbf{JG}_j\mathbf{D}_j^{-1/2} - \mathbf{G}_j(\mathbf{I} - \mathbf{11}'\mathbf{D}_j/I)\mathbf{D}_j^{-1/2}||^2 = ||\mathbf{JG}_j\mathbf{D}_j^{-1/2} - \mathbf{JG}_j\mathbf{D}_j^{-1/2}||^2 = 0, \quad (21)$$

hence this solution finds the minimum of the loss function.

In case the number of clusters is bigger than the number of categories of variable j ($K > c_j = Q$), a similarly trivial solution can be obtained by starting out from the above $K = c_j = Q$ solution (i.e., $\mathbf{U} = \mathbf{G}_j$ and $\mathbf{Y} = (\mathbf{I} - \mathbf{11}'\mathbf{D}_j/I)\mathbf{D}_j^{-1/2}$), and then arbitrarily splitting one or more categories of the variable, and calling the associated indicator matrix $\mathbf{U}^*$. Then $\mathbf{U} = \mathbf{U}^*\mathbf{P}$, where $\mathbf{P}$ is a matrix with ones and zeros that indicates which clusters of $\mathbf{U}^*$ combined give the clusters of $\mathbf{U}$, e.g., to split cluster 1 in two clusters, we would need $\mathbf{P} = [1 \ 0; 1 \ 0; 0 \ 1]$. Now taking $\mathbf{Y}^* = \mathbf{PY}$, we obtain $\mathbf{U}^*\mathbf{Y}^* = \mathbf{UY} = \mathbf{JG}_j\mathbf{D}_j^{-1/2}$, while the number of clusters in $\mathbf{U}^*$ is bigger than the number of categories in variable j .

In case the number of clusters is smaller than the number of categories of variable j ($K < c_j = Q$), we again start out from the $K = c_j = Q$ solution (i.e., $\mathbf{U} = \mathbf{G}_j$ and $\mathbf{Y} = (\mathbf{I} - \mathbf{11}'\mathbf{D}_j/I)\mathbf{D}_j^{-1/2}$), but now arbitrarily combine some clusters in $\mathbf{U}$, so we take $\mathbf{U}^* = \mathbf{UP}$ for some combining matrix $\mathbf{P}$, and define $\mathbf{D}_{\mathbf{U}^*} = \mathbf{U}^*\mathbf{U}^*$. Now taking

$\mathbf{U}^* = \mathbf{U}\mathbf{P}$, $\mathbf{Y}^* = (\mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D}_U^*/I)\mathbf{D}_U^{*-1/2}\mathbf{T}$, $\mathbf{A}^* = \mathbf{D}_j^{1/2}\mathbf{P}\mathbf{D}_U^{*-1/2}\mathbf{T}$ for the block of $\mathbf{A}$ related to variable j , where $\mathbf{T}$ is chosen such that $\mathbf{A}^*\mathbf{A}^* = \mathbf{I}$, again gives a loss of 0:

$$\begin{aligned} & \|\mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1/2}\mathbf{D}_j^{1/2}\mathbf{P}(\mathbf{D}_U^*)^{-1/2}\mathbf{T} - \mathbf{U}^*(\mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D}_U^*/I)(\mathbf{D}_U^*)^{-1/2}\mathbf{T}\|^2 \\ &= \|\mathbf{J}\mathbf{U}\mathbf{P}(\mathbf{D}_U^*)^{-1/2}\mathbf{T} - \mathbf{U}^*(\mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D}_U^*/I)(\mathbf{D}_U^*)^{-1/2}\mathbf{T}\|^2 \\ &= \|\mathbf{J}\mathbf{U}^*(\mathbf{D}_U^*)^{-1/2}\mathbf{T} - \mathbf{J}\mathbf{U}^*(\mathbf{D}_U^*)^{-1/2}\mathbf{T}\|^2 = 0 \end{aligned}$$

Thus, when $Q = c_j$, for some variable j , always a perfectly fitting solution can be found. Trivially, from these solutions we can also immediately find a solution for $Q < c_j$, since all we have to do then is to take a subset of the columns of $\mathbf{A}$ and $\mathbf{Y}$ and still find a loss function of 0.

Such solutions, although perfectly satisfying the objective of factorial K -means (finding partitions anywhere in the space), from a substantive point of view are not at all interesting, because they pertain to only one categorical variable. Thus, the partition obtained by CDR with $\alpha = 0$ provides nothing more than information that was immediately available, without any analysis. It does not give partitions that to some extent summarize the information in the different variables. Indeed, this situation illustrates a fundamental characteristic of factorial K -means, which searches any dimension in high dimensional space along which a (near) perfect clustering exists. With categorical variables, perfect partitions by definition exist somewhere in space, and will be found, even when they relate to only one single variable. Factorial K -means by no means ensures that dimensions are found that also summarize the variables well, and the case in point demonstrates this clearly.

In practice, one will quite often have categorical variables with at least as many categories as the number of dimensions (Q) one is interested in, so the above problem would arise quite often in practice, if one would analyse (partially) categorical data by CDR with $\alpha = 0$. For the situation where $Q > c_j$, no such trivial perfect solutions can be expected to be found. However, suppose $c_j = 3$ and $Q = 4$, and suppose there is at least one other categorical variable. Then $\mathbf{A}$ must have four columns. When in $\mathbf{A}$ the block corresponding to categorical variable j is taken equal to $\mathbf{I}_3$ and in the fourth column, the vector associated with a different categorical variable (say the k th) is taken equal to $\mathbf{D}_k^{1/2}\mathbf{1}(\mathbf{1}'\mathbf{D}_k\mathbf{1})^{-1/2}$, then $\mathbf{A}'\mathbf{A} = \mathbf{I}$, and $\mathbf{X}\mathbf{A} = \mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1/2}$ (because the fourth column in $\mathbf{X}\mathbf{A}$ equals $\mathbf{J}\mathbf{G}_k\mathbf{D}_k^{-1/2}\mathbf{D}_k^{1/2}\mathbf{1}(\mathbf{1}'\mathbf{D}_k\mathbf{1})^{-1/2} = \mathbf{J}\mathbf{G}_k\mathbf{1}(\mathbf{1}'\mathbf{D}_k\mathbf{1})^{-1/2} = \mathbf{J}\mathbf{1}(\mathbf{1}'\mathbf{D}_k\mathbf{1})^{-1/2} = \mathbf{0}$, so again a perfect solution can be found (with the last column of $\mathbf{Y}$ filled with zeros). So it appears that, in the presence of categorical variables, only rarely nontrivial solutions will be found when $\alpha = 0$. Therefore, we avoid taking $\alpha = 0$ with CDR applied to (among others) categorical variables.

4.3 Assessing α , the number of clusters K and the number of dimensions Q

The choice of α can be made on methodological grounds, for instance, by deciding a priori that the desired method is to be right in the middle of factorial K -means and PPC/GROUPALS/RKM (i.e., choosing $\alpha = 0.25$), or for instance, right in the middle of PCA and PPC/GROUPALS/RKM (i.e., choosing $\alpha = 0.75$). Alternatively, several

choices of α could be evaluated and the most attractive of these (e.g., the most stable or best interpretable one) could be chosen.

To select the model in terms of a number of clusters, dimensions and α , we propose to use a generalization of the pseudo F index (pF) (Calinski and Harabasz 1974, also proposed by Timmerman et al. (2010) as one possible way of handling this problem), which measures the quality of a partition in the reduced space taking into account the model complexity:

$$pF = \frac{\text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A})/d_b}{(\text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{A}'\mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}))/d_w}, \quad (22)$$

where $d_b = (K - 1)Q + (J - Q)Q$ and $d_w = IJ - KQ - (J - Q)Q$ are the degrees of freedom. Here J is the total number of variables, where each categorical variables counts as its number of categories minus 1. Hence, in cases involving categorical variables, J equals the number of columns of the supermatrix $[\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{J}\mathbf{G}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_P\mathbf{D}_P^{-1/2}, \mathbf{Z}]$ minus the number of categorical variables. The pF in fact gives the ratio of between cluster variance of component scores (in $\mathbf{X}\mathbf{A}$) to the within cluster variance, both corrected for their associated degrees of freedom. We propose to select that model that maximizes this ratio. We use this model selection approach in the simulation study as well as in the example analyses in the next section.

5 Equivalence of GROUPALS and CDR with $\alpha = 1/2$ applied to mixed variables

When $\alpha = 1/2$, CDR boils down to minimizing loss function (2)

$$f(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{U}\mathbf{Y}\mathbf{A}'\|^2 \quad (23)$$

over $\mathbf{A}$, $\mathbf{U}$, $\mathbf{Y}$, and $\mathbf{s}_1, \dots, \mathbf{s}_T$, where $\mathbf{X} = [\mathbf{q}_1, \dots, \mathbf{q}_T, \mathbf{J}\mathbf{G}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_P\mathbf{D}_P^{-1/2}, \mathbf{Z}]$ and $\mathbf{q}_j = \mathbf{G}\mathbf{s}_j$, subject to the constraints $\mathbf{A}'\mathbf{A} = \mathbf{I}$ and the monotonicity constraints that the elements of $\mathbf{s}_j$ are weakly ordered from low to high, $\|\mathbf{G}_j\mathbf{s}_j\|^2 = 1$, and $\mathbf{1}'\mathbf{G}_j\mathbf{s}_j = 0, j = 1, \dots, T$. We will now show that this method gives the same solutions as GROUPALS (Van Buuren and Heiser 1989) which minimizes

$$g(\mathbf{G}_c, \mathbf{Y}, \mathbf{Y}_1, \dots, \mathbf{Y}_P) = \sum_j \|\mathbf{G}_j\mathbf{Y}_j - \mathbf{G}_c\mathbf{Y}\|^2 \quad (24)$$

subject to the constraint that $\mathbf{Y}'\mathbf{D}_c\mathbf{Y} = \mathbf{I}$ and $\mathbf{1}'\mathbf{G}_c\mathbf{Y} = \mathbf{0}$, where $\mathbf{D}_c = \mathbf{G}_c'\mathbf{G}_c$. Here $\mathbf{Y}_j$ is left unconstrained in case of categorical variables, while in case of numerical or ordinal variables, $\mathbf{Y}_j = \mathbf{y}_j\mathbf{a}_j'$. In case $\mathbf{Y}_j = \mathbf{y}_j\mathbf{a}_j'$, for ordinal variables, the monotonicity constraint is imposed on $\mathbf{y}_j$, see above, while for numerical variables, $\mathbf{y}_j$ is constrained to have values equal to (a positive linear transformation of) the original

scores. It will be shown that the solutions for $\mathbf{U}$ and $\mathbf{Y}$ minimizing (23) equal those for $\mathbf{G}_c$ and $\mathbf{Y}$ minimizing (24).

First, we consider the case with only categorical variables. Then we can reformulate the GROUPALS problem as follows. The optimal $\mathbf{Y}_j$ is given by $\mathbf{Y}_j = \mathbf{D}_j^{-1} \mathbf{G}'_j \mathbf{G}_c \mathbf{Y}$. Substituting this expression in (24) and elaborating gives:

$$\begin{aligned} g(\mathbf{G}_c, \mathbf{Y}) &= \sum_j \left\| \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j \mathbf{G}_c \mathbf{Y} - \mathbf{G}_c \mathbf{Y} \right\|^2 \\ &= \sum_j \left\| \left(\mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j - \mathbf{I} \right) \mathbf{G}_c \mathbf{Y} \right\|^2 \\ &= \sum_j \text{tr} \mathbf{Y}' \mathbf{G}'_c \left(\mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j - \mathbf{I} \right)^2 \mathbf{G}_c \mathbf{Y} \\ &= \text{tr} \mathbf{Y}' \mathbf{G}'_c \sum_j \left(\mathbf{I} - \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j \right) \mathbf{G}_c \mathbf{Y} \end{aligned} \quad (25)$$

where in the last step it is used that $(\mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j - \mathbf{I})^2 = \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j - \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j - \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j + \mathbf{I} = (\mathbf{I} - \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j)$. It follows that minimizing (24) is equivalent to maximizing

$$g^*(\mathbf{G}_c, \mathbf{Y}) = \text{tr} \mathbf{Y}' \mathbf{G}'_c \left(\sum_j \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j \right) \mathbf{G}_c \mathbf{Y} \quad (26)$$

subject to the constraints mentioned above.

Now we will show that maximizing (26) is equivalent to applying the CDR approach with $\alpha = 1/2$, to $\mathbf{X} = [\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{G}_p \mathbf{D}_p^{-1/2}]$, that is, using the *uncentered* version of the indicator matrices. Also we use that, minimizing $f(\mathbf{A}, \mathbf{U}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{U} \mathbf{Y} \mathbf{A}'\|^2$ subject to $\mathbf{A}' \mathbf{A} = \mathbf{I}$ is equivalent to minimizing this loss function subject to $\mathbf{Y}' \mathbf{D} \mathbf{Y} = \mathbf{I}$, where $\mathbf{D} = \mathbf{U}' \mathbf{U}$. This is so, because one can always multiply $\mathbf{A}$ by a nonsingular matrix $\mathbf{T}'$ and $\mathbf{Y}$ by $\mathbf{T}^{-1}$ (without loss of fit) such that $\mathbf{Y}^* = \mathbf{Y} \mathbf{T}^{-1}$ satisfies $\mathbf{Y}^{*'} \mathbf{D} \mathbf{Y}^* = \mathbf{I}$. This transition from a constraint on $\mathbf{A}$ to a constraint on $\mathbf{Y}$ is analogous to the transfer of normalization used by Van Buuren and Heiser (1989). Now first minimizing (23) over $\mathbf{A}'$ we obtain, using that $(\mathbf{U} \mathbf{Y})' \mathbf{U} \mathbf{Y} = \mathbf{Y}' \mathbf{U}' \mathbf{U} \mathbf{Y} = \mathbf{Y}' \mathbf{D} \mathbf{Y} = \mathbf{I}$, that $\mathbf{A}' = (\mathbf{U} \mathbf{Y})' \mathbf{X}$, and substituting this in (23) gives:

$$\begin{aligned} f(\mathbf{U}, \mathbf{Y}) &= \|\mathbf{X} - \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{X}\|^2 \\ &= \text{tr} \mathbf{X}' \mathbf{X} - 2 \text{tr} \mathbf{X}' \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{X} + \text{tr} \mathbf{X}' \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{X} \\ &= \text{tr} \mathbf{X}' \mathbf{X} - 2 \text{tr} \mathbf{X}' \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{X} + \text{tr} \mathbf{X}' \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{X} \\ &= \text{tr} \mathbf{X}' \mathbf{X} - \text{tr} \mathbf{Y}' \mathbf{U}' \mathbf{X} \mathbf{X}' \mathbf{U} \mathbf{Y}. \end{aligned} \quad (27)$$

Minimizing (27) is equivalent to maximizing

$$f^*(\mathbf{U}, \mathbf{Y}) = \text{tr} \mathbf{Y}' \mathbf{U}' \mathbf{X} \mathbf{X}' \mathbf{U} \mathbf{Y} \quad (28)$$

which upon substituting $\mathbf{X} = [\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{G}_p \mathbf{D}_p^{-1/2}]$ boils down to maximizing

$$f^*(\mathbf{U}, \mathbf{Y}) = \text{tr} \mathbf{Y}' \mathbf{U}' \left(\sum_j \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}_j' \right) \mathbf{U} \mathbf{Y}. \quad (29)$$

Clearly, (26) and (29) are the same functions of $\mathbf{G}_c$ and $\mathbf{Y}$, and $\mathbf{U}$ and $\mathbf{Y}$, respectively, so GROUPALS and CDR with $\alpha = 1/2$ applied to $\mathbf{X}$ give the same solution when the same constraints are imposed on $\mathbf{G}_c$ and $\mathbf{U}$, and on $\mathbf{Y}$ in both methods. In GROUPALS the constraints $(\mathbf{G}_c \mathbf{Y})' \mathbf{G}_c \mathbf{Y} = \mathbf{I}$ and $\mathbf{1}' \mathbf{G}_c \mathbf{Y} = \mathbf{0}$ are imposed, in CDR only the constraint $(\mathbf{U} \mathbf{Y})' \mathbf{U} \mathbf{Y} = \mathbf{I}$. However, imposing also the constraint $\mathbf{1}' \mathbf{U} \mathbf{Y} = \mathbf{0}$ in CDR has an interesting implication. To minimize $f(\mathbf{U}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{X}\|^2$ subject to the constraint $\mathbf{1}' \mathbf{U} \mathbf{Y} = \mathbf{0}$, we write $\mathbf{X} = (\mathbf{J} \mathbf{X} + \mathbf{1} \mathbf{1}' \mathbf{X} / I)$, and elaborate $f(\mathbf{U}, \mathbf{Y})$ as:

$$\begin{aligned} f(\mathbf{U}, \mathbf{Y}) &= \|\mathbf{X} - \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{X}\|^2 \\ &= \|(\mathbf{J} \mathbf{X} + \mathbf{1} \mathbf{1}' \mathbf{X} / I) - \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' (\mathbf{J} \mathbf{X} + \mathbf{1} \mathbf{1}' \mathbf{X} / I)\|^2 \\ &= \|\mathbf{J} \mathbf{X} - \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{J} \mathbf{X} + \mathbf{1} \mathbf{1}' \mathbf{X} / I\|^2 \\ &= \|\mathbf{J} \mathbf{X} - \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{J} \mathbf{X}\|^2 + \|\mathbf{1} \mathbf{1}' \mathbf{X} / I\|^2 \\ &= \text{constant} - 2 \text{tr} (\mathbf{J} \mathbf{X})' \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{J} \mathbf{X} + \text{tr} (\mathbf{J} \mathbf{X})' \mathbf{U} \mathbf{Y} (\mathbf{U} \mathbf{Y})' \mathbf{J} \mathbf{X}, \end{aligned} \quad (30)$$

where in the third and the forth step it is used that $\mathbf{1}' (\mathbf{U} \mathbf{Y}) = \mathbf{0}$. It follows that CDR with $\alpha = 1/2$ applied to $\mathbf{X} = [\mathbf{G}_1 \mathbf{D}_1^{-1/2}, \dots, \mathbf{G}_p \mathbf{D}_p^{-1/2}]$, subject to the additional constraint $\mathbf{1}' (\mathbf{U} \mathbf{Y}) = \mathbf{0}$, is equivalent to maximizing

$$f^{**}(\mathbf{U}, \mathbf{Y}) = \text{tr} \mathbf{Y}' \mathbf{U}' (\mathbf{J} \mathbf{X}) (\mathbf{J} \mathbf{X})' \mathbf{U} \mathbf{Y} = \text{tr} \mathbf{Y}' \mathbf{U}' \left(\sum_j \mathbf{J} \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}_j' \mathbf{J} \right) \mathbf{U} \mathbf{Y} \quad (31)$$

subject to $\mathbf{1}' (\mathbf{U} \mathbf{Y}) = \mathbf{0}$. This constraint, however, is inactive. That is, if we maximize (31) without imposing the constraint, it will be satisfied automatically, as can be seen as follows. When we define $\mathbf{Y}^* = \mathbf{D}^{1/2} \mathbf{Y}$, for any $\mathbf{U}$, the problem comes down to maximizing $\text{tr} \mathbf{Y}^{*'} (\mathbf{D}^{-1/2} \mathbf{U}' \sum_j \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}_j' \mathbf{U} \mathbf{D}^{-1/2}) \mathbf{Y}^*$ subject to the constraint $\mathbf{Y}^{*'} \mathbf{Y}^* = \mathbf{I}$. The solution for $\mathbf{Y}^*$ is given by the first Q eigenvectors of $(\mathbf{D}^{-1/2} \mathbf{U}' \sum_j \mathbf{J} \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}_j' \mathbf{J} \mathbf{U} \mathbf{D}^{-1/2})$ so $\mathbf{Y}^*$ is in the subspace spanned by $(\mathbf{D}^{-1/2} \mathbf{U}' \sum_j \mathbf{J} \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}_j')$. Now the optimal $\mathbf{Y}$ is given by $\mathbf{D}^{-1/2} \mathbf{Y}^*$, which hence is in the subspace of $(\mathbf{D}^{-1} \mathbf{U}' \sum_j \mathbf{J} \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}_j')$. In other words, $\mathbf{Y} = (\mathbf{D}^{-1} \mathbf{U}' \sum_j \mathbf{J} \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}_j') \mathbf{T}$ for a particular matrix $\mathbf{T}$. To see that this solution satisfies the constraint, we simply elaborate $\mathbf{1}' \mathbf{U} \mathbf{Y}$ as follows:

$$\begin{aligned}
\mathbf{1}'\mathbf{U}\mathbf{Y} &= \mathbf{1}'\mathbf{U}\mathbf{D}^{-1}\mathbf{U}'\left(\sum_j \mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\right)\mathbf{T} \\
&= \mathbf{1}'\mathbf{D}\mathbf{D}^{-1}\mathbf{U}'\left(\sum_j \mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\right)\mathbf{T} \\
&= \mathbf{1}'\mathbf{U}'\left(\sum_j \mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\right)\mathbf{T} \\
&= \mathbf{1}'\left(\sum_j \mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\right)\mathbf{T} = \mathbf{0}.
\end{aligned} \tag{32}$$

Interestingly, maximizing $f^*(\mathbf{U}, \mathbf{Y}) = \text{tr } \mathbf{Y}'\mathbf{U}'(\sum_j \mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{J})\mathbf{U}\mathbf{Y}$ over $\mathbf{U}$ and $\mathbf{Y}$, subject to $(\mathbf{U}\mathbf{Y})'\mathbf{U}\mathbf{Y} = \mathbf{I}$, is equivalent to CDR with $\alpha = 1/2$ applied to $\mathbf{X} = [\mathbf{J}\mathbf{G}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_p\mathbf{D}_p^{-1/2}]$, as follows from substituting this expression for $\mathbf{X}$ in (28).

Now we have the following chain of equivalences: CDR with $\alpha = 1/2$ applied to $\mathbf{X} = [\mathbf{J}\mathbf{G}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{J}\mathbf{G}_p\mathbf{D}_p^{-1/2}]$, subject to $(\mathbf{U}\mathbf{Y})'\mathbf{U}\mathbf{Y} = \mathbf{I}$, is equivalent to CDR with $\alpha = 1/2$ applied to $\mathbf{X} = [\mathbf{G}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{G}_p\mathbf{D}_p^{-1/2}]$ subject to $(\mathbf{U}\mathbf{Y})'\mathbf{U}\mathbf{Y} = \mathbf{I}$ and $\mathbf{1}'(\mathbf{U}\mathbf{Y}) = \mathbf{0}$, which in turn is equivalent to maximizing g^* over $\mathbf{G}_c$ and $\mathbf{Y}$ subject to $(\mathbf{G}_c\mathbf{Y})'\mathbf{G}_c\mathbf{Y} = \mathbf{I}$ and $\mathbf{1}'(\mathbf{G}_c\mathbf{Y}) = \mathbf{0}$, which is equivalent to GROUPALS applied to only categorical variables.

Now consider the case with only ordinal or numerical variables. Then GROUPALS is defined as minimizing $g(\mathbf{G}_c, \mathbf{Y}, \mathbf{Y}_1, \dots, \mathbf{Y}_p) = \sum_j \|\mathbf{G}_j\mathbf{Y}_j - \mathbf{G}_c\mathbf{Y}\|^2$ with $\mathbf{Y}_j = \mathbf{y}_j\mathbf{a}_j'$. This function can be elaborated as:

$$g(\mathbf{G}_c, \mathbf{Y}, \mathbf{Y}_1, \dots, \mathbf{Y}_p) = \sum_j \|\mathbf{G}_j\mathbf{y}_j\mathbf{a}_j' - \mathbf{G}_c\mathbf{Y}\|^2 = \sum_j \|\mathbf{q}_j\mathbf{a}_j' - \mathbf{G}_c\mathbf{Y}\|^2, \tag{33}$$

where $\mathbf{q}_j = \mathbf{G}_j\mathbf{y}_j$. Upon using the GROUPALS identification constraint $\mathbf{q}_j'\mathbf{q}_j = \mathbf{y}_j'\mathbf{D}_j\mathbf{y}_j = 1$, we easily see that the values for $\mathbf{a}_j'$ minimizing g are given by $\mathbf{a}_j' = \mathbf{q}_j'\mathbf{G}_c\mathbf{Y}$. Substituting these into (33) gives:

$$\begin{aligned}
g(\mathbf{G}_c, \mathbf{Y}, \mathbf{q}_1, \dots, \mathbf{q}_p) &= \sum_j \|\mathbf{q}_j\mathbf{q}_j'\mathbf{G}_c\mathbf{Y} - \mathbf{G}_c\mathbf{Y}\|^2 \\
&= \sum_j \|(\mathbf{q}_j\mathbf{q}_j' - \mathbf{I})\mathbf{G}_c\mathbf{Y}\|^2 \\
&= \sum_j \text{tr} \mathbf{Y}'\mathbf{G}_c' \sum_j (\mathbf{q}_j\mathbf{q}_j' - \mathbf{I})^2 \mathbf{G}_c\mathbf{Y} \\
&= \sum_j \text{tr} \mathbf{Y}'\mathbf{G}_c' \sum_j (\mathbf{I} - \mathbf{q}_j\mathbf{q}_j') \mathbf{G}_c\mathbf{Y},
\end{aligned} \tag{34}$$

hence GROUPALS for ordinal or numerical variables is equivalent to maximizing:

$$g^*(\mathbf{G}_c, \mathbf{Y}, \mathbf{q}_1, \dots, \mathbf{q}_P) = \sum_j \text{tr} \mathbf{Y}' \mathbf{G}'_c \left(\sum_j \mathbf{q}_j \mathbf{q}'_j \right) \mathbf{G}_c \mathbf{Y}. \quad (35)$$

In case of ordinal variables, $\mathbf{q}_j = \mathbf{G}_j \mathbf{y}_j$, with monotonicity constraints on $\mathbf{y}_j$, while in case of numerical variables, $\mathbf{q}_j = \mathbf{z}_j$. This is equivalent to CDR with $\alpha = 1/2$, which comes down to maximizing $f^*(\mathbf{U}, \mathbf{Y}) = \text{tr} \mathbf{Y}' \mathbf{U}' \mathbf{X} \mathbf{X}' \mathbf{U} \mathbf{Y}$ (see (28)), with $\mathbf{X}$ containing columns of the form $\mathbf{q}_j = \mathbf{G}_j \mathbf{s}_j$ for ordinal variables, and $\mathbf{q}_j = \mathbf{z}_j$ for numerical variables, where we use that $\mathbf{X} \mathbf{X}' = \sum_j \mathbf{q}_j \mathbf{q}'_j$. For the ordinal case, we employ the constraints $\mathbf{1}' \mathbf{G}_j \mathbf{s}_j = 0$, while GROUPALS does not seem to do so (Van Buuren and Heiser (1989) do not specify this). However, as GROUPALS uses the constraint $\mathbf{1}' \mathbf{G}_c \mathbf{Y} = 0$, we presume that the monotone regression minimizing $\| \mathbf{G}_j \mathbf{y}_j \mathbf{a}'_j - \mathbf{G}_c \mathbf{Y} \|^2$ automatically ensures $\mathbf{1}' \mathbf{G}_j \mathbf{y}_j = 0$, for $j = 1, \dots, T$.

Finally, if $\mathbf{X}$ contains a mixture of categorical, ordinal and numerical variables, then from the above it follows that GROUPALS is equivalent to maximizing:

$$g^*(\mathbf{G}_c, \mathbf{Y}, \mathbf{q}_1, \dots, \mathbf{q}_P) = \text{tr} \mathbf{Y}' \mathbf{G}'_c \left(\sum_j \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j + \sum_j \mathbf{q}_j \mathbf{q}'_j \right) \mathbf{G}_c \mathbf{Y}, \quad (36)$$

while CDR with $\alpha = 1/2$ comes down to maximizing

$$f^*(\mathbf{G}_c, \mathbf{Y}, \mathbf{q}_1, \dots, \mathbf{q}_P) = \text{tr} \mathbf{Y}' \mathbf{U}' \left(\sum_j \mathbf{J} \mathbf{G}_j \mathbf{D}_j^{-1} \mathbf{G}'_j \mathbf{J} + \sum_j \mathbf{q}_j \mathbf{q}'_j \right) \mathbf{U} \mathbf{Y} \quad (37)$$

subject to all the above-mentioned constraints.

Having now seen that GROUPALS and CDR with $\alpha = 1/2$ are fully equivalent, one might think that it is possible to develop an equivalent general approach (i.e., for other values of α) taking the GROUPALS loss function as a basis. Unfortunately, we do not see how this would be possible, since to show the equivalence, a crucial step was to transfer the constraint from $\mathbf{A}$ to $\mathbf{Y}$. This was possible because in the expression $\| \mathbf{X} - \mathbf{U} \mathbf{Y} \mathbf{A}' \|^2$ the constraint $\mathbf{A}' \mathbf{A} = \mathbf{I}$ served as just an identification constraint, and could be replaced by the constraint $\mathbf{Y}' \mathbf{D} \mathbf{Y} = \mathbf{I}$. However, in the full CDR loss function, $\alpha \| \mathbf{X} - \mathbf{X} \mathbf{A} \mathbf{A}' \|^2 + (1 - \alpha) \| \mathbf{X} \mathbf{A} - \mathbf{U} \mathbf{Y} \|^2$, replacing the constraint $\mathbf{A}' \mathbf{A} = \mathbf{I}$ by the constraint $\mathbf{Y}' \mathbf{D} \mathbf{Y} = \mathbf{I}$ would lead to a modified problem, since $\mathbf{A}' \mathbf{A} = \mathbf{I}$ is more than just an identification constraint.

Finally, we wish to remark that, having seen that GROUPALS and CDR with $\alpha = 1/2$ are fully equivalent, there also is a practical gain for interpreting GROUPALS solutions now. That is, from the CDR loss function we see that the projected scores given in $\mathbf{X} \mathbf{A}$ are fitted by the centroids, and by inspecting these projected scores themselves, we can see to what extent they are actually fitted well by the centroids. This can be done by inspecting the plot of $\mathbf{X} \mathbf{A}$ and seeing to what extent

there actually are clear clusters, or, if so desired, it can be done numerically by, for instance, assessing the percentage of misfit as $\| \mathbf{XA} - \mathbf{UY} \|^2 / \| \mathbf{XA} \|^2$.

6 Using CDR in practice: simulation study and applications

6.1 Simulation study

A Monte Carlo experiment was performed to test how well the algorithm performs in different situations of error and dimension reduction. A number of data sets of sample size $I=100$ was generated containing 3 quantitative variables, 3 ordinal variables and one categorical variable. To obtain $K=4$ well-separated groups of objects, the 3 quantitative variables were randomly generated following the procedure proposed by Milligan (1985),² where the generated groups are well-separated, (mildly) truncated multivariate normal mixtures. As such, the resulting structure consists of groups which exhibit the properties of external isolation and internal cohesion. In addition to the 3 classification quantitative variables (representing the generated “true” dimensions), we retained the true membership matrix $\mathbf{U}$, which defines the categorical variable in the data. Without loss of generality, we included only one categorical variable in each generated data set, because from any combination of several categorical variables, a new categorical one can generally be built defining a partition. Three ordinal variables were also generated, with random permutations of the scores as scores on the categorical variable. In this way, all the variables contain the same classification pattern (the same “true” partition of the objects).

To consider a more realistic situation, both categorical and ordinal variables were perturbed at different levels. Specifically, for any categorical and ordinal variable, a percentage of objects was randomly chosen to be perturbed as follows:

- for each sampled object, the “true” category was randomly changed into one of the others with probability equal to the observed relative weights of the categories;
- for each sampled object, the “true” ordinal category was randomly changed into one of the others assigning higher probabilities to the adjacent categories than the far ones as follows: one number among $[-(K-1), -(K-2), \dots, -2, -1, 1, 2, \dots, (K-2), (K-1)]$ was sampled from a binomial distribution $\text{Bin}(2(K-1)-1, 0.5)$ and added to the category label by checking that the number of categories still varies between 1 and K .

The objects to be perturbed were chosen independently for any variable, to generate random errors in the classification structure and avoiding to create outliers.

² In detail, $(I \times 3)$ data matrices were generated following the generating process by Milligan (1985) by setting the parameters $\text{rd}=0.01$ and $\text{rad}=2$.

Moreover, 3 noise quantitative (multinormal) variables and noise categorical variables were included, representing dimensions with no clustering structure which could possibly affect the performance of the methodology in recovering the true partition.

Three experimental factors were considered:

- (a) different percentages (%Error=2%, 4%, 6%, 8%, 10%) of objects to be perturbed for the classification ordinal and categorical variables;
- (b) different correlation level between quantitative noise variables ($\rho=0, 0.9$);
- (c) number of noise categorical variables (NP=0, 1, 3).

The model was fitted to each data set by setting different values for the number of dimensions ($Q=2, 3$) and for the parameter α ($\alpha=0.05, 0.10, 0.15, 0.20, 0.35, 0.50, 0.80, 0.95$).

The algorithm was randomly initialized and the best solution in terms of objective function in 50 different runs of the algorithm was retained to prevent from falling in local optima due to the starting solutions.

In the simulation experiment for each of the 5 (error levels) $\times$ 2 (correlation levels) $\times$ 3 (error levels) factor combinations, 200 data sets were constructed, giving a total of 6000 data sets. The algorithm was run 6000 (data sets) $\times$ 2 (dimensions) $\times$ 50 (random starting values) = 600,000 times for each of the values of α .

The performance of the method has been evaluated by considering in addition to the average loss function, the average modified rand index (Hubert and Arabie 1985), between true and fitted membership matrices of the partition, over the 200 replications of each factor combination.

Tables 1 and 2 display the results for the different cases of correlation, number of categorical noise variables and error level in the classification variables. The performance in terms of recovery of the true classification structure is generally good. In fact the MRand index never goes under 0.5 even in presence of the highest number of perturbed objects in the classification variables. From Table 1, it is evident that when noise quantitative variables are highly correlated, the performance becomes better when α increases, but for the highest levels of α it decreases because a large amount of variability is accounted for, related to the noise variables. It has to be noted that the average MRand index achieves the highest values for $0.15 \leq \alpha \leq 0.35$, sometimes with very slight differences. Increasing the number of categorical noise variables, the best performances generally correspond to lower α values, especially for the case with 3 dimensions.

When the noise quantitative variables are uncorrelated (Table 2), it is necessary to take into account a higher amount of variance (large α values) to get a better recovery, especially with only 2 dimensions, while with 3 dimensions the best results are achieved for $0.15 \leq \alpha \leq 0.35$ depending on the presence of categorical noise variables.

To test how the choice of the right model can be done in real situations where the true clustering structure is unknown, the generalized pseudo F index was used.

Table 1 Results of the simulation study in case of 3 noise-correlated quantitative variables (correlation coefficient 0.9) and (a) 0 categorical noise variables, (b) 1 categorical noise variable, (c) 3 categorical noise variables

		%Error	Average MRand index						
(a)	α	0.05	0.10	0.15	0.20	0.35	0.50	0.80	0.95
$Q=2$	2	0.85	0.92	0.94	0.95	0.96	0.94	0.67	0.58
	4	0.78	0.87	0.91	0.93	0.95	0.92	0.63	0.56
	6	0.72	0.85	0.90	0.91	0.94	0.92	0.61	0.55
	8	0.64	0.80	0.87	0.89	0.93	0.92	0.60	0.53
	10	0.58	0.77	0.86	0.89	0.93	0.91	0.59	0.52
$Q=3$	2	0.94	0.99	0.99	0.99	0.99	0.99	0.72	0.73
	4	0.85	0.98	0.99	0.99	0.99	0.98	0.71	0.73
	6	0.79	0.96	0.98	0.98	0.98	0.98	0.71	0.73
	8	0.73	0.91	0.98	0.98	0.98	0.97	0.69	0.70
	10	0.69	0.87	0.96	0.96	0.96	0.96	0.70	0.72
(b)	α	0.05	0.10	0.15	0.20	0.35	0.50	0.80	0.95
$Q=2$	2	0.77	0.91	0.95	0.97	0.97	0.96	0.66	0.56
	4	0.67	0.87	0.93	0.94	0.96	0.95	0.59	0.55
	6	0.63	0.80	0.89	0.93	0.95	0.92	0.59	0.53
	8	0.55	0.78	0.87	0.90	0.93	0.91	0.59	0.53
	10	0.50	0.74	0.85	0.88	0.91	0.89	0.57	0.51
$Q=3$	2	0.84	0.99	1.00	1.00	0.99	0.99	0.72	0.75
	4	0.74	0.96	0.99	0.99	0.99	0.98	0.70	0.72
	6	0.65	0.90	0.98	0.98	0.98	0.98	0.70	0.72
	8	0.56	0.84	0.97	0.98	0.98	0.97	0.70	0.71
	10	0.51	0.78	0.95	0.96	0.97	0.97	0.69	0.70
(c)	α	0.05	0.10	0.15	0.20	0.35	0.50	0.80	0.95
$Q=2$	2	0.52	0.87	0.95	0.97	0.96	0.93	0.66	0.58
	4	0.47	0.80	0.94	0.95	0.96	0.94	0.60	0.55
	6	0.42	0.72	0.89	0.92	0.93	0.91	0.59	0.53
	8	0.39	0.67	0.86	0.91	0.93	0.90	0.57	0.52
	10	0.36	0.63	0.81	0.89	0.91	0.88	0.60	0.52
$Q=3$	2	0.56	0.92	1.00	1.00	0.99	0.99	0.71	0.72
	4	0.53	0.85	0.99	0.99	0.99	0.98	0.71	0.73
	6	0.47	0.78	0.97	0.98	0.98	0.98	0.71	0.72
	8	0.42	0.71	0.95	0.98	0.98	0.97	0.69	0.70
	10	0.37	0.67	0.93	0.97	0.97	0.96	0.69	0.71

The average of the MRand indices across alpha values is computed over 200 runs of the algorithm

The idea was to first locate the alpha for which pF was maximum and retain the so-called “indicated alpha value” α^* for which the corresponding value of the MRand index was considered. Table 3 displays the percentages (computed on 200 runs) on

Table 2 Results of the simulation study in case of 3 noise uncorrelated quantitative variables (correlation coefficient 0) and (a) 0 categorical noise variables, (b) 1 categorical noise variable, (c) 3 categorical noise variables

		%Error	Average MRand index							
(a)	α	0.05	0.10	0.15	0.20	0.35	0.50	0.80	0.95	
$Q=2$	2	0.82	0.92	0.94	0.95	0.96	0.96	0.97	0.98	
	4	0.74	0.86	0.90	0.93	0.95	0.96	0.97	0.97	
	6	0.67	0.84	0.90	0.91	0.94	0.95	0.96	0.96	
	8	0.60	0.78	0.87	0.90	0.93	0.93	0.94	0.95	
	10	0.51	0.76	0.86	0.89	0.92	0.93	0.95	0.95	
$Q=3$	2	0.90	0.99	0.99	0.99	0.99	0.99	0.99	0.99	
	4	0.81	0.98	0.99	0.99	0.99	0.99	0.98	0.98	
	6	0.75	0.95	0.98	0.98	0.98	0.98	0.98	0.98	
	8	0.68	0.89	0.98	0.98	0.98	0.98	0.98	0.97	
	10	0.64	0.83	0.96	0.97	0.97	0.97	0.97	0.96	
(b)	α	0.05	0.10	0.15	0.20	0.35	0.50	0.80	0.95	
$Q=2$	2	0.73	0.92	0.95	0.97	0.97	0.98	0.98	0.98	
	4	0.64	0.86	0.93	0.94	0.96	0.96	0.97	0.97	
	6	0.59	0.79	0.90	0.93	0.94	0.95	0.96	0.96	
	8	0.52	0.76	0.87	0.91	0.93	0.93	0.94	0.95	
	10	0.48	0.72	0.85	0.89	0.91	0.92	0.93	0.94	
$Q=3$	2	0.78	0.99	1.00	0.99	0.99	0.99	0.99	0.98	
	4	0.70	0.95	0.99	0.99	0.99	0.99	0.99	0.98	
	6	0.60	0.90	0.98	0.98	0.98	0.98	0.98	0.97	
	8	0.56	0.83	0.97	0.98	0.98	0.98	0.97	0.97	
	10	0.50	0.77	0.95	0.97	0.97	0.96	0.96	0.96	
(c)	α	0.05	0.10	0.15	0.20	0.35	0.50	0.80	0.95	
$Q=2$	2	0.48	0.85	0.95	0.97	0.97	0.97	0.97	0.97	
	4	0.43	0.79	0.93	0.95	0.96	0.96	0.96	0.97	
	6	0.37	0.70	0.89	0.92	0.94	0.94	0.95	0.95	
	8	0.33	0.66	0.87	0.91	0.92	0.93	0.94	0.93	
	10	0.33	0.62	0.83	0.89	0.92	0.93	0.93	0.93	
$Q=3$	2	0.50	0.93	0.99	0.99	0.99	0.99	0.98	0.96	
	4	0.50	0.86	0.99	0.99	0.99	0.99	0.97	0.95	
	6	0.43	0.76	0.97	0.98	0.98	0.98	0.97	0.94	
	8	0.40	0.71	0.96	0.98	0.98	0.97	0.96	0.94	
	10	0.36	0.66	0.91	0.97	0.97	0.96	0.95	0.93	

The average of the MRand indices across alpha values is computed over 200 runs of the algorithm

- (a) how often α^* corresponds to a MRand value up to 0.05 lower than the highest MRand value;
 (b) how often α^* gives the highest MRand value (in parentheses).

Table 3 Results of the simulation study. Percentage (over 200 runs) on how often the α value indicated by pseudo F index corresponds to a MRand value up to 0.05 lower than the highest MRand index or corresponds to the highest MRand value (in parentheses)

Correlation between quan- titative noise variables	$\rho = 0.9$			$\rho = 0.0$		
	0	1	3	0	1	3
Number of categorical noise variables						
%Error						
$Q=2$						
2	90.5 (79.5)	91.0 (78.5)	83.5 (75.5)	94.0 (85.0)	93.5 (83.0)	92.5 (82.0)
4	84.5 (76.0)	84.0 (70.0)	86.0 (73.0)	88.5 (77.0)	89.0 (75.5)	87.5 (77.0)
6	85.5 (71.0)	85.0 (65.0)	86.5 (66.5)	91.5 (74.5)	91.0 (71.0)	86.5 (70.0)
8	82.0 (60.5)	82.5 (68.5)	85.0 (66.0)	88.5 (70.5)	87.0 (68.0)	87.0 (66.0)
10	79.5 (57.5)	82.0 (59.5)	83.0 (54.5)	85.5 (66.5)	82.5 (67.5)	82.0 (59.5)
$Q=3$						
2	99.0 (90.0)	99.5 (91.0)	99.0 (88.5)	97.0 (87.5)	99.0 (86.5)	99.0 (87.0)
4	97.0 (80.5)	95.0 (84.0)	96.0 (82.0)	94.5 (82.5)	97.0 (82.5)	95.5 (81.0)
6	90.5 (81.0)	94.0 (77.0)	92.5 (72.5)	93.5 (83.5)	92.5 (72.5)	92.0 (75.0)
8	89.5 (77.5)	92.5 (71.5)	87.0 (68.0)	93.5 (72.0)	92.0 (71.0)	89.5 (69.0)
10	83.0 (63.5)	90.0 (67.5)	85.0 (67.0)	86.0 (66.5)	82.5 (56.5)	86.5 (61.5)

So, the percentage in parentheses indicates the special case where the MRand value at the “indicated alpha” α^* equals the maximum MRand value for that condition.

As expected, the percentages decrease when the percentage of error in classification variables increases, regardless of the possible presence of correlation between quantitative noise variables and regardless of the number of categorical noise variables, maintaining in all cases high values and giving a useful tool for the choice of the α value. It is worthwhile to note that for $Q=2$ dimensions the equality percentages are higher for cases with uncorrelated quantitative noise variables than when these variables were correlated.

6.2 First empirical example analysis: Australian crab data

The Australian Crab data set is a well-known data set (Campbell and Mahon 1974; Venables and Ripley 2002) describing 5 morphological measurements on a total of 200 specimens of crabs, specifically 50 crabs each of two colour forms (Blue and Orange) and both sexes, of the species *Leptograpsus variegatus* collected at Fremantle (W. Australia). The five measurements related to their carapace (shell) are: F =size of the frontal lobe, R =rear width, L =carapace length, W =carapace width, B =body depth.

Fig. 1 Crab data example: loss functions across alpha values

The (200, 7) data matrix containing the 5 quantitative variables and the 2 categorical variables representing species and gender has been analysed by our method. It was chosen to analyse these data using $K=3$ clusters according to the results from previous analyses; the number of dimensions was set to $Q=2$ to adequately capture the structure in a visual display.

The algorithm has been run several times corresponding to alpha values from 0.05 to 0.95 by 0.05 and retaining the best solution in 400 runs starting from random values for each alpha value. The plot for the CDR loss function is displayed in Fig. 1 where the CDR loss function is plotted across the alpha values, together with its two constituents FKM and CATPCA loss functions. While the CDR loss function increases with alpha values, the CATPCA loss decreases, and shows two elbows corresponding to $\alpha=0.20$ and $\alpha=0.45$ and becomes stable for $\alpha>0.55$. The pseudo F index computed across alpha values (see Fig. 2) as well as the loss functions (see Fig. 1) suggest to select $\alpha=0.5$.

The results obtained for $\alpha=0.50$ are displayed in Fig. 3 where the crab data points projected onto the plane spanned by the two dimensions are represented in 3 different colours corresponding to the clusters they belong to. The coordinates are given by $\mathbf{F}=[\mathbf{JG}_1\mathbf{D}_1^{-1/2}, \dots, \mathbf{JG}_5\mathbf{D}_5^{-1/2}, \mathbf{Z}]\mathbf{A}$. Clearly, the clusters are mainly discriminated by the first dimension. The variables are also plotted for the same alpha value to better interpret the new dimensions and clusters. The first dimension represents the highly correlated quantitative variables, identifying the “size” of the crabs. The obtained clustering of crabs is thus primarily related to their size. On the second dimension, species and gender are discriminated to quite a big extent. Such a dimension discriminates female (f) vs. male (m) and orange (o) vs. blue (b) crabs. In particular, the second dimension presents a contrast between F, B, R vs. L, W , and hence indicates to what extent crabs have relatively big body measures compared to their carapace sizes, which confirms what is known about the crabs with respect to species and gender. The scatter plot of the crab points in the left panel of Fig. 3 displays the groups of crabs (in different colours). It is evident that the distinction between the

Fig. 2 Crab data example: pseudo F index plotted against α . **a** total range, **b** zoomed in on interesting area

Fig. 3 Crab data example: data points and variables projected onto the plane spanned by the two dimensions

groups is related to the first dimension, hence to the sizes of the crabs (see above). The presence of the categorical information obtained by crossing the 2 categorical variables species and gender is evident, because the second dimension (in the left panel of Fig. 3) clearly shows four clouds, that can be identified to correspond to the combinations “of”, “bf”, “bm”, and “om”.

6.3 Second empirical example analysis: pricing the Cs of diamond stones

The factors that influence the price of a diamond stone are the 4 Cs of diamond grading: carat, clarity, colour and cut. Specifically, diamonds with no inclusion under a

loupe with a 10 power magnification are labelled IF (“internally flawless”). Lesser diamonds are categorised in descending order as “very very slightly imperfect” VVS1 or VVS2 and “very slightly imperfect” VS1 or VS2. The most prized diamonds display colour purity, they are not contaminated with yellow or brown tones. Top colour purity attracts a grade of D and subsequent degrees of colour purity are rated E, F, G, etc., all the way down the alphabet ladder. To assist shoppers, independent certification bodies assay diamond stones and provide each of them with a certificate listing their caratage and their grades of clarity, colour and cut.

Here, data have been analysed that pertain to a number of round diamond stones, for which details on the caratage, clarity, colour and price have been assessed (http://www2.amstat.org/publications/jse/jse_data_archive.htm). In addition, the variable indicating the certification body is available, namely New York-based Gemmological Institute of America (GIA), Antwerp-based International Gemmological Institute (IGI) and Hoge Raad Voor Diamant (HRD). Their reputations could be a factor in the pricing of the diamond stones. Specifically, five variables are available for $I=308$ round diamonds:

$$\begin{aligned}\text{Carat} &= \text{Weight of diamond stones in carat units;} \\ \text{Colour} &= \text{D, E, F, G, H, I;} \\ \text{Clarity} &= \text{IF, VVS1, VVS2, VS1, VS2;} \\ \text{Certification body} &= \text{GIA, IGI or HRD;} \\ \text{Price} &= \text{Price of diamond stones (Singapore \$).}\end{aligned}$$

Since stones may be divided into 3 clusters due to their size, namely small (less than 0.5 carats), medium (0.5 to less than 1 carat) and large (1 carat and over), three binary variables have been built representing the three caratage ranges. Following Chu (2001), instead of using directly the variable carats, its interactions with such binary variables have been employed by building three quantitative variables denoted (small, medium, large). Specifically, three binary variables have been built for the three sizes of diamonds and three quantitative variables (denoted small, medium, large) have been derived by multiplying such binary variables by carats. So, the “Small” variable has nonzero values (i.e., the carat values) only for the smallest diamonds (less than 0.5 carats), and likewise for the other two variables. Thus, these variables are weighted binary variables. Furthermore, colour and clarity have been recoded on an ordinal scale, while price has been left out of the analysis because it is actually the target of the study.

We have analysed the (308×6) data matrix containing 6 variables: one categorical variable (certification), 2 ordinal variables (colour, clarity), and three quantitative ones (small, medium, large). It was chosen to analyse these data using $K=3$ clusters (hopefully reflecting the expected three price levels); the number of dimensions was set to $Q=2$ to adequately capture the structure in a visual display. The algorithm was run several times corresponding to alpha values from 0.05 to 0.95 by 0.05 and retaining the best solution in 400 runs starting from random starting points.

Fig. 4 Diamonds' example: loss functions across alpha values

Fig. 5 Diamonds' example: pseudo F index plotted against alpha

In Fig. 4, the CDR loss function is plotted across the alpha values, together with its two constituents FKM and CATPCA loss functions. While the loss function increases with alpha values, the CATPCA loss decreases, showing an elbow corresponding to $\alpha=0.1$ and after that decreasing slowly. The pseudo F index (see Fig. 5) increases across the alpha values and reaches its maximum around $\alpha=0.55$.

It is worthwhile noticing that for low values of α (where in the CDR loss function the weight of FKM is higher), corresponding to the elbow in the CATPCA loss function, the solution is dominated by the categories of the qualitative variable certification, but it does not discriminate with respect to the price; in fact, as suggested by the pseudo F index assuming low values, even a too small amount of total variance is accounted for.

The results obtained for $\alpha=0.55$ are displayed in Fig. 6 where the diamonds data points projected onto the plane spanned by the two dimensions are represented in 3

Fig. 6 Diamonds' example: data points and variables projected onto the plane spanned by the two dimensions

different colours corresponding to the clusters they belong to. In the right side of Fig. 6, the variables are also plotted to better interpret the new dimensions and clusters. The first dimension discriminates the diamonds with respect to their sizes (less vs. more than 0.5 carats) and the certification body (IGI vs. GIA/HRD), while the second dimension is mainly related to differentiate between medium and large diamonds. Note that the result is consistent with previous analyses that IGI certification differs from the other two. In the left side of Fig. 6 the clusters of points, labelled by their price levels, happen to correspond exactly to three price levels and exhibit a good separation in discriminating diamond prices.

7 Discussion

The proposed clustering and dimension reduction (CDR) model allows to find a partition of objects and a simultaneous dimension reduction in the very general case where we have both qualitative (nominal and ordinal) and quantitative variables. The model includes in a single framework several known methods (K -means clustering, factorial K -means, projection pursuit clustering or reduced K -means, principal component analysis), but also intermediate cases of object clustering and variable reduction by setting the value of a constant α . Thus, CDR can be seen to cover a whole range of methods from FKM via PPC/GROUPALS/RKM to PCA, by varying α from 0 to 1. The choice of the model in terms of number of clusters, dimensions and value of α can be made by a generalization of the pseudo F index also presented.

An efficient algorithm in the least-squares framework is presented to minimize the loss function which is a compromise between PCA and FKM, taking into account the nature of the mixed variables. The simulation study has shown the

capability to recover the classification structure of the data and a good performance in different conditions of correlation between quantitative noise variables and possible presence of qualitative noise variables. Markos et al. (2018) have developed an R package called **clustrd** for various methods mentioned in the introduction of this paper, and announced (in their very last sentence) that the methods developed in the present paper “might be included in future versions of the package, after further investigation”.

Acknowledgements The authors thank three anonymous reviewers for their important comments on an earlier version of this paper, the most important of which is the equivalence of GROUPALS to the special case of CDR with $\alpha = 1/2$. Also Marieke Timmerman is acknowledged for her important comments on the problematic case leading to perfectly fitting trivial solutions.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Bock HH (1987) On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In: Bozdogan H, Gupta AK (eds) *Multivariate statistical modeling and data analysis*. Reidel, New York, pp 17–34
- Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. *Commun Stat Theory Methods* 3:1–27
- Campbell NA, Mahon RJ (1974) A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. *Australian J Zool* 22:417–425
- Chu S (2001) Pricing the C's of diamond stones. *J Stat Educ* 9:2
- De Leeuw J (1973) *Canonical Analysis of Categorical Data*. University of Leiden, Leiden
- De Leeuw J, Van Rijckevorsel JLA (1980) HOMALS and PRINCALS, some generalizations of principal components analysis. In: Diday E et al (eds) *Data analysis and informatics II*. Elsevier Science Publishers, Amsterdam, pp 231–242
- De Soete G, Carroll JD (1994) K-means clustering in a low-dimensional Euclidean space. In: Diday E et al (eds) *New approaches in classification and data analysis*. Springer, Heidelberg, pp 212–219
- Escofier B (1979) Traitement simultané, de variables qualitatives et quantitatives en analyse factorielle. *Cahiers de l'Analyse des Données* 4:137–146
- Gifi A (1990) *Nonlinear multivariate analysis*. Wiley, Chichester
- Hubert L, Arabie P (1985) Comparing Partitions. *J Classification* 2:193–218
- Hwang H, Dillon WR, Takane Y (2006) An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika* 71:161–171
- Kiers HAL (1991) Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika* 56:197–212
- Linting M, Meulman JJ, Groenen PJF, van der Kooij AJ (2007) Nonlinear principal components analysis: introduction and application. *Psychol Methods* 12:336–358
- Markos A, Iodice D'Enza A, van de Velden M (2018) Beyond tandem analysis: joint dimension reduction and clustering in R. *J Stat Softw* (**manuscript conditionally accepted**)
- Meulman JJ, Heiser WJ, SPSS Inc. (2009) *PASW Categories 18.0*. SPSS Inc., Chicago
- Milligan GW (1985) An algorithm for generating artificial test clusters. *Psychometrika* 50(1):123–127
- Nishisato S (1980) *Analysis of categorical data: dual scaling and its applications*. University Press, Toronto
- Timmerman ME, Ceulemans E, Kiers HAL, Vichi M (2010) Factorial and reduced K-means reconsidered. *Comput Stat Data Anal* 54:1858–1871

- Uno K, Satomura H, Adachi K (2016) Fixed factor analysis with clustered factor score constraint. *Comput Stat Data Anal* 94:265–274
- Van Buuren S, Heiser WJ (1989) Clustering n objects into k groups under optimal scaling of variables. *Psychometrika* 54:699–706
- van de Velden M, Iodice D’Enza A, Palumbo F (2017) Cluster correspondence analysis. *Psychometrika* 82:158–185
- Venables WN, Ripley BD (2002) *Modern applied statistics with S*, 4th edn. Springer, New York
- Vichi M, Kiers HAL (2001) Factorial k -means analysis for two-way data. *Comput Stat Data Anal* 37:49–64
- Vichi M, Rocci R, Kiers HAL (2007) Simultaneous component and clustering models for three-way data: within and between approaches. *J Classification* 24(1):71–98
- Yamamoto M, Hwang H (2014) A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika* 41:115–129

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.