Applied Mathematics and Computation 221 (2013) 571–580

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Finding cluster centers and sizes via multinomial parameterization

Stan Lipovetsky

GfK Custom Research North America, 8401 Golden Valley Road, Minneapolis, MN 55427, USA

ARTICLE INFO

Keywords:
Clusters parameters
Nonlinear optimization
Multinomial parameterization

ABSTRACT

The clustering problem consists in dividing a data set into groups of observations that are similar within but different across. This paper presents a method for assessment the clusters centers and sizes in a non-linear least squares optimization with multinomial parameterization. The method is especially useful for large data sets as it operates on the summary statistics only. This approach also works for the problem of finding clusters' centers and sizes by the covariance matrix when the original data is not available. Estimation of the clusters centers and sizes can be followed by actual clustering. Example of application to marketing research problem is discussed.

1. Introduction

Modern cluster analysis presents a huge area of theoretical methods and applications related to pattern recognition, segmentation, data mining, and machine learning [2,5,9,10,19,20,22,24]. Numerous works are published in dozens of professional journals on various clustering problems (see [7,8,20,21] and the references within). Various works suggest different theoretical considerations of the question which data can potentially be divided into groups of observations more closely related within each group in comparison with the relations between groups [1,6,25]. The current paper considers this problem using an optimizing procedure for the objective of nonlinear approximation of the covariance matrix by the total of the outer products of the distances from the cluster centers to the total center for the data.

To keep the relations between the clusters and the total parameters for size and center, the clusters' coordinates and sizes are parameterized via multinomial shares of their relations to the total center and the sample base [15–18]. The problem reduces to a nonlinear regression model which opens up the possibility of using various criteria from regression analysis for estimation of the model quality for clustering purposes. These exploratory results can be further complemented by the actual clustering, so this approach also suggests a convenient way to obtain the clustering solutions. One of the important features of this approach consists of finding the cluster centers and sizes simply by using the variance–covariance matrix. This means that when the matrix is constructed, the algorithm actually does not depend on the observations themselves or on their number. Thus, it can be used for difficult clustering task on huge data sets from data bases and for data mining problems. This, of course, also comes in handy when the data itself is not readily available since the clusters' centers and sizes can be found by the covariance matrix only.

The paper is organized as follows. Section 2 describes the suggested method for finding cluster parameters, Section 3 considers the multinomial parameterization for this technique and criteria for the quality of cluster models. Numerical example is given in Section 4, and Section 5 summarizes. The Newton–Raphson optimizing technique for finding centers and sizes of clusters is given in Appendix A.

E-mail address: stan.lipovetsky@gfk.com

0096-3003/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amc.2013.06.098

2. Criterion for finding clusters' centers and sizes

Let X denote a data matrix of N by n order consisting of $i = 1,2,\ldots,N$ rows of observations by n variables in columns $x_1,x_2,\ldots,x_n$. The total matrix S_{tot} of second-moments is defined as the cross-product of the centered data, so the elements of this matrix are:

$$(S_{tot})_{jk} = \sum_{i=1}^{N}(x_{ij} - M_j)(x_{ik} - M_k), \tag{1}$$

where Mj denotes the mean value of each xj by the total sample. Suppose the total set of observations can be divided into K subsets of grouped observations, or clusters, and these clusters are numbered as q = 1,2,...,K. Also let each q-th cluster have Nq number of observations, so the sum of them equals the sample base:

$$N_1 + N_2 + \cdots + N_K = N. \tag{2}$$

For sizes of groups the index of cluster q will be used as the lower index.
Consider the decomposition of the cross-product (1) into the items related to the clustered data subsets with sizes (2). Such a transformation is known in the analysis of variance and can be presented as follows [12,13]:

$$\begin{aligned}\sum_{i=1}^{N}(x_{ij} - M_j)(x_{ik} - M_k) &= \sum_{q=1}^{K}\sum_{i=1}^{N_q}\left[\left(x_{ij}^q - m_j^q\right) + \left(m_j^q - M_j\right)\right]x_{ik}^q = \sum_{q=1}^{K}\sum_{i=1}^{N_q}\left(x_{ij}^q - m_j^q\right)x_{ik}^q + \sum_{q=1}^{K}\sum_{i=1}^{N_q}\left(m_j^q - M_j\right)x_{ik}^q \\ &= \sum_{q=1}^{K}\sum_{i=1}^{N_q}\left(x_{ij}^q - m_j^q\right)\left(x_{ik}^q - m_k^q\right) + \sum_{q=1}^{K}N_q\left(m_j^q - M_j\right)\left(m_k^q - M_k\right),\end{aligned} \tag{3}$$

where the upper index denotes the q-th group, so an i-th observation by a j-th variable x_{ij}^q belongs to the q-th cluster, and m_j^q is the mean value of each j-th variable within the q-th cluster. There is an evident relation between the subsets and the total means for each j-th variable:

$$N_1 m_j^1 + N_2 m_j^2 + \cdots + N_K m_j^K = NM_j, \quad j = 1,2,\ldots,n, \tag{4}$$

where both sides simply express the total by each x_j. The double sum obtained in (3) equals the pooled second moment within each cluster:

$$(S_{within})_{jk} = \sum_{q=1}^{K}\left(S_{within}^q\right)_{jk} = \sum_{q=1}^{K}\sum_{i=1}^{N_q}\left(x_{ij}^q - m_j^q\right)\left(x_{ik}^q - m_k^q\right). \tag{5}$$

The last sum in (3) corresponds to the weighted (by sub-sample sizes) second moment between the clusters' means centered by the total means of the variables:

$$(S_{between})_{jk} = \sum_{q=1}^{K}N_q\left(m_j^q - M_j\right)\left(m_k^q - M_k\right). \tag{6}$$

So we can rewrite (3) as the matrix sum:

$$S_{tot} = S_{within} + S_{between} = S_{within} + \sum_{q=1}^{K}N_q(m^q - M)(m^q - M)', \tag{7}$$

where all the matrices elements are defined by the relations (1), (5), and (6). In the right-hand side (7) we have the outer product of the vectors of distances from the centers m^q of each cluster to the total center M. Note that each vector m^q consists of the means m_j^q by all the variables, and similarly the vector M contains all the total means M_j.

One of the main ideas of the cluster analysis consists in finding such subsets of observations that the total sum of squares on the diagonal of S_{within} (5) representing the total distance within the clusters' observations to the clusters' centers is minimized [6,11,23]. For a given matrix S_{tot} (7), it corresponds to maximizing the distances between the clusters, or the total of the diagonal elements in $S_{between}$ (6). In this work we modify this idea to a formulation which permits further analytical consideration as follows. If all the observations within each q-th cluster are collapsing to one point of its center, the elements of the matrix S_{within} (5) are reaching zero.

This is the situation of perfect split when each cluster is strictly defined by one location point of the observations collapsed to it. To find such values of the clusters centers we can minimize the Euclidean norm of the whole matrix S_{within}. Or in other words, we can minimize the squared norm between the known matrix S_{tot} and the unknown matrix of the total outer product $S_{between}$ from (7), so the objective is:

$$F = \left\| S_{tot} - \sum_{q=1}^{K}N_q(m^q - M)(m^q - M)' \right\|^2 \to \min. \tag{8}$$