Robust Reduced k-Means and Factorial k-Means by trimming

Un approccio robusto a Reduced k-Means e Factorial k-Means

Luca Greco and Antonio Lucadamo and Pietro Amenta

Abstract In this paper we propose a robust version of Reduced and Factorial k-means, based on a trimming strategy. Reduced and Factorial k-means are data reduction techniques for simultaneous dimension reduction and clustering. The occurrence of data inadequacies can invalidate standard analyses. An appealing approach to develop robust counterparts of Reduced and Factorial k-means is given by impartial trimming. The idea is to discard a fraction of observations that are selected as the most distant from the centroids.

Abstract *In questo lavoro viene proposta una versione robusta di Reduced e Factorial k-means, basata su una procedura di trimming. Reduced e Factorial k-means sono tecniche che simultaneamente realizzano una riduzione della dimensionalitá e della numerosità, mediante analisi in componenti principali e k-means, rispettivamente. La presenza di contaminazione nei dati puó invalidare le analisi standard. Un approccio utile per sviluppare una procedura robusta alla presenza di valori anomali é rappresentato dal trimming, che si basa sull'idea di eliminare le osservazioni piú distanti dai centroidi stimati.*

Key words: Clustering, Factorial k-means, Reduced k-means, Trimmed k-means

Luca Greco
DEMM University of Sannio, Piazza Arechi II Benevento, e-mail: lucgreco@unisannio.it

Antonio Lucadamo
DEMM University of Sannio, Piazza Arechi II Benevento, e-mail: antonio.lucadamo@unisannio.it

Pietro Amenta
DEMM University of Sannio, Piazza Arechi II Benevento, e-mail: amenta@unisannio.it

1 Introduction

Reduced (De Soete and Carroll, 1994) and Factorial k-means (Vichi and Kiers, 2001) (RKM and FKM, respectively, hereafter) are data reduction techniques aimed at performing principal components and k-means clustering simultaneously. The main idea is that cluster centroids are located in a low dimensional subspace determined by the most relevant features.

Let $\mathbf{X}$ be the $n \times p$ zero centered data matrix, where n denotes the number of objects and p the number of variables, k be the number of clusters and $q < p$ the number of components, with $k \geq q+1$. We denote by $\mathbf{U}$ the $n \times k$ membership matrix whose i^{th} row has a one corresponding to the cluster assignment of the i^{th} object and zero otherwise, whereas $\mathbf{A}$ is the $p \times q$ loadings matrix and $\mathbf{Y} = \mathbf{XA}$ is the $n \times q$ scores matrix. RKM looks for centroids in a low dimensional subspace that minimize the distance of the data points from such centroids. The optimization problem connected with RKM can be expressed as

$$\min_{A,\bar{Y}} ||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}||^2 = \min_{A,\bar{Y}} \sum_{i=1}^{n} \min_{c=1,\ldots,k} \sum_{j=1}^{p} \left(x_{ij} - \sum_{j'=1}^{q} \bar{y}_{cj'} a_{j'j} \right)^2 , \tag{1}$$

where $\bar{\mathbf{Y}}$ is the $k \times q$ matrix collecting centroids $\bar{\mathbf{y}}_{\mathbf{c}} = (\bar{y}_{c1}, \ldots, \bar{y}_{cq})$. In a complementary fashion, FKM finds low dimensional centroids such that the scores, rather than the original data, are closest to them, that is

$$\min_{A,\bar{Y}} ||\mathbf{XA} - \mathbf{U}\bar{\mathbf{Y}}||^2 = \min_{A,\bar{Y}} \sum_{i=1}^{n} \min_{c=1,\ldots,k} \sum_{j=1}^{q} (y_{ij} - \bar{y}_{cj})^2 \ . \tag{2}$$

Both RKM and FKM are built on conventional k-means that can be badly affected by the occurrence of contamination in the data (the reader is pointed to Farcomeni and Greco (2016) for a recent account on robustness issues in data reduction). In this paper, we aim at developing a robust counterpart of RKM and FKM stemming from trimmed k-means (Cuesta-Albertos et al, 1997). Let us assume that a fraction α of data points is prone to contamination and therefore discarded. The remaining part of clean objects is then used to solve the k-means optimization problem. Trimmed data are not assigned to any cluster and do not contribute to centroids estimation. The key feature is that trimming and estimation are performed simultaneously: this approach is usually referred as impartial trimming (Gordaliza, 1991). Here, in a similar fashion, it is suggested to introduce impartial trimming into problems (1) and (2). The interesting features of the proposed methodologies rely on the ability of the method, on the one hand, to detect the anomalous data and rightly assign the remaining ones to clusters and, on the other hand, to estimate clusters' centroids in the presence of outliers. These objectives are shown in the following illustrative examples. Since the complementary nature of the RKM and FKM models, we only consider the latter model both for simulated and real data.

Figure 1 displays the result of applying trimmed FKM (tFKM) to three simulated datasets of size 400×8. All the panels display the scores lying in a two dimensional subspace, where the data are assumed to be clustered. A sample of 320 genuine scores has been drawn from a mixture of three bivariate normal with standardized components and null correlation, that is we are dealing with spherical clusters. Then 80 outliers were added, that are simulated from a different random mechanism. Genuine data have been generated according to the scheme described in Timmerman et al (2010). Three different scenarios have been considered. In the top left panel, the anomalous data have been randomly generated from a bivariate normal centered at the mean of the centroids corresponding to the three original groups with a dispersion large enough to produce both inner and outer contaminations. In the top middle panel, outliers are clustered to form a well separated group from the genuine observations. In the top right panel, outliers are clustered along some linear structures. In all scenarios, we observe the capability of the proposed method both in detecting outliers and adapting to the true underlying clustering structure. On the contrary, in the first and third data configuration, the standard procedure allocates outliers in the three clusters leading to biased centroids' estimates, inflated within-group variances and lack of separation among them, whereas in the second scenario, two well separated clusters are wrongly merged together.

2 Trimmed RKM and FKM

The optimization problems connected with trimmed RKM and FKM (tRKM and tFKM, hereafter) can be expressed as follows, respectively:

$$\min_{z\in Z}\min_{A,\bar{Y}}\sum_{i=1}^{n} z_i \min_{c=1,\ldots,k}\sum_{j=1}^{p}\left(x_{ij}-\sum_{j'=1}^{q}\bar{y}_{cj'}a_{j'j}\right)^2 \tag{3}$$

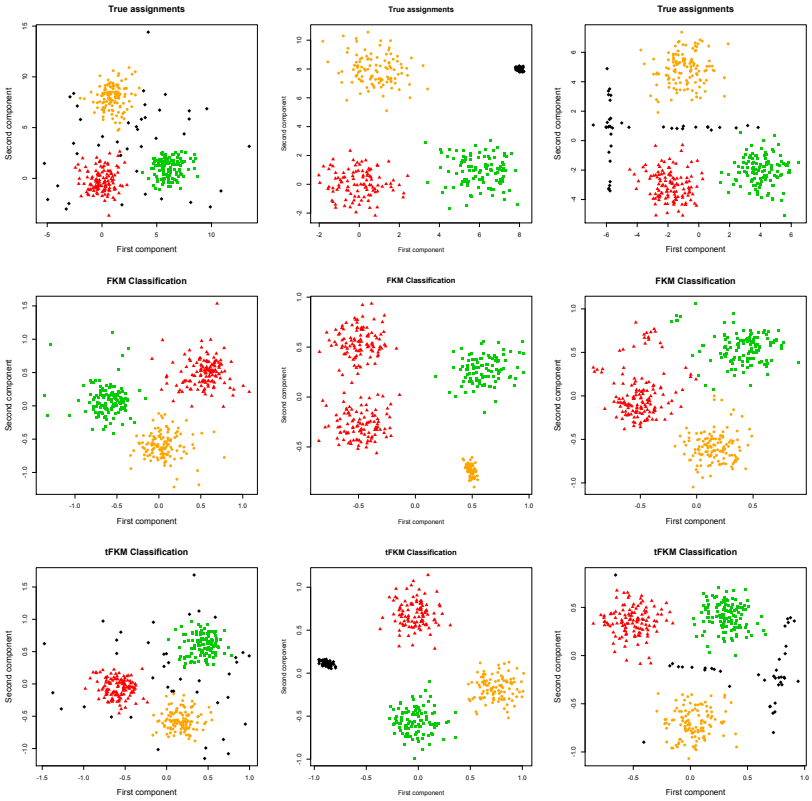
and

$$\min_{z\in Z}\min_{A,\bar{Y}}\sum_{i=1}^{n} z_i \min_{c=1,\ldots,k}\sum_{j=1}^{q}\left(y_{ij}-\bar{y}_{cj}\right)^2, \tag{4}$$

where $\mathbf{z}=(z_1,z_2,\ldots,z_n)^{\mathrm{T}}$ is a binary vector and $Z=\{z:\sum_{i=1}^{n} z_i=\lfloor n(1-\alpha)\rfloor\}$. The objective functions (3) and (4) are such that data points for which $z_i=0$ do not contribute to their minimization. It is worth noting that classical RKM and FKM are included in the above definitions as limiting cases when $\alpha=0$.

Minimization of the objective functions (3) and (4) requires an iterative algorithm characterized by the introduction of concentration steps. The concentration step (Rousseeuw and Driessen, 1999; Gallegos and Ritter, 2005; Farcomeni, 2009) is meant to detect the $n\alpha$ largest distances with respect to the current closest centroid and trim the corresponding observations. Then, the loss functions in (1) or (2) are minimized based on the observations not flagged for trimming. The final $\mathbf{U}$ obtained at convergence simultaneously identifies the optimal cluster for each

Fig. 1 Three simulated data sets with true assignments (top row), classical classification from FKM (middle row), robust classification from tFKM (bottom row). The symbol ⧫ is used to denote true outliers in the top row and trimmed observations in the other rows.



observation that is not trimmed and trimmed observations, which correspond to a constant row of zeros. Both algorithms have been implemented into the statistical environment `R` by combining the main features of the functions `cluspca` from package `clustrd` and `tkmeans` from package `tclust`.

3 Selecting the number of clusters, components and the trimming level

The selection of the number of clusters k can be pursued by paralleling the approach described in García-Escudero et al (2011): the strategy could be to display the objective function at convergence against the trimming level α for different choices of

k. Then, the number of clusters should be set equal to the minimum k for which there is no substantial improvement in the loss function when adding one group more.

In order to choose the number of components q, it is suggested to explore the quality of the fitted model by varying q from 1 to $(k-1)$. For instance, this task could be exploited by looking at the rate of explained robust variance or the adjusted Rand index. Then, the number of components can be augmented until there is no more significant increase in the selected criterion.

The selection of α could be based on the inspection of the G-statistic or of the generalized G-statistic introduced in Farcomeni (2009). Parameters' estimates or the objective function itself are monitored by varying α. Then, we select a trimming level above which the differences in parameters' estimates or in the objective function become negligible.

4 Macroeconomic data

This is a 20×6 data set, already analyzed in Vichi and Kiers (2001), concerning the macroeconomic performance of national economies in September 1999. Six main economic indicators, that measure the percentage change from the previous year, have been considered: gross domestic product (GDP), leading indicator (LI), unemployment rate (UR), interest rate (IR), trade balance (TB), net national savings (NNS). A classification of the countries into $k = 3$ groups is considered, that is expected to reflect the striking features of economic development and to take into account the differences in growth among them. The G-statistic leads to select $\alpha = 0.15$ (i.e. 3 trimmed observations). Figure 2 gives the classification resulting from FKM and tFKM. There are remarkable differences between the classical and the robust analysis, mainly due to the three outlying countries that have been detected. The cluster profiles and the raw score measurements for the outlying countries are given in Table 1. It can be seen that the three clusters are well separated, even if Cluster 2 and Cluster 3 are separated only w.r.t. the second component. The first component is mainly determined by NNS (positive sign), UR (positive sign) and TB (negative sign), whereas the second is dominated by GDP. The first cluster is composed by 4 countries that are characterized by the largest values on the first component, that is countries showing large values of NNS or UR and small values of TB. The second cluster is composed by 5 countries. It is characterized by the largest values on the second component, that is countries with large GDP. The third cluster is composed by 8 countries, that are those exhibiting the lowest growth in GDP. The outlying countries Sweden and Japan are easily explained. Sweden is well spotted on the left side and it actually shows the lowest NNS; Japan, on the contrary, is well detected along the first component since it exhibits the minimum growth in GDP. The explanation for Denmark is more complicated: it could be included in Cluster 3 but it shows an unexpected low growth in UR and NNS.

Fig. 2 Macroeconomic data: classification from FKM (left) and tFKM (right). Trimmed observations in the right panel are denoted by ⧫.

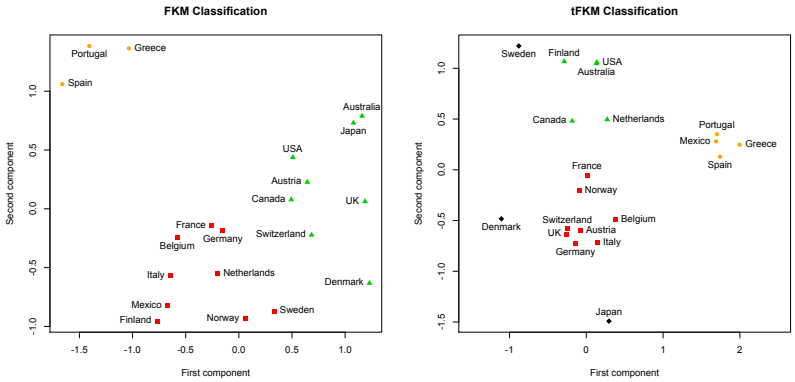


Table 1 Macroeconomic data: cluster profiles and raw scores for the outlying countries.

	Comp. 1	Comp. 2		Comp.1	Comp. 2
Cluster 1	1.782	0.252	Denmark	-1.106	-0.483
Cluster 2	0.014	0.830	Japan	0.296	-1.491
Cluster 3	-0.034	-0.500	Sweden	-0.879	1.220

References

Cuesta-Albertos J, Gordaliza A, Matrán C (1997) Trimmed *k*-means: An attempt to robustify quantizers. The Annals of Statistics 25(2):553–576

De Soete G, Carroll JD (1994) K-means clustering in a low-dimensional euclidean space. In: New approaches in classification and data analysis, pp 212–219

Farcomeni A (2009) Robust double clustering: a method based on alternating concentration steps. Journal of Classification 26(1):77–101

Farcomeni A, Greco L (2016) Robust methods for data reduction. CRC press

Gallegos M, Ritter G (2005) A robust method for cluster analysis. Annals of Statistics pp 347–380

García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2011) Exploring the number of groups in robust model-based clustering. Statistics and Computing 21(4):585–599

Gordaliza A (1991) Best approximations to random variables based on trimming procedures. Journal of Approximation Theory 64(2):162–180

Rousseeuw P, Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223

Timmerman M, Ceulemans E, Kiers HA, Vichi M (2010) Factorial and reduced k-means reconsidered. Computational Statistics & Data Analysis 54(7):1858–1871

Vichi M, Kiers H (2001) Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis 37(1):49–64