Multiple Correspondence *K*-Means: simultaneous vs sequential approach for dimension reduction and clustering

Mario Fordellone and Maurizio Vichi

Abstract In this work, a discrete model for clustering and a continuous factorial one for dimension reduction are simultaneously fitted to categorical data, with the aim of identifying the best partition of the objects, described by the best orthogonal linear combinations of the factors, according to the least-squares criterion. This new methodology named Multiple Correspondence *K*-Means is an useful alternative to the Tandem Analysis in the case of categorical data. Then, this approach has a double objective: data reduction and synthesis, simultaneously in the direction of rows and columns of the data matrix.

1 Introduction

In the era of "*big data*" complex phenomena - representing reality in economic, social and many other fields - are frequently described by a large number of statistical units and variables. Researchers who have to deal with this abundance of information are often interested to explore and extract the relevant relationships by detecting a reduced set of prototype units and a reduced set of prototype latent variables, both representing the "*golden knowledge*" mined from the observed data. This dimensionality reduction of units and variables is frequently achieved through the application of two types of methodologies: a discrete classification method, producing hierarchical or non-hierarchical clustering and a latent model, creating factors. The two methodologies, generally are not independently applied. In fact, first, the factorial method is used to determine a reduced set of latent variables and then the clustering algorithm is computed on the achieved factors. This sequential strategy

Mario Fordellone
Sapienza University of Rome (Italy), e-mail: mario.fordellone@uniroma1.it

Maurizio Vichi
Sapienza University of Rome (Italy), e-mail: maurizio.vichi@uniroma1.it

of analysis has been called Tandem Analysis (TA) by Arabie and Hubert [1]. By applying first the factorial method it is believed that all the relevant information regarding the relationships of variables is selected by the factorial method, while, the residual information represents noise that can be discarded. Then, the clustering of units complete the dimensionality reduction of data by producing prototype units generally described by centroids, that is, mean profiles of units belonging to clusters.

However, some authors have noted that TA in some situations cannot be reliable because the factorial models applied first may identify factors that do not necessarily include all the information on the clustering structure of units (De Sarbo et al. [4]). In other terms the factorial method may filter out some of the relevant information for the subsequent clustering. A solution to this problem is given by a methodology that includes the simultaneous detection of factors and clusters on the observed data. Many alternative methods combining cluster analysis and the search for a reduced set of factors have been proposed, focusing on factorial methods, multidimensional scaling or unfolding analysis and clustering (e.g., Heiser [8]; De Soete and Heiser [6]). De Soete and Carroll [5] proposed an alternative to the K-Means procedure, named Reduced K-means (RKM), which appeared to equal the earlier proposed Projection Pursuit Clustering (PPC) (Bolton and Krzanowski [3]). RKM simultaneously searches for a clustering of objects, based on the K-means criterion (MacQueen [10]), and a dimension reduction of the variables, based on component analysis. However, this approach may fail to recover the clustering of objects when the data contain much variance in directions orthogonal to the subspace of the data in which the clusters reside (Timmerman at al. [11]). To solve this problem, Vichi and Kiers [12], proposed the Factorial K-Means (FKM) model. FKM combines K-means cluster analysis with PCA, then finding the best subspace that best represents the clustering structure in the data. In other terms FKM selects the most relevant variables by producing factors that best identify the clustering structure in the data. Both RKM and FKM proposals are good alternative to the TA in the case numeric variables have been considered.

When categorical (nominal) variables are observed TA corresponds to apply first Multiple Correspondence Analysis (MCA) and subsequently the K-means clustering on the achieved factors. As far as we know there are no studies that verify if this TA has the same problems observed for quantitative variables. Thus, the first aim of this paper is to discuss if there are limits of the TA in the case of categorical data. The second and most relevant aim of the paper is to present a methodology, named Multiple Correspondence K-Means (MCKM), for simultaneous dimension reduction and clustering in the case of categorical data. The work is structured as follows: in section 2 a background on the sequential and simultaneous approaches is provided, showing an example where TA for categorical data fails to identify the correct clusters. This is a good motivating example that justifies the use of a simultaneous methodology. In section 3 details on the MCKM model are shown, in section 4 the Alternative Least-Square (ALS) algorithm is proposed for MCKM. In section 5 the main theoretical and applied proprieties of the MCKM are discussed

and finally, in section 6, and application on a real benchmark data is given to show the characteristics of MCKM.

2 Statistics background and motivating example

Let $\mathbf{X} = [x_{ij}]$ be a $N \times J$ data matrix corresponding to N units (objects) on which J categorical (nominal) variables have been observed. Tandem Analysis (TA) (Arabie and Hubert [1]; De Sarbo et al. [4]) is the statistical multivariate procedure that uses two methodologies: *(i)* a dimension reduction (factorial) method for finding a set of P factors (generally, $P < J$) better reconstructing the J observed variables (for example by using Principal Component Analysis (PCA) or Factor Analysis (FA)); and *(ii)* a clustering method that partitions the N multivariate objects into K homogeneous and isolated clusters (for example by considering K-Means, or Gaussian Mixture Models). In TA the factorial method is applied first to compute a matrix of component scores; then, the clustering method is applied, sequentially, on the component score matrix. The first methodology detects the maximal part of the total variance by using a reduced set of P components; while the second method maximizes the between variance of the total variance explained in the first analysis. Thus, the variance explained by the factorial method could not be all the between variance of the original variables necessary for the successive clustering methodology. Actually, it may happen that some noise masking the successive clustering could have been included in the P components. Vichi and Kiers [12] show an instructive example where a data set formed by variables with a clustering structure, together with other variables without clustering structure (noise), but having high variance, has been considered. When TA is applied on this typology of data the PCA generally explains also part of the nose data. These last tend to mask the observed clustering structure, and as a consequence, several units are misclassified.

If the J variables considered in the matrix $\mathbf{X}$ are categorical, then TA corresponds, usually, to the application of Multiple Correspondence Analysis (MCA) and K-Means (KM), this last sequentially applied on the factors identified by MCA. The researcher may ask if this TA for the categorical variables has the same limits discussed for the quantitative case. Before considering this, let us first formalize TA in the categorical data case.
The MCA model can be written as

$$J^{1/2}\mathbf{JBL}^{1/2} = \mathbf{YA}' + \mathbf{E}_{MCA}\,, \tag{1}$$

where $\mathbf{Y} = J^{1/2}\mathbf{JBL}^{1/2}\mathbf{A}$ is the $N \times P$ score matrix of the MCA; $\mathbf{A}$ is the $J \times P$ column-wise orthonormal loadings matrix (i.e., $\mathbf{A}'\mathbf{A} = \mathbf{I}_P$); $J^{1/2}\mathbf{JBL}^{1/2} = \mathbf{X}$ is the centered data matrix corresponding to the J qualitative variables, with the binary block matrix $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_j]$ formed by J indicator binary matrices $\mathbf{B}_j$ with elements $b_{ijm} = 1$ if the i^{th} has assumed category m for variable j, $b_{ijm} = 0$ otherwise; $\mathbf{L} = diag(\mathbf{B}'\mathbf{1}_N)$; $\mathbf{J} = \mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'$ is the idempotent centering matrix with $\mathbf{1}_N$ the

N-dimensional vector of unitary elements.
The KM applied on the MCA score matrix $\hat{\mathbf{Y}} = J^{1/2}\mathbf{JBL}^{1/2}\hat{\mathbf{A}}$ can be written as

$$\hat{\mathbf{Y}} = \mathbf{U}\bar{\mathbf{Y}} + \mathbf{E}_{KM} \ , \tag{2}$$

where $\mathbf{U}$ is the $N \times K$ binary and row stochastic memberships matrix, i.e., $u_{ik} \in \{0,1\}$ with $i = 1,\ldots,N$ and $k = 1,\ldots,K$ and $\mathbf{U1}_K = \mathbf{1}_N$, identifying a partition of objects and $\bar{\mathbf{Y}}$ is the $K \times P$ corresponding centroid matrix in the P-dimensional space. Note that $\mathbf{Y} = \mathbf{XA}$, while $\hat{\mathbf{Y}} = \bar{\mathbf{X}}\mathbf{A}$. Finally, $\mathbf{E}_M CA$ and $\mathbf{E}_K M$ are the $N \times J$ error matrices of MCA and KM, respectively.
The Least-Squares (LS) estimation of model (1) corresponds to minimize the loss function

$$\begin{cases} ||J^{1/2}\mathbf{JBL}^{1/2} - \mathbf{YA}'||^2 \underset{\mathbf{A}}{\rightarrow} min \\ \mathbf{A}'\mathbf{A} = \mathbf{I}_P \\ \mathbf{Y} = J^{1/2}\mathbf{JBL}^{1/2} \end{cases} , \tag{3}$$

while LS estimation of model (2) relates to minimize the loss function

$$\begin{cases} ||\hat{\mathbf{Y}} - \mathbf{U}\bar{\mathbf{Y}}||^2 \underset{\mathbf{U},\bar{\mathbf{Y}}}{\rightarrow} min \\ \mathbf{U} \in \{0,1\} \\ \mathbf{U1}_K = \mathbf{1}_N \end{cases} , \tag{4}$$

Thus, given the LS estimates $\hat{\mathbf{A}}$, $\hat{\mathbf{U}}$, $\hat{\bar{\mathbf{Y}}}$ of MCA and KM and considering $\mathbf{Y} = J^{1/2}\mathbf{JBL}^{1/2}\hat{\mathbf{A}}$, the TA procedure has an overall objective function equal to the sum (or mean) of the two objective functions of MCA and KM; formally,

$$f(\hat{\mathbf{Y}},\hat{\mathbf{A}},\hat{\mathbf{U}},\hat{\bar{\mathbf{Y}}}) = \frac{1}{2}\left(||J^{1/2}\mathbf{JBL}^{1/2} - \hat{\mathbf{Y}}\hat{\mathbf{A}}'||^2 + ||\hat{\mathbf{Y}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}||^2\right) . \tag{5}$$

Therefore, TA is the procedure that optimizes sequentially the two objective functions of MCA and KM, which loss can be summarized by (5). However, we now show with an example that this sequential estimation has some limits similar to those evidenced in the quantitative case. In Figure 1, the heat-map of the data matrix of 90 units according to 6 qualitative categorical variables, each one with 9 categories, is shown.

This is a synthetic data set formed by considering multinomial distributions. The first two variables are a mixture of three multinomial distributions with values from 1 to 3, from 4 to 6 and from 7 to 9, respectively, thus defining three clusters of units, each one with equal size (30 units). The other four variables are multinomial distributions with values from 1 to 9 with equal probabilities, thus these do not define clusters of units. We suppose that this is an example of a simulated data set of 90 customers who have expressed their preferences on 6 products on the basis of a Likert scale from 1 (like extremely) to 9 (dislike extremely), passing through 5 (neither like nor dislike). The heat-map in Figure 1 is a graphical representation of data where the individual values contained in the matrix are represented as different levels of grey from white (value 1) to black (value 9) (1 like extremely, 2 like very

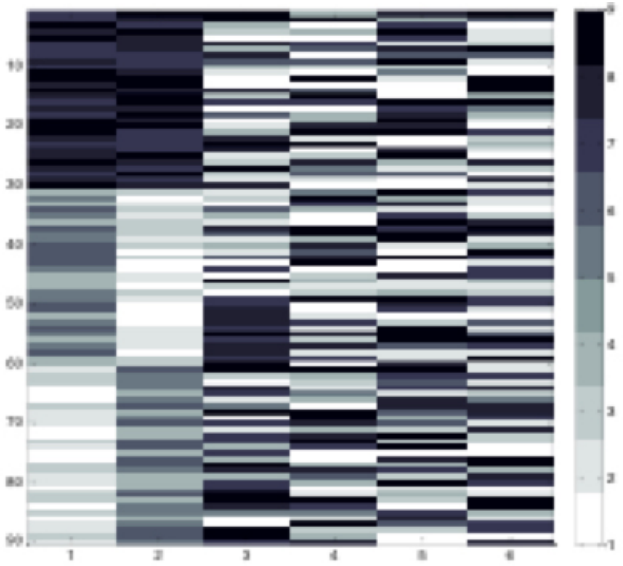

Fig. 1 Heat-map of the 90×6 categorical variables with 9 categories for each variable

much, 3 like moderately, 4 like slightly, 5 neither like nor dislike, 6 dislike slightly, 7 dislike moderately, 8 dislike very much, 9 dislike extremely). By examining the columns of the heat-map (corresponding to products) it can be confirmed that the first two (products A, B)) have a well-defined clustering structure. In fact, the first 30 customers dislike (moderately, very much and extremely), the two products having chosen attributes from 7 to 9, for both products. Customers from 31 to 60 having values from 4 to 6 and from 1 to 3, for the first and second column, respectively, are almost neutral on the first product (like slightly, nether like nor dislike, dislike slightly), but they like the second product (extremely, very much or moderately). Finally, customers from 61 to 90 have values from 1 to 3 and from 4 to 6 in the first and second column, respectively, thus, they like the first product and are substantially neutral for the second. For the other four products (C, D, E, F) the 90 customers do not show a systematic clustering pattern with values that range randomly with equal probability from 1 to 9. Therefore, the 90 customers have two patterns of preferences: "clustered" for products A, B and "random" for products C, D, E and F. On the 90×6 data matrix so defined, the TA was applied by computing first the MCA and successively, by calculating the *K*-means algorithm on the first two components identified by the MCA.

Figure 2, shows the Biplot of categories of the 6 variables named A, B, C, D, E, F and followed by a number between 1 and 9 to distinguish categories. The total loss (5) is 7.39.

It can be clearly seen from the Biplot that the most relevant categories are those of the two variables A and B together with other categories e.g., F7, C7, E9, D1 from variables F, C, E and D. Thus, the clustered and the random patterns of the

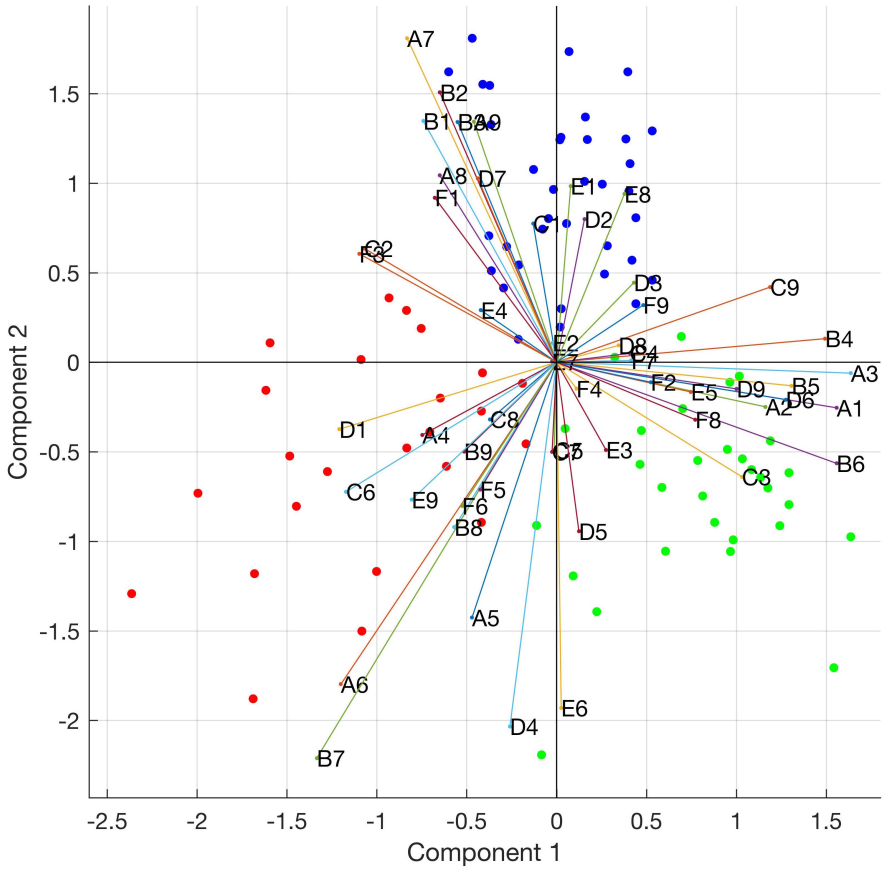

Fig. 2 Biplot of the 90×6 qualitative variables (A, B, C, D, E, F) with categories from 1 to 9. The three clusters are represented by three different colors

customers are assorted and not clearly distinguishable in the Biplot. Furthermore, TA tends to mask the three clusters of costumers, each one originally formed by 30 customers, as shown in the Table 1.

Table 1 Contingency table between K-Means groups and simulated groups

		K-Means			
		Group 1	Group 2	Group 3	Total
Simulated groups	Group 1	30	0	0	30
	Group 2	3	27	0	30
	Group 3	7	1	22	30
	Total	40	28	22	90

In fact, the points classified in the three groups are 40, 28 and 22, respectively. Thus,

11 customers (12%) are misclassified (3 from the second cluster and 8 from the last cluster). The Adjusted Rand Index (ARI) between the generated three clusters and the three clusters obtained by K-means is $ARI = 0.6579$. Then, TA describes imprecisely the three clusters and defines components which do not clearly distinguish the two different preference patterns: the clustered, for products A, B and the random for the products C, D, E, F.

3 Multiple Correspondence K-Means model

Hwang et al. [9] propose a convex combination the homogeneity criterion for MCA and the criterion for K-means; in this paper let us use a different approach by specifying a model for the data, replacing equation (2) into equation (1). Thus, it follows that

$$J^{1/2}\mathbf{JBL}^{1/2} = (\mathbf{U}\bar{\mathbf{Y}} + \mathbf{E}_{KM})\mathbf{A}' + \mathbf{E}_{MCA}\ , \tag{6}$$

and rewriting the error term $\mathbf{E}_{MCKM} = \mathbf{E}_{KM}\mathbf{A}' + \mathbf{E}_{MCA}$, the resulting equation is here named Multiple Correspondence K-Means (MCKM) model:

$$J^{1/2}\mathbf{JBL}^{1/2} = (\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}' + \mathbf{E}_{MCKM})\ . \tag{7}$$

MCKM model identifies, simultaneously, the best partition of the N objects described by the best orthogonal linear combination of variables according to a single objective function. The coordinates of the projections onto the basis are given by the components y_{ip} collected in the matrix $\mathbf{Y} = \mathbf{XA}$. Within this subspace, hence, with these components, a partition of objects is sought such that the objects are "closest" to the centroids of the clusters (Vichi and Kiers [12]). When $\mathbf{X} = J^{1/2}\mathbf{JBL}^{1/2}$ is actually a quantitative data matrix the Least-Squares (LS) estimation of model (7) is equal to the Reduced K-Means (RKM) model, proposed by De Soete and Carroll [5]. Additionally, when equation (7) is post-multiplied both sides by A, the RKM model is transformed into the Factorial K-Means (FKM) model, proposed by Vichi and Kiers [12]. Both models have been formalized for numeric data.

The LS estimation of MCKM corresponds to minimize the objective function

$$\begin{cases} ||J^{1/2}\mathbf{JBL}^{1/2} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'||^2 \underset{\mathbf{A},\mathbf{U},\bar{\mathbf{Y}}}{\rightarrow} min \\ \mathbf{A}'\mathbf{A} = \mathbf{I}_P \\ \mathbf{U} \in \{0,1\} \\ \mathbf{U}\mathbf{1}_K = \mathbf{1}_N \end{cases} . \tag{8}$$

4 Alternating Least-Squares algorithm

The quadratic constrained problem of minimizing (8) can been solved by an Alternative Least-Squares (ALS) algorithm, which is structured on three steps, as follows:

Step 0: Firstly, initial values are chosen for $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$; in particular, initial values for $\mathbf{A}$ and $\mathbf{U}$ can be chosen randomly satisfying the constraints shown in (8), while initial values for are then given at once by $(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{Y}$.
Step 1: Minimize $F([u_{ik}]) = ||J^{1/2}\mathbf{JBL}^{1/2} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'||^2$ with respect to $\mathbf{U}$, given the current values of $\mathbf{A}$ and $\bar{\mathbf{Y}}$. The problem is solved for the rows of $\mathbf{U}$ independently by taking $u_{ik} = 1$ if $F([u_{ik}]) = min\{F([u_{iv}]) : v = 1,\ldots,P; (v \neq k)\}$; $u_{ik} = 0$, otherwise.
Step 2: Given $\mathbf{U}$, update $\mathbf{A}$ and implicitly $\bar{\mathbf{Y}}$ by minimizing (8). The problem is solved by taking the first p eigenvectors of $\mathbf{X}'(\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}')\mathbf{X}$ (e.g., see Vichi M., Kiers H.A.L. [12]).
Step 3: Compute the objective function (8) for the current values of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$. When the updates of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$ have decreased the function value, repeat the step 1 and 2; otherwise, the process has converged.

ALS algorithm monotonically decreases the loss function and, because the constraints on $\mathbf{U}$, the method can be expected to be rather sensitive to local optima. For this reasons, it is recommended the use of many randomly started runs to find the best solution. In some test, it has been valued that, for a good solution (a good local optimal value), the use of 500 random starts usually suffices.

5 Theoretical and applied properties

5.1 Theoretical Property

PROPERTY 1: The LS solution of MCKM obtained by solving the quadratic problem (8) subject to constraints $\mathbf{A}'\mathbf{A} = \mathbf{I}_P$, $\mathbf{U} \in \{0,1\}$, and $\mathbf{U}\mathbf{1}_K = \mathbf{1}_N$ is equivalent to the minimization of the objective function (5) used to give an overall estimation of the loss produced by Tandem Analysis results. In other terms, it can be proved the equality

$$2f(\hat{\mathbf{Y}},\hat{\mathbf{A}},\hat{\mathbf{U}},\hat{\bar{\mathbf{Y}}}) = ||J^{1/2}\mathbf{JBL}^{1/2} - \hat{\mathbf{Y}}\hat{\mathbf{A}}'||^2 + ||\hat{\mathbf{Y}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}||^2 = ||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'||^2\,, \quad (9)$$

where $\mathbf{X} = J^{1/2}\mathbf{JBL}^{1/2}$.
Prof. In fact, after some algebra the objective function of MCKM can be written as

$$||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'||^2 = ||\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'||^2 = tr(\mathbf{X}'\mathbf{X}) - tr(\mathbf{X}'\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}')\,. \quad (10)$$

Thus, it is necessary to prove that the objective function of the TA is equal to (10).

$$
\begin{aligned}
&||\mathbf{X}-\mathbf{XAA}'||^2+||\mathbf{XA}-\mathbf{U}\bar{\mathbf{X}}\mathbf{A}||^2= \\
&tr\{(\mathbf{X}-\mathbf{XAA}')'(\mathbf{X}-\mathbf{XAA}')\}+tr\{(\mathbf{XA}-\mathbf{U}\bar{\mathbf{X}}\mathbf{A})'(\mathbf{XA}-\mathbf{U}\bar{\mathbf{X}}\mathbf{A})\}= \\
&tr(\mathbf{X}'\mathbf{X})-tr(\mathbf{X}'\mathbf{XAA}')-tr(\mathbf{AA}'\mathbf{X}'\mathbf{X})+tr(\mathbf{AA}'\mathbf{X}'\mathbf{XAA}')+ \\
&+tr(\mathbf{A}'\mathbf{X}'\mathbf{XA})-tr(\mathbf{A}'\mathbf{X}'\mathbf{U}\bar{\mathbf{X}}\mathbf{A})-tr(\mathbf{A}'\bar{\mathbf{X}}'\mathbf{U}'\mathbf{XA})+tr(\mathbf{A}'\bar{\mathbf{X}}'\mathbf{U}'\mathbf{U}\bar{\mathbf{X}}\mathbf{A})\ .
\end{aligned}
\tag{11}
$$

Now, knowing that $\mathbf{U}\bar{\mathbf{X}}=\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}=\mathbf{P}_{\mathbf{U}}\mathbf{X}$, where $\mathbf{P}_{\mathbf{U}}$ the idempotent projector of matrix $\mathbf{U}$, equation (11) can be written as

$$
\begin{aligned}
&tr(\mathbf{X}'\mathbf{X})-tr(\mathbf{A}'\mathbf{X}'\mathbf{XA})-tr(\mathbf{AA}'\mathbf{X}'\mathbf{X})+tr(\mathbf{A}'\mathbf{X}'\mathbf{XA})+ \\
&+tr(\mathbf{AA}'\mathbf{X}'\mathbf{X})-tr(\mathbf{A}'\mathbf{X}'\mathbf{P}_{\mathbf{U}}\mathbf{XA})-tr(\mathbf{A}'\mathbf{X}'\mathbf{P}_{\mathbf{U}}\mathbf{XA})+tr(\mathbf{A}'\mathbf{X}'\mathbf{P}_{\mathbf{U}}\mathbf{P}_{\mathbf{U}}\mathbf{XA})= \\
&=tr(\mathbf{X}'\mathbf{X})-tr(\mathbf{A}'\mathbf{X}'\mathbf{P}_{\mathbf{U}}\mathbf{XA})-tr(\mathbf{A}'\mathbf{X}'\mathbf{P}_{\mathbf{U}}\mathbf{XA})+tr(\mathbf{A}'\mathbf{X}'\mathbf{P}_{\mathbf{U}}\mathbf{XA})= \\
&=tr(\mathbf{X}'\mathbf{X})-tr(\mathbf{X}'\mathbf{U}\bar{\mathbf{X}}\mathbf{AA}')\ ,
\end{aligned}
\tag{12}
$$

which complete the prof.

5.2 Applied Property

Let us apply the Multiple Correspondence K-means on the 90×6 data set used in section 2 to show the limits of the Tandem Analysis in case categorical data are considered. The loss function (8) is equal to 7.23, better than the loss of the TA, with an improvement of the loss function of 2%. Even if the improvement seems small this time the biplot of Multiple Correspondence K-means in Figure 3 shows a very clear synthesis of the data. Categories of products A and B are well-distinguished from categories of products C, D, E, F and therefore the clustered and random patterns of preferences of customers are clearly differentiated. Furthermore the clustering structure of the customers is well represented in the Biplot. In fact, the three clusters are formed each one by 30 customers, as expected, and they are more homogeneous and well-separated with respect to the clusters in the Biplot of TA (Figure 3).

The red cluster is formed by customers who like products A and are neutral on the product B (the first 30 rows, in the data set). The blue cluster is formed by customers who like the second product B and dislike the first product A (the second 30 rows of the data set). Finally, the green cluster of customers is formed by persons that dislike the product B and are neutral of on product A (the third and last 30 rows of the data set). So this time no misclassifications are observed for the clusters (see Table 2) and the two different patterns of products are differently represented in the plot as expected.

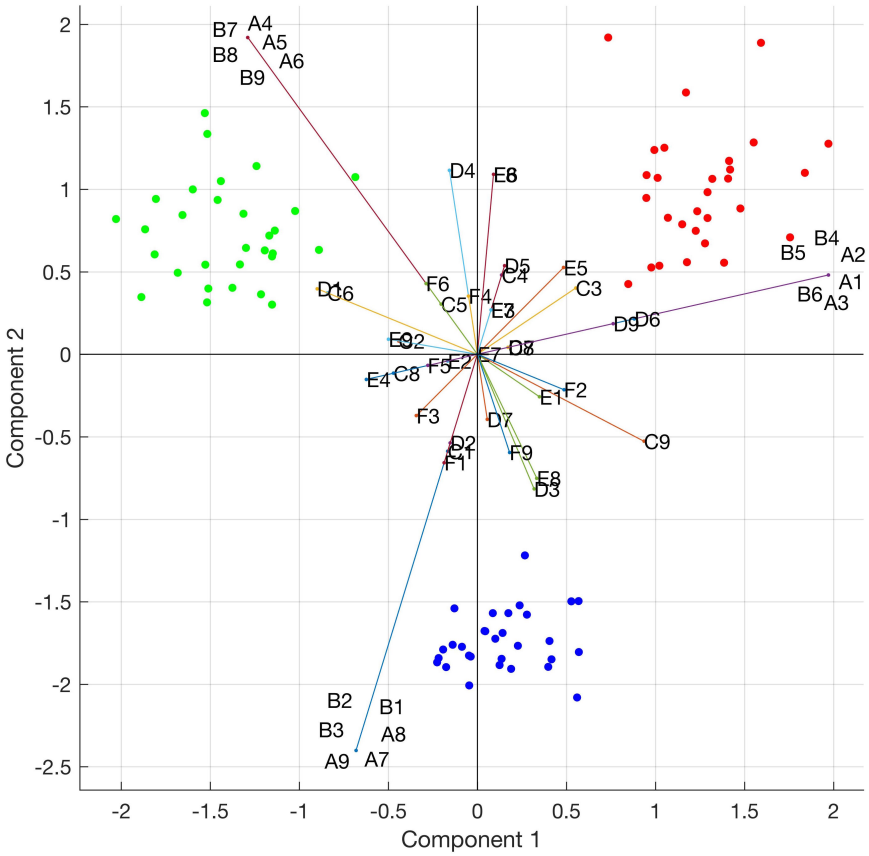

Fig. 3 Biplot of the multiple correspondence K-means . It can be clearly observed that the three cluster are homogeneous and well-separated

Table 2 Contingency table between MCKM groups and simulated groups

		K-Means			
		Group 1	Group 2	Group 3	Total
Simulated groups	Group 1	30	0	0	30
	Group 2	0	30	0	30
	Group 3	0	0	30	30
	Total	30	30	30	90

6 Application on South Korean underwear manufacturer

The empirical data presented in this section, is part of a large survey conducted by a South Korean underwear manufacturer in 1997 (Hwang et al. [9]), where 664 South Korean consumers were asked to provide responses for three multiple-choice items.

In particular, the first item asked which of eight brands of underwear the consumer most prefers (A): (A01) BYC, (A02) TRY, (A03) VICMAN, (A04) James Dean, (A05) Michiko-London, (A06) Benetton, (A07) Bodyguard, and (A08) Calvin Klein; then, both domestic (A01, A02, A03, A04, and A07) and international (A05, A06, and A08) brands were included. The second item asked the attribute of underwear most sought by the consumers (B): (B01) comfortable, (B02) smooth, (B03) superior fabrics, (B04) reasonable price, (B05) fashionable design, (B06) favourable advertisements, (B07) trendy colour, (B08) good design, (B09) various colours, (B10) elastic, (B11) store is near, (B12) excellent fit, (B13) design quality, (B14) youth appeal, and (B15) various sizes. The last item asked the age class of each consumer (C): (C01) 10–29, (C02) 30–49, and (C03) 50 and over. In Table 3 the frequency distributions of the three categorical variables is shown.

Table 3 Frequency distributions of the South Korean underwear manufacturer data

BRAND (A)		ATTRIBUTES (B)		AGE (C)	
A01. BYC	201	B01. Comfortable	398	C01. 10 - 29	239
A02. TRY	131	B02. Smooth	65	C02. 30 - 49	242
A03. VICMAN	30	B03. Superior fabrics	29	C03. 50 and over	183
A04. James Dean	72	B04. Reasonable price	33		
A05. Michiko-London	11	B05. Fashionable design	67		
A06. Benetton	13	B06. Favorable advertisements	7		
A07. Bodyguard	166	B07. Trendy color	15		
A08. Calvin Klein	40	B08. Good design	4		
		B09. Various colors	4		
		B10. Elastic	11		
		B11. Store is near	3		
		B12. Excellent fit	20		
		B13. Design quality	6		
		B14. Youth appeal	1		
		B15. Various sizes	1		

The analysis starts with the application of Multiple Correspondence Analysis and, subsequently, the application of *K*-Means on the computed scores (Tandem Analysis). Hwang et al. [9], suggested to apply MCA by fixing the number of components equal to 2 since sizes of the adjusted inertias appeared to decrease slowly after the first two. The results obtained by the MCA are shown in the Table 4.

Table 4 Results of the MCA model applied on the South Korean underwear manufacturer data

Singular Value	Inertia	Chi-square	Inertia (%)	Cum. Inertia (%)
0.726	0.527	1048.930	6.870	6.870
0.644	0.414	824.878	5.400	12.270
Total	0.941	1873.808	12.270	-

P-value= 0 Degrees of freedom= 196

From Table 4, it is worthy to note that the non-revaluated explained variance of the two computed factors is equal to 12.27% of the total inertia (note that Greenacre [7], recommends to adjust the inertias greater than 1/J using Benzécri's [2] formula). In the Table 5 it is possible to observe the computed loadings among the two components and each category of the data.

Table 5 Loading matrix of the MCA model applied on the South Korean underwear manufacturer data

Component 1			Component 2		
Brand	Attributes	Age	Brand	Attributes	Age
-0.250	-0.133	0.467	0.177	-0.152	0.102
-0.302	-0.065	-0.163	0.090	0.184	-0.374
-0.134	-0.008	-0.346	-0.363	0.285	0.312
0.135	-0.047	-	-0.291	0.234	-
0.161	0.373	-	0.311	0.064	-
0.181	-0.046	-	-0.031	-0.036	-
0.334	0.108	-	0.038	0.030	-
0.175	0.123	-	-0.077	0.017	-
-	-0.097	-	-	0.027	-
-	-0.082	-	-	-0.278	-
-	-0.020	-	-	0.162	-
-	-0.002	-	-	-0.164	-
-	0.152	-	-	-0.231	-
-	0.099	-	-	0.049	-
-	-0.067	-	-	-0.073	-

From the table, it easy to note that the categories with bigger contributions on the first component are: the first two brands of underwear (A01 and A02) and the seventh brand (A07); the fifth attribute (B05); the first and third class of the age (C01 and C03). Whereas, the categories with bigger contribution on the second component are: the third, fourth and fifth brand (A03, A04 and A05); the third, fourth, tenth and thirteenth attribute (B03, B04, B10 and B13); second and third class of the age (C01 and C03). Then, the two component scores represent a very high number of the categories. However, the variables brands (A) and age (C) are more represented than attributes (B). Subsequently, according to the TA approach, the K-Means model on the two component scores has been applied. The fixed number of groups is $K = 3$ as suggested by Hwang et al. [9]. The plot in Figure 4 shows the projection of the single category on the bi-dimensional factorial plane and the distributions of the computed scores. We can note that the three defined groups are underlined with different colours.
The biplot shows that the groups are not well separated and they are characterized by an high inside heterogeneity. In fact, it is very hard to understand the preferences of the consumers that belong to the three groups.

Different results have been obtained with the Multiple Correspondence K-Means approach. Fixing the same number of components and groups, the explained variance of the two components are around to 20%. The component loadings of the MCKM are represented in the Table 6.

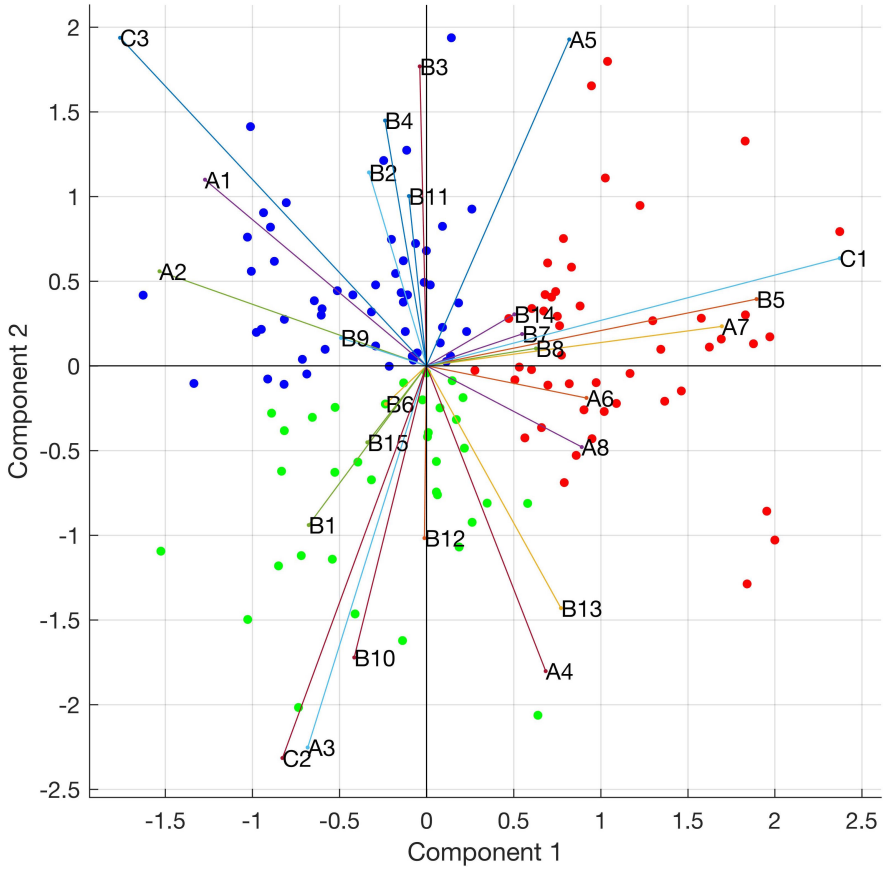

Fig. 4 Biplot of the sequential approach applied on South Korean underwear manufacturer data

In the MCKM model the categories with bigger contributions on the first component are: the first two brands of underwear (A01 and A02) and the seventh brand (A07); the first and the third class of the age (C01 and C03). The categories with bigger contribution on the second component are the fourth, fifth, sixth, seventh and eighth brand (A04, A05, A06, A07 and A8) only. Then, unlike TA, in the MCKM model the variable attributes (B) do not give a relevant contribution.

In the Figure 5 is shown the biplot where are represented the component scores and the three defined groups.

From the plot we can note that the groups are well separated and homogeneous. In fact, it easy to note that the green group (166 observations) are the consumers that prefer the seventh brand (A07); the blue group (361 observations) are the consumers that prefer the first three brands (A01, A02 and A03) and they have mainly an age of 50 years and over (C03); finally the red groups (137 observations) are the consumers that prefer the fourth, fifth, sixth and eight brand (A04, A05, A06, and A08). It is

Table 6 Loading matrix of the MCKM model applied on the South Korean underwear manufacturer data

Component 1			Component 2		
Brand	Attributes	Age	Brand	Attributes	Age
0.429	0.029	-0.252	0.159	0.040	-0.057
0.346	0.028	0.062	0.128	0.068	-0.018
0.158	0.034	0.216	0.046	-0.045	0.086
-0.123	0.025	-	-0.609	0.007	-
-0.048	-0.161	-	-0.238	-0.074	-
-0.052	0.031	-	-0.259	0.007	-
-0.694	-0.016	-	0.449	-0.018	-
-0.092	-0.046	-	-0.454	0.005	-
-	0.061	-	-	0.022	-
-	0.011	-	-	0.034	-
-	0.052	-	-	0.019	-
-	0.036	-	-	-0.093	-
-	-0.052	-	-	-0.132	-
-	-0.054	-	-	0.035	-
-	0.030	-	-	0.011	-

possible to verify these results observing the frequency distributions of the three categorical variables shown in Table 3.

7 Conclusions

Tandem Analysis (TA) is a well-known sequential procedure for clustering and dimensional reduction. It is frequently used in applications for quantitative data, however is has several limitations. In particular, it can fail to find the correct clustering structure with a reduced set of factors (Vichi and Kiers [12]). TA is also frequently used when categorical variables are considered. It corresponds to apply MCA on the original data and successively K-means clustering on the component score matrix of MCA. In this paper it was proved that also this TA has serious problems to correctly classify units and synthesize the relationships of the observed categorical variables. Thus, a model called Multiple Correspondence K-means (MCKM) was proposed and estimated in the LS by using an ALS algorithm. Property 1 proves that the LS estimation of MCKM corresponds to optimize the loss function of the TA which is only imprecisely estimated by the sequential application of MCA and K-means.

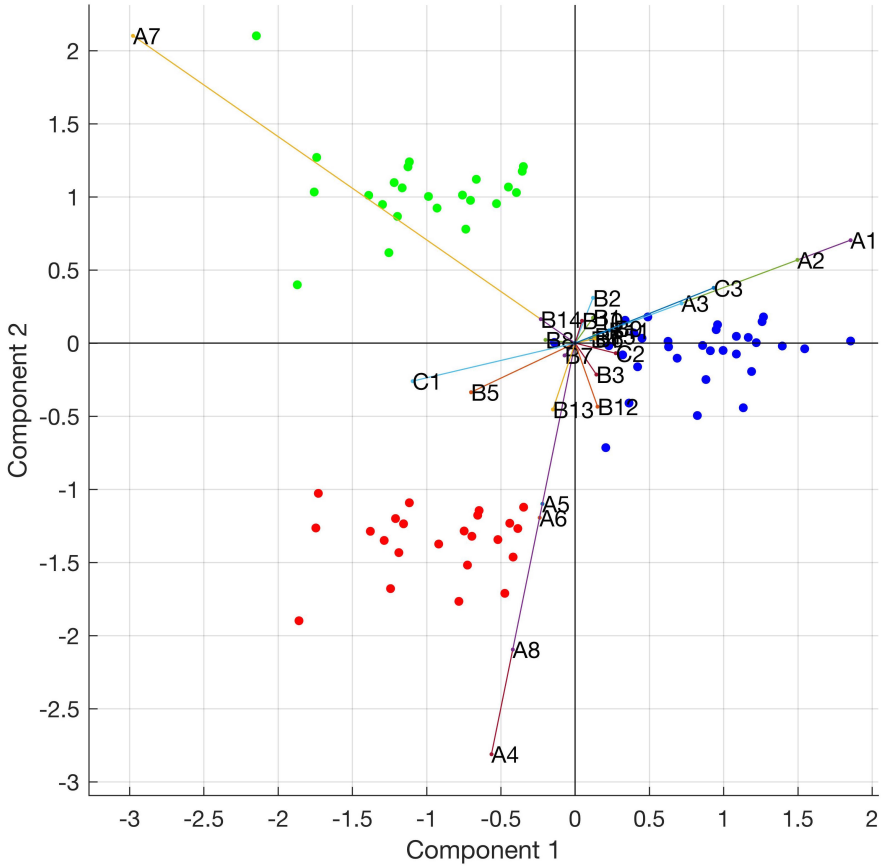

Fig. 5 Biplot of the simultaneous approach applied on South Korean underwear manufacturer data

References

1. Arabie, P.: Cluster analysis in marketing research. *Advanced methods in marketing research*, pp. 160–189 (1994). 2, 3
2. Benzécri, J.P.: Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT.]. *Les cahiers de l'analyse des données*, **4(3)**, pp. 377–378 (1979). 12
3. Bolton, R. J., Krzanowski, W. J.: Projection pursuit clustering for exploratory data analysis. *Journal of Computational and Graphical Statistics* (2012). 2
4. Desarbo, W., Jedidi, K., Cool, K., Schendel, D.: Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters*, **2(2)**, pp. 129–146 (1991). 2, 3
5. De Soete, G., Carroll, J. D.: *K*-means clustering in a low-dimensional Euclidean space. In New approaches in classification and data analysis. *Springer Berlin Heidelberg*, pp. 212–219 (1994). 2, 7
6. De Soete, G., Heiser, W. J.: A latent class unfolding model for analyzing single stimulus preference ratings. *Psychometrika*, **58(4)**, pp. 545–565 (1993). 2

7. Greenacre, Michael J: Theory and applications of correspondence analysis, (1984). 12
8. Heiser, W. J.: Clustering in low-dimensional space. In Information and Classification. *Springer Berlin Heidelberg*, pp. 162–173 (1993). 2
9. Hwang, H., Dillon, W. R., Takane, Y.: An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, **71(1)**, pp. 161–171 (2006). 7, 10, 11, 12
10. MacQueen, J.: Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, **1(14)**, pp. 281–297 (1967). 2
11. Timmerman, M. E., Ceulemans, E., Kiers, H. A., Vichi, M.: Factorial and reduced *K*-means reconsidered. *Computational Statistics & Data Analysis*, **54(7)**, pp. 1858–1871 (2010). 2
12. Vichi, M., Kiers, H. A.: Factorial *K*-means analysis for two-way data. *Computational Statistics & Data Analysis*, **37(1)**, pp. 49–64 (2001). 2, 3, 7, 8, 14
Bioinformatics **17(9)**, 763–774 (2001)