Use of Correspondence Analysis in Clustering a Mixed-Scale Data Set with Missing Data

Michael Greenacre

Abstract Correspondence analysis is a method of dimension reduction for categorical data, providing many tools that can handle complex data sets. Observations on different measurement scales can be coded to be analysed together and missing data can also be handled in the categorical framework. In this study, the method's ability to cope with these problematic issues is illustrated, showing how a valid continuous sample space for a cluster analysis can be constructed from the complex data set from the IFCS 2017 Cluster Challenge.

1 Introduction

This short article details my approach to the clustering of 928 lower back pain patients using the dataset of self-reported baseline assessments from the IFCS Cluster Challenge in 2017. Since the majority of the data are categorical, I have taken the route of correspondence analysis and related methods (CARME) to arrive at a solution, in the process making use of multiple correspondence

Michael Greenacre
Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, Barcelona.
✉ michael.greenacre@upf.edu

Archives of Data Science, Series B
(Online First)
KIT Scientific Publishing
Vol. 1, No. 1, 2019

DOI 10.5445/KSP/1000085952/04
ISSN 2510-0564

analysis, subset correspondence analysis, fuzzy coding, and k-means clustering in reduced-dimensional space. In the course of the methodological description, I point out the benefits and drawbacks of each step of the process. For the most part, the drawbacks can give rise to interesting problems as side-issues, which are suitable for masters-level projects.

2 The data and data recoding

The data for each of the 928 patients consists of 112 variables, whose descriptions are further detailed in this special issue by van Mechelen and Vach (2018). For the statistical treatment, the variables fall into the following categories (numbers of variables in each case):

Continuous	8
Dichotomous	64
Multistate Nominal	9
Ordinal	30
Trichotomous	1

To reduce all data to a categorical scale, the eight continuous variables were fuzzy coded (Aşan and Greenacre, 2011) into four fuzzy categories plus a missing value category. This is a generalization of dummy variable, or "crisp", coding. Instead of cutting up the range of the continuous variable into intervals and coding a particular value strictly into one of the intervals, the value is coded in a fuzzy way into adjacent intervals.

For example, a value might be coded into the five-category variable as [0 0.15 0.85 0 0] to show that it is 15 % in fuzzy category 2 and 85 % in fuzzy category 3. A missing value would be coded as [0 0 0 0 1]. Hence, for each continuous observation, this coding produces four values between 0 and 1 (inclusive), summing to 1, precisely coding the continuous value in the four categories, and a missing value in the fifth. The fuzzy values are determined from a continuous value using a set of membership functions — for a detailed description, see Aşan and Greenacre (2011) or Greenacre and Primicerio (2013, chapter 3).

The only variable that presented a problem with this coding was obeh0, which has 333 values of 0 (i.e. 333 respondents completely agree that treatment

is necessary to decrease their pain), so this variable was coded into a crisp category for the zeros, and 3 fuzzy categories for the other positive values. The other 104 categorical variables generated small numbers of categories, the counts of each of which are as follows:

Variables with 2 categories	64
Variables with 3 categories	25
Variables with 4 categories	3
Variables with 5 categories	1
Variables with 6 categories	2
Variables with 7 categories	8
Variables with 8 categories	1

These were all treated as nominal for the subsequent correspondence analysis (CA) approach, even though many of them (30) are ordinal. Notice that for each variable that has missing values, a separate code (in this case a "9" was used) for missing values and thus a separate category. Since our methodology will be purely nominal, combined with the fuzzy categories, the missing value categories are just additional categories of the data set.

Benefits : All variables are reduced to the same measurement scale. The fuzzy coding especially is useful to categorize continuous variables without losing information (the fuzzy-coded variables can be back-transformed to their original continuous values) and also allow non-linear inter-relationships to be taken into account. The missing value problem is obviated by the simple addition of missing value categories where necessary for the corresponding variables.

Drawbacks : The ordinal information in the 30 ordinal variables is lost. For ordinal variables with many categories, e.g. those with 7 and 8 categories, a fuzzy coding could have been contemplated. On the other hand, a completely different approach, based on Gower's mixed-scale distance function could be used (Gower, 1971), implemented in the R package `cluster`, for example.

3 Methods

3.1 Step 1 – Subset multiple correspondence analysis

Multiple correspondence analysis (MCA) is the analogue of principal component analysis (PCA) for multivariate categorical data, leading to optimal quantifications of the categories and a multivariate Euclidean space of the individuals. The number of categories is 430, consisting of dummy variables for the categorical variables, fuzzy categories for the continuous variables and missing value categories where necessary. The inclusion of missing value categories invariably leads to them dominating the solution, because of their strong associations in the responses. This would be fine if the idea were to study patterns of missing values, but in the present case respondent groups are required based on substantive answers to the survey, not on the pattern of missing responses.

A variant of MCA called subset MCA, where certain chosen categories can be suppressed, is designed for this situation, described by Greenacre and Pardo (2006a,b). In the subset version, missing value categories are not omitted in the original matrix but rather in the matrix of respondent profiles, which are the rows of the crisp and fuzzy values of the dummy variables divided by their respective totals. Then the usual chi-square normalization inherent in correspondence analysis and the dimension reduction calculations are performed on the retained columns – see Greenacre (2016a, chapter 21) for further examples. Hence, in this step, subset MCA is performed, declaring the missing value categories out of the subset.

3.2 Step 2 – Choice of dimensionality

Whereas the number of categories in the recoded data is 430, the actual number of dimensions of this analysis is 309, due to the many linear relationships, one for each categorical variable, and several missing categories being out of the subset. The rule for determining the number of "true" dimensions in a regular MCA is applied here and we remove all dimensions below the eigenvalue threshold of $1/Q$ where Q = number of variables, i.e. 1/112. See Greenacre (2016a, chapter 19) and the conjecture on the last page of this book that the true dimensionality of a multivariate categorical data matrix is determined by this

rule. This conjecture is based on the idea that the dimensionality of a categorical data set is the number of dimensions required to fit all pairwise cross-tabulations exactly in a joint correspondence analysis (JCA), which holds true for simple correspondence analysis. This rule suggests that the dimensionality is 73, so in the following the first 73 dimensions are used, which situates all the respondents in a 73-dimensional continuous Euclidean space, which can be handled by standard methods.

3.3 Step 3 – Clustering using k-means

Non-hierarchical k-means clustering is performed in the 73-dimensional space obtained above. To determine the number of clusters, the clustering is performed for 2, 3, 4, ..., 15 clusters, each repeated with one of 50 random starting points, and up to 500 iterations maximum. The improvement in the between-cluster sum-of-squares is used to decide on the number of clusters, similar to the scree plot in PCA, making use of the elbow "rule-of-thumb".

3.4 Step 4 – Cluster interpretation using correspondence analysis

The clusters are interpreted by cross-tabulating them with all the variables, and performing a correspondence analysis on this concatenated table. This table has the clusters as column categories and the rows as all the categories of the variables used to establish the clusters. Greenacre and Primicerio (2013) call this a CA centroid discriminant analysis, which is in fact a variation of canonical correspondence analysis (CCA) (Ter Braak, 1986), where the constraining variable is the set of dummies (i.e. crisp categorical variables) for the clusters. (If fuzzy clustering is performed, the fuzzy coded memberships can be similarly used.) The variables that make the highest contributions to separating the clusters are identified in an ordered list and provide an indication of the cluster characteristics.

Benefits : The coding and the MCA have brought the whole data set into Euclidean space, and thus the remainder of the exercise is straightforward. The subset idea is put to good use here, and allows all respondents to be used without the inevitable associations amongst the missing value categories, which interfere with the results.

Drawbacks : The conjecture about the true dimensionality of the data set was used. But the conjecture has so far stood up to empirical justification, and no convincing counterexample has been found. Also, the data matrix was not of "pure" categorical data, with dummy variables all crisply coded, but contained 8 fuzzy-coded variables – this is an aspect worth following up.

4 Results

Having extracted the first 73 dimensions of the recoded data set in the subset MCA of the crisp- and fuzzy-coded variables, the k-means clustering gave the following results (figure 1).

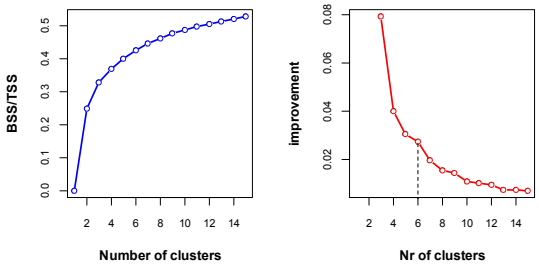

Figure 1: On the left, the between-cluster sum of squares (BSS) relative to total sum of squares (TSS) for increasing number of clusters (starting from one cluster where BSS=0). On the right, the increment (improvement), starting from the benefit from two to three clusters. The six-cluster solution is chosen..

From seven clusters onwards the improvements are small and tailing off, so the six-cluster solution is preferred, with 42.5 % of the total variance explained by between-cluster variance. The number of respondents in each cluster are as follows:

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
90	189	191	177	47	234

For interpretation of the clusters, the CA of the six clusters cross-tabulated with all 112 variables yields an analysis with five dimensions, and the following eigenvalues (output from R package **ca** by Nenadić and Greenacre (2006) reported here):

```
dim    value      %    cum%   scree plot
1      0.093964  56.8  56.8   **************
2      0.033384  20.2  76.9   *****
3      0.022624  13.7  90.6   ***
4      0.009153   5.5  96.1   *
5      0.006378   3.9 100.0   *
       -------- -----
Total: 0.165503 100.0
```

Using the "elbow" rule of thumb (which is essentially what was used in figure 1, on the right, to decide on the number of clusters), it seems that the result is three-dimensional, with 9.4 % of the inertia that can be ascribed to random fluctuations.

Figure 2 shows the CA maps of the respondents, coloured according to their corresponding clusters, in dimensions 1 and 2, and then rotated around the second axis to show dimensions 3 and 2. In Figs 2a and 2b the peeled convex hulls, enclosing approximately 95 % of the points in each cluster are shown, with respect to the two planar projections. In Figs 2c and 2d, the 99 % confidence ellipses for the means of each cluster are shown. As supplementary material an animation is given in a GIF file (see Greenacre (2016b) for example, for a description of the meaning of these ellipses and 3D ellipsoids).The categories of the top 10 most important variables contributing to the separation of the clusters in the projection onto dimensions 1 & 2, are shown in the CA map of figure 3a, while those of the top 10 that separate mainly cluster 5 from the others along dimension 3 are shown in figure 3b (Greenacre, 2013). We can thus give a basic interpretation of the main features of the clusters as follows (labels are given of each category referred to in figure 3):

Cluster 1 (90 people): [rm170:1] Yes: more irritable with people than usual; [fabq100:2] Agree: Work makes/would make pain worse.

Cluster 2 (189 people): [start20:0] No: shoulder/neck pain in last 2 weeks; [tlda0:1] Less than 30 days of LBP in last year.

Cluster 3 (191 people, very acute sufferers) [rm200:1] Yes: Decreased sexual activity; [vas10/4] fuzzy category 4 of LBP intensity (this is highest fuzzy category, thus the highest scale values on this variable — note that the fuzzy categories of this variable in figure 3a go from lowest category at top right to highest category at bottom left); [rm60:1] Yes: Hold on to something to get out

of an easy chair; [rm40:1] Yes: Not doing usual jobs around the house; [start90:1] Very to extremely bothersome pain.

Cluster 4 (177 people, more or less the opposite of cluster 2) [start20:1] Yes: shoulder/neck pain in last 2 weeks; [tlda0:2] More than 30 days of LBP last year.

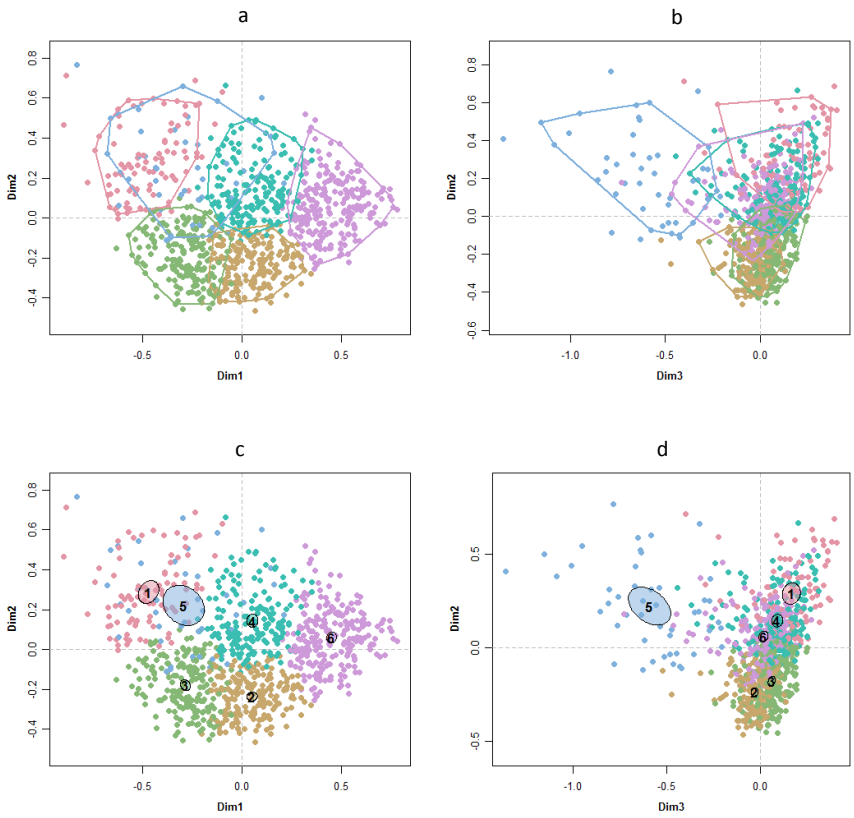

Figure 2: The 928 patients, colour-coded by cluster, and showing (a) and (b) convex hulls enclosing approximately 95 % of the data points in each cluster, with respect to dimensions 1 & 2 and 3 & 2 respectively; (c) and (d) 99 % confidence ellipses for the means of the clusters in the two respective projections. All clusters separate significantly in both of these projections, with the smaller group 5 (hence with largest confidence ellipse) separating on the third dimension..

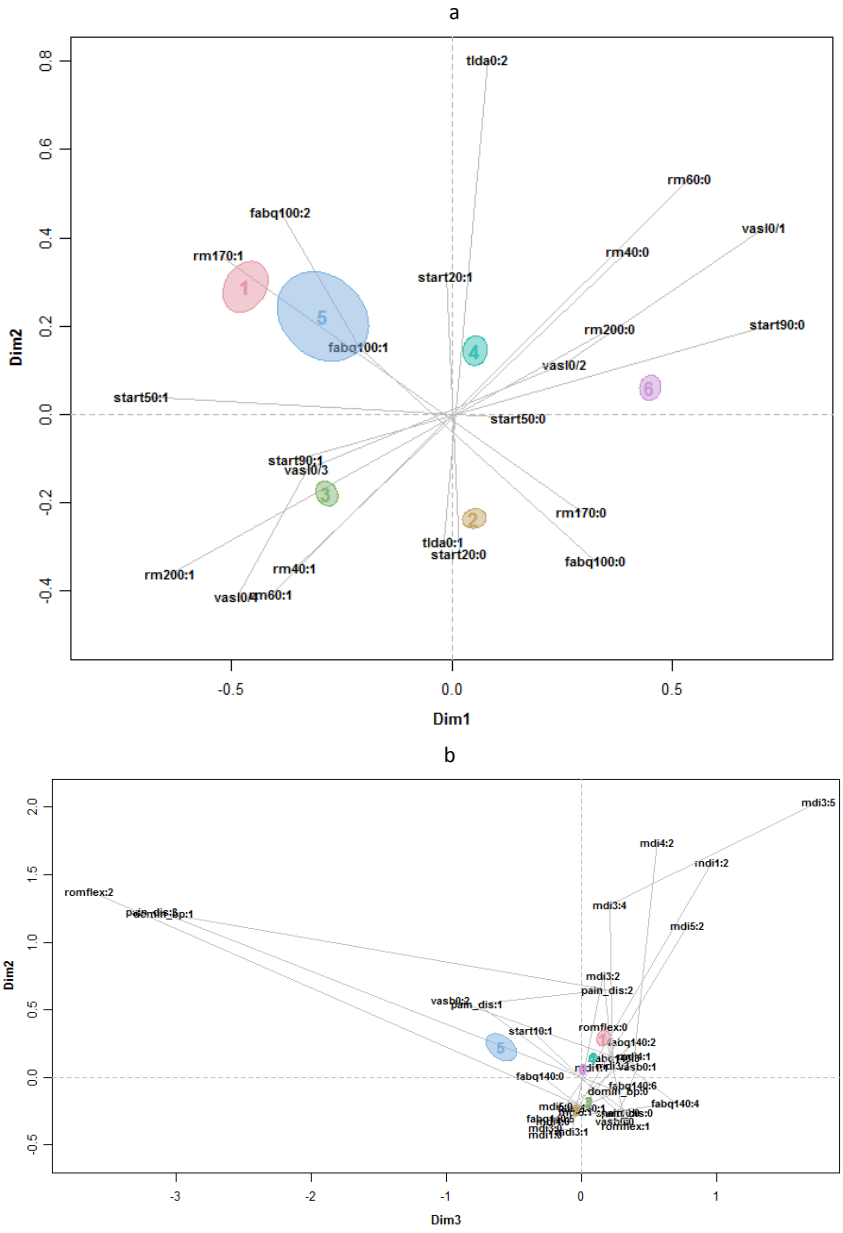

Figure 3: (a) The 10 most important contributors to the two dimensional CA solution that separates the clusters; (b) the 10 most important contributors to the separation of cluster 5 on the third axis..

Cluster 5 (47 people — this interpretation from figure 3b): [romflex:2] Pain on flexion category 2: leg pain with or without back pain; [paindis:3] Pain distribution category 3: leg pain only; [dominbp:1] Yes: LBP not dominating; [vasb0:2] Leg pain intensity category 2: moderate to worst pain imaginable.

Cluster 6: (234 people, opposite of cluster 3, less acute sufferers) [start90:0] No-to-moderately bothersome back pain in last 2 weeks; [vas10/1] lowest fuzzy category of LBP intensity; [rm60:0] No: Hold on to something to get out of an easy chair (i.e. not having to hold on to something); [rm40:0] No: Not doing usual jobs around the house (i.e. doing usual jobs...); [rm200:0] No: Decreased sexual activity (i.e. not decreased...).

In summary, cluster 3 is the acute suffering group of LBP, cluster 6 is the less acute suffering group. Cluster 4 has shoulder/neck pain in addition to more frequent LBP, while cluster 2 has less shoulder/neck pain along with less frequent LBP. Cluster 5 is a smaller group of sufferers of leg pain, not so much LBP, while cluster 1 appears to suffer more from the affects of the pain, e.g. becoming more irritable and with pain that would be aggravated by work.

5 Discussion

The approach adopted in this study is ro reduce all the observed mixed-scale data to a common categorical level, and then profit from the versatility of correspondence analysis in quantifying multivariate categorical data and bringing the data set into a valid continuous space. Once the samples are embedded in this space, an algorithm such as k-means clustering can be implemented in a standard way. The clusters that have been revealed by this approach have a substantive interpretation in terms of the variables that are determinant in defining the clusters. Moreover, this interpretation is facilitated by using correspondence analysis again in the final stage, in order to map the revealed clusters and the categorical variables. Thus, correspondence analysis assists in both analysing the data and analysing the results, which are themselves quite complex, composed of six clusters constructed from over 112 variables coded into a total of 430 categories.

Supplementary material

A Flash animation (`.swf` format) shows the rotation of the individuals and cluster confidence ellipsoids in three-dimensional space.

A video presentation of this article can be found at the CARMEnetwork YouTube channel (CARME = Correspondence Analysis and Related Methods), specifically at `https://youtu.be/C`$_{\text{e}}$`HO4SYKF8`. The video includes some additional results, comparing the clusters according to the demographical variables sex and age, as well as according to the three longitudinal outcomes observed during the one-year period after these survey data were collected: global perceived improvement, LBP intensity and Roland-Morris score.

There are two small errata in the video. On a summary slide of the clusters, at time 8:50, cluster 1 at the top left is erroneously labelled cluster 2. Furthermore, the interpretation of clusters 2 and 4 suffers from a data coding problem which was only recently discovered in the supplied data file. The description of these clusters has been corrected in the present article.

References

Aşan Z, Greenacre M (2011) Biplots of fuzzy coded data. Fuzzy Sets and Systems 183(1):57–71, DOI doi.org/10.1016/j.fss.2011.03.007

Gower J (1971) A general coefficient of similarity and some of its properties. Biometrics 27(4):857–871, DOI 10.2307/2528823

Greenacre M (2013) Contribution biplots. Journal of Computational and Graphical Statistics 22(1):107–122, Taylor & Francis, DOI 10.1080/10618600.2012.702494

Greenacre M (2016a) Correspondence Analysis in Practice, 3rd edn. Chapman & Hall/CRC, New York. ISBN: 978-1-498731-77-5

Greenacre M (2016b) Data reporting and visualization in ecology. Polar Biology 39(11):2189–2205, DOI 10.1007/s00300-016-2047-2

Greenacre M, Pardo R (2006a) Multiple correspondence analysis of subsets of response categories. In: Multiple Correspondence Analysis and Related Methods, Greenacre M, Blasius J (eds), Chapman & Hall/CRC, Boca Raton, FL, p. 197–217, DOI 10.2139/ssrn.847647

Greenacre M, Pardo R (2006b) Subset correspondence analysis: Visualization of selected response categories in a questionnaire survey. Sociological Methods and Research 35(2):193–218, DOI 10.1177/0049124106290316

Greenacre M, Primicerio R (2013) Multivariate Analysis of Ecological Data. BBVA Foundation, Bilbao. ISBN: 978-8-492937-50-9

van Mechelen I, Vach W (2018) Cluster analyses of a target data set in the IFCS cluster benchmark data repository: Introduction to the special issue. Archives of Data Science 1(1):1–12, DOI 10.5445/KSP/1000085952/01

Nenadić O, Greenacre M (2006) Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. Journal of Statistical Software 20:1–13, DOI 10.18637/jss.v020.i03, URL http://www.jstatsoft.org/v20/i03

Ter Braak C (1986) Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. Ecology 67(5):1167–1179, DOI 10.2307/1938672