Robust Diagnostics of Fuzzy Clustering Results Using the Compositional Approach

Karel Hron and Peter Filzmoser

Abstract. Fuzzy clustering, like the known fuzzy k-means method, allows to incorporate imprecision when classifying multivariate observations into clusters. In contrast to hard clustering, when the data are divided into distinct clusters and each data point belongs to exactly one cluster, in fuzzy clustering the observations can belong to more than one cluster. The strength of the association to each cluster is measured by a vector of membership coefficients. Usually, an observation is assigned to a cluster with the highest membership coefficient. On the other hand, the refinement of the hard membership coefficients enables to consider also the possibility of assigning to another cluster according to prior knowledge or specific data structure of the membership coefficients. The aim of the paper is to introduce a methodology to reveal the real data structure of multivariate membership coefficient vectors, based on the logratio approach to compositional data, and show how to display them in presence of outlying observations using loadings and scores of robust principal component analysis.

Keywords: Compositional biplot, compositional data, fuzzy clustering, robust principal component analysis.

1 Overview of Fuzzy Clustering

In fuzzy clustering, each assignment of an object is distributed proportionally to all clusters through membership coefficients according to the similarity to

Karel Hron
Palacký University, 77146 Olomouc, Czech Republic
e-mail: hronk@seznam.cz

Peter Filzmoser
Vienna University of Technology, 1040 Vienna, Austria
e-mail: p.filzmoser@tuwien.ac.at

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 245–253.
springerlink.com

each of the clusters. The number of clusters k for the n objects needs to be provided in advance. Then an objective function

$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^2}, \tag{1}$$

that contains only the similarity measure $d(i,j)$ and the desired membership coefficients u_{iv} of the i-th object to the v-th cluster, needs to be minimized. The measure $d(i,j)$ can be chosen e.g. as squared Euclidean distance when the fuzzy k-means method is applied [3, 4]; an alternative choice is described in [10]. Each object is usually assigned to a cluster with the highest membership coefficient. On the other hand, the refinement of the hard clustering result enables to consider also the possibility of assigning to another cluster according to prior knowledge or specific data structure of the membership coefficients. It means that although an observation belongs to a certain cluster according to the classification rule, the data structure of the membership coefficients implies its pertinence rather to another cluster.

Obviously, the sum of the membership coefficients equals 1 or 100 (in case of proportions or percentages, respectively), so their sample space can be considered to be a k-part simplex,

$$\mathcal{S}^k = \{\mathbf{u} = (u_1, \ldots, u_k)', \, u_i > 0, \, \sum_{i=1}^{k} u_i = 1\}, \tag{2}$$

the prime stands for a transpose. Here we have excluded the case of zero membership values since then the predefined number of clusters obviously needs to be revisited. The important difference of fuzzy clustering to hard clustering methods is contained in the fact that with the latter we obtain a detailed information about the data structure. On the other hand, with an increasing number of the involved groups the results become quite complex so that the obtained information cannot be easily processed further.

For this reason, in this paper we focus on the case of more clusters involved into the analysis and provide a tool to display the multivariate data structure of the membership coefficients using a biplot of loadings and scores from principal component analysis [9]. Hereat we consider in particular a specific data structure of the coefficients, that contain naturally only relative information, and can thus be identified with the concept of compositional data [1]. In addition, we apply a robust counterpart of principal component analysis to ensure that the obtained diagnostics tool will not be influenced by outlying observations. The next section provides a brief review on compositional data and the log-ratio approach for their statistical analysis. Then we introduce classical and robust principal component analysis to construct a biplot and demonstrate how it can be applied in case of compositional data. Finally, the theoretical results will be applied to a real-world example.

2 Relative Information and Compositional Data

Each vector of membership coefficients contains exclusively relative information, thus only ratios between its parts are informative. In the context of fuzzy clustering, the coefficients are normalized to a prescribed constant sum constraint (proportions, percentages). However, this is not a necessary condition but rather a proper representation of the observations, also a positive constant multiple of the vector would provide exactly the same information. In addition, also the concept of relative scale plays an important role here: if a membership coefficient of a certain group increases from 0.1 to 0.2 (two times), it is not the same as an increase from 0.5 to 0.6 (1.2 times), although the Euclidean distances are the same in both cases. All these above properties can be found in the concept of compositional data as introduced in the early 1980s by John Aitchison [1]. The properties of this kind of observations induce a special geometry of compositional data, the Aitchison geometry on the simplex [6] that forms for k-part compositional data, a Euclidean space of dimension $k-1$. Then the main goal is to represent compositional data in orthonormal coordinates with respect to the Aitchison geometry and to perform usual multivariate methods for their statistical analysis. This concept is closely connected with the family of isometric log-ratio (ilr) transformations from the $\mathcal{S}^k$ to the $(k-1)$-dimensional real space $\mathbf{R}^{k-1}$ [5]. One popular choice results for a composition $\mathbf{u} = (u_1, \ldots, u_k)'$ in ilr coordinates $\mathbf{z} = (z_1, \ldots, z_{k-1})'$, where

$$z_i = \sqrt{\frac{k-i}{k-i+1}} \ln \frac{u_i}{\sqrt[k-i]{\prod_{j=i+1}^{k} u_j}}, \; i = 1, \ldots, k-1. \tag{3}$$

Obviously, the ilr transformations move the Aitchison geometry on the simplex isometrically to the usual Euclidean geometry in real space, i.e. to the geometry that we are used to work in. This has also consequences for visualization of the compositional data structure. Three-part compositions are traditionally displayed in a ternary diagram. The ternary diagram is an equilateral triangle $U_1U_2U_3$ such that a composition $\mathbf{u} = (u_1, u_2, u_3)'$ is plotted at a distance u_1 from the opposite side of vertex U_1, at a distance u_2 from the opposite side of vertex U_2, and at a distance u_3 from the opposite side of the vertex U_3 (see, e.g., [1, 12]).

An example can be seen in Fig. 1 with the well-known Iris data set [8] that contains measurements for 50 flowers from each of 3 species of iris. Fuzzy k-means clustering was applied with $k = 3$. The ternary diagram (left) shows the resulting membership coefficients, where the lines correspond to equal coefficients in two groups. The lines can thus be considered as separation lines for a hard cluster assignment. The plot symbols correspond to the true group memberships. One of the clusters (circles) is clearly distinguishable, but the other two clusters show some overlap that leads to a misclassification. The

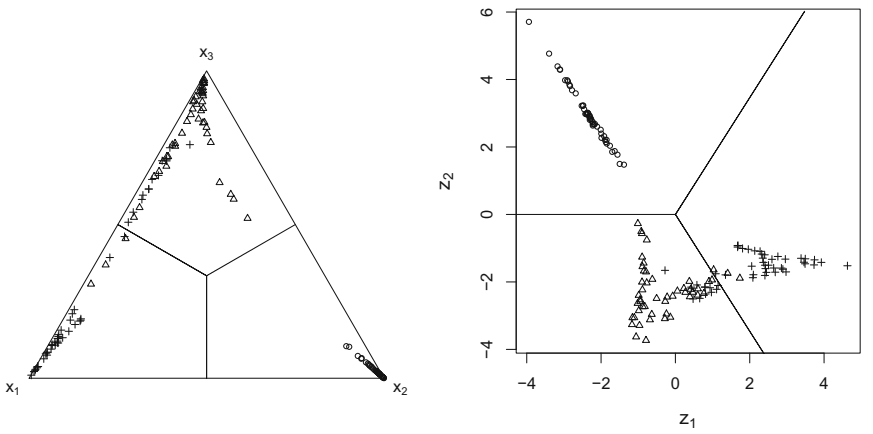

Fig. 1 Membership coefficients of the Iris data in the ternary diagram (left) and after ilr transformation (right) are displayed together with borders (lines) for the classification rule. The symbols correspond to the true memberships.

right plot panel shows the ilr-transformed results, again with the separating lines. The misclassified observations are of course still the same, but the data structure is much better visible in the overlapping region. In this plot the distances are in terms of the usual Euclidean geometry, while in the ternary diagram one has to think in the Aitchison geometry.

Although the ilr transformation has nice geometrical properties, an interpretation of the orthonormal coordinates is sometimes quite complex. Thus, for the purpose of a compositional biplot introduced in the next section, a representation of compositions in a special generating system is more appropriate. The resulting coordinates correspond to the centred logratio (clr) transformation [1], given for a k-part composition $\mathbf{u}$ as

$$(y_1, \ldots, y_k)' = \left(\frac{u_1}{\sqrt[k]{\prod_{i=1}^{k} u_i}}, \ldots, \frac{u_k}{\sqrt[k]{\prod_{i=1}^{k} u_i}} \right)'. \tag{4}$$

The clr transformation seems easier to handle than the ilr transformation, however, it leads to a singular covariance matrix, because the sum of y_i, $i = 1, \ldots, k$, equals zero. This makes the use of robust statistical methods not possible. In the next section we show how the ilr transformation can be utilized in this case.

3 Diagnostics Using a Robust Compositional Biplot

Unfortunately, for more than three-part compositional data it is not possible to visualize them in a planar graph without dimension reduction. A proper tool for this purpose seems to be the compositional biplot [2]. It displays both samples and variables of a data matrix graphically in the form of scores and loadings of the first two principal components [9]. Note that the well-known principal component analysis is appropriate for this purpose, because it explains most of the variability of the original multivariate data by only few new variables (the mentioned principal components). Usually, samples in the biplot are displayed as points while variables are displayed either as vectors or rays. For compositional data, one would intuitively construct the biplot for ilr-transformed data, however, due to the complex interpretation of the new variables it is common to construct the compositional biplot for clr-transformed compositions as proposed in [2]. The scores represent the structure of the compositional data set in the Euclidean real space, so they can be used to see patterns and clusters in the data. The loadings (rays) represent the corresponding clr-variables. In the compositional biplot, the main interest is concentrated to links (distances between vertices of the rays); concretely, for the rays i and j, $i, j = 1, \ldots, k$, the link approximates the (usual) variance $\mathrm{var}(\ln \frac{u_i}{u_j})$ of the logratio between the compositional parts (clusters) u_i and u_j. Hence, when the vertices coincide, or nearly so, then the ratio between u_i and u_j is constant, or nearly so, and the corresponding clusters are redundant. In addition, directions of the rays signalize where observations with dominance of the clusters are located. Although the dimension reduction, caused by taking only the first two principal components, naturally leads to some inconsistencies (observations from different clusters may overlap, also the display of classification boundaries is not meaningful), the biplot can be used to reconstruct the multivariate data structure and reveal reasons for misclassification within fuzzy clustering.

However, through all the advantages of the compositional biplot, outliers can substantially affect results of the underlying principal component analysis and depreciate the predicative value of the biplot. For this reason, a robust version of the biplot is needed. Because the principal component analysis is based on the estimation of location and covariance, we need to find proper alternatives to the standard choice, represented by the arithmetic mean and the sample covariance matrix that can be strongly influenced by outlying observations. Among the various proposed robust estimators of multivariate location and covariance, the MCD (Minimum Covariance Determinant) estimator (see, e.g., [11]) became very popular because of its good robustness properties and a fast algorithm for its computation [13]. The MCD estimator looks for a subset h out of n observations with the smallest determinant of their sample covariance matrix. A robust estimator of location is the arithmetic mean of these observations, and a robust estimator of covariance is the sample covariance matrix of the h observations, multiplied by a factor for

consistency at normal distribution. The subset size h can vary between half the sample size and n, and it will determine the robustness of the estimates, but also their efficiency.

Besides robustness properties the property of affine equivariance of the estimators of location and covariance plays an important role. The location estimator T and the covariance estimator C are called affine equivariant, if for a sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of n observations (e.g. ilr-transformed membership vectors) in $\mathbf{R}^{D-1}$, any nonsingular $(D-1) \times (D-1)$ matrix $\mathbf{A}$ and for any vector $\mathbf{b} \in \mathbf{R}^{D-1}$ the conditions

$$T(\mathbf{A}\mathbf{z}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{z}_n + \mathbf{b}) = \mathbf{A}T(\mathbf{z}_1, \ldots, \mathbf{z}_n) + \mathbf{b},$$
$$C(\mathbf{A}\mathbf{z}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{z}_n + \mathbf{b}) = \mathbf{A}C(\mathbf{z}_1, \ldots, \mathbf{z}_n)\mathbf{A}'$$

are fulfilled. The MCD estimator shares the property of affine equivariance for both the resulting location and covariance estimator.

Because the robust statistical methods cannot work with singular data, the robust scores and loadings must be computed from ilr-transformed compositions before their representation in the clr space. Below we provide some technical details according to paper [7].

Given an $n \times k$ data matrix $\mathbf{U}_{n,k}$ with n membership coefficient vectors $\mathbf{u}_i'$, $i = 1, \ldots, n$, in its rows. Applying the clr transformation to each row results in the clr-transformed matrix $\mathbf{Y}$. The relation

$$\mathbf{Z} = \mathbf{Y}\mathbf{V} \tag{5}$$

for the ilr-transformed data matrix $\mathbf{Z}$ of dimension $n \times (k-1)$ follows from the relation between clr and ilr transformations where the columns of the $k \times (k-1)$ matrix $\mathbf{V}$ contain orthonormal basis vectors of the hyperplane $y_1 + \cdots + y_k = 0$, $\mathbf{V}'\mathbf{V} = \mathbf{I}_{k-1}$ (identity matrix of order $k-1$) [5]. Using the location estimator $T(\mathbf{Z})$ and the covariance estimator $C(\mathbf{Z})$ for the ilr-transformed data, the principal component analysis transformation is defined as

$$\mathbf{Z}^* = [\mathbf{Z} - \mathbf{1}\,T(\mathbf{Z})']\mathbf{G}_\mathbf{z}. \tag{6}$$

The $(k-1) \times (k-1)$ matrix $\mathbf{G}_\mathbf{z}$ results from the spectral decomposition of

$$C(\mathbf{Z}) = \mathbf{G}_\mathbf{z}\mathbf{L}_\mathbf{z}\mathbf{G}_\mathbf{z}', \tag{7}$$

where the matrix $\mathbf{L}_\mathbf{z}$ is made up of the sorted eigenvalues of matrix $C(\mathbf{Z})$.

If the original data matrix has rank $k-1$, the matrix $\mathbf{Z}$ will also have full rank $k-1$, and an affine equivariant estimator like MCD can be used for $T(\mathbf{Z})$ and $C(\mathbf{Z})$, resulting in robust principal component scores $\mathbf{Z}^*$ and loadings $\mathbf{G}_\mathbf{z}$. However, since these are no longer easily interpretable, we have to back-transform the results to the clr space. The scores in the clr space, $\mathbf{Y}^*$, are identical to the scores $\mathbf{Z}^*$ of the ilr space, except that the additional last column of the clr score matrix has entries of zero. For obtaining the

back-transformed loading matrix we can use relation (5). For an affine equivariant scatter estimator we have

$$C(\mathbf{Y}) = C(\mathbf{ZV}') = \mathbf{V}\, C(\mathbf{Z})\, \mathbf{V}' = \mathbf{V}\, \mathbf{G_z} \mathbf{L_z} \mathbf{G_z}'\, \mathbf{V}', \tag{8}$$

and thus the matrix

$$\mathbf{G_y} = \mathbf{VG_z} \tag{9}$$

represents the matrix of eigenvectors to the *nonzero* eigenvalues of $C(\mathbf{Y})$ (with the property $\mathbf{G}_\mathbf{y}'\mathbf{G_y} = \mathbf{I}_{k-1}$). The nonzero eigenvalues of $C(\mathbf{Y})$ are the same as for $C(\mathbf{Z})$ and consequently the explained variance with the chosen number of principal components remains unchanged. Finally, the robust loadings and scores can be used to obtain a robust biplot for compositional data.

The above introduced theoretical framework is applied to geochemical data originated from a 120 km transect running through Oslo. In total, 360 samples from nine different plant species (40 samples for each species) were analyzed for the concentration of 25 chemical elements. The data set is available in the R package `rrcov` as object `OsloTransect`. Here we only used the variables with reasonable data quality, namely Ba, Ca, Cr, Cu, La, LOI, Mg, Mn, P, Pb, Sr and Zn. Since the data set is of compositional nature itself, we first used the ilr-transformation and afterwards applied fuzzy k-means clustering with $k = 9$ (number of different plant species in the data set). This results in nine-part membership coefficients, and thus their visualization in a ternary diagram is no longer possible.

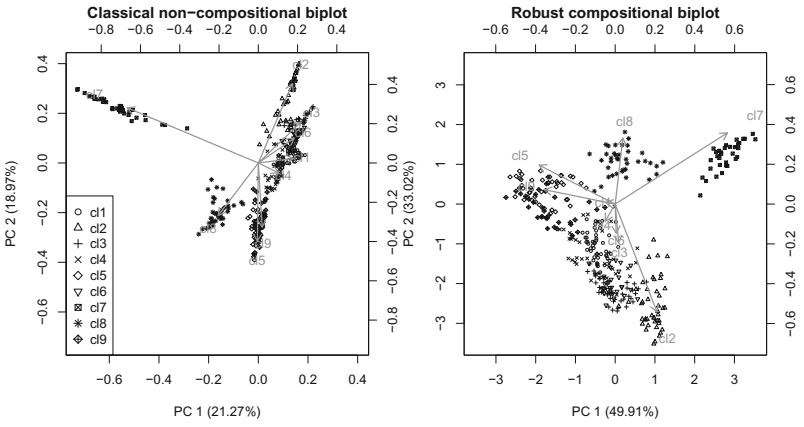

Fig. 2 Biplot resulting from an application to the untransformed membership coefficients (left), and robust biplot resulting from transformed membership coefficients (right)

Without being aware of the above approach based on compositional data analysis, one would probably try to summarize the information contained in the matrix of membership coefficients by principal component analysis (PCA). This procedure is applied here for comparison, and the resulting biplot is presented in Fig. 2 left. The symbols refer to the clusters that have been found with k-means clustering. One can see that there is a certain grouping structure, but there is a lot of overlap of the groups. This is due to an application of PCA in a inappropriate space, the simplex sample space. Note that a robust PCA applied in this space would not lead to an improvement.

Next we apply the procedure as proposed above, by first transforming the membership coefficients, and then applying robust PCA. The resulting robust compositional biplot is displayed in Fig. 2 right. This plot allows for a much better visual inspection. In contrast to the previous biplot, here the first two principal components explain more than 80% of the total variance. It can be seen that fuzzy k-means clustering indeed gave membership coefficients that correspond to relatively clearly separated groups. This also verifies that the algorithm worked well, and that the clustering structure in the data is clearly present. Here we do not further analyse if the correct groups (plant species) were identified, since we are not evaluating the clustering procedure itself.

Acknowledgements. This work was supported by the grant Matematické modely a struktury, PrF_2011_022.

References

1. Aitchison, J.: The statistical analysis of compositional data. Chapman & Hall, London (1986)
2. Aitchison, J., Greenacre, M.: Biplots of compositional data. Applied Statistics 51, 375–392 (2002)
3. Bezdek, J.C.: Cluster validity with fuzzy sets. J. Cybernetics 3, 58–73 (1973)
4. Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. J. Cybernetics 3, 32–57 (1973)
5. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric logratio transformations for compositional data analysis. Mathematical Geology 35, 279–300 (2003)
6. Egozcue, J.J., Pawlowsky-Glahn, V.: Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds.) Compositional Data in the Geosciences: From Theory to Practice, Geological Society, London (2006)
7. Filzmoser, P., Hron, K., Reimann, C.: Principal component analysis for compositional data with outliers. Environmetrics 20, 621–632 (2009)
8. Fisher, R.A.: The use of multiple measurements in axonomic problems. Annals of Eugenics 7, 179–188 (1936)
9. Gabriel, K.R.: The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453–467 (1971)

10. Kaufman, L., Rousseeuw, P.J.: Finding groups in data. John Wiley & Sons, New York (1990)
11. Maronna, R., Martin, R.D., Yohai, V.J.: Robust statistics: theory and methods. John Wiley & Sons, New York (2006)
12. Mocz, G.: Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins. Protein Science 4, 1178–1187 (1995)
13. Rousseeuw, P., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223 (1999)