One-mode Additive Clustering of Multiway Data

Dirk Depril and Iven Van Mechelen

KULeuven
Tiensestraat 103
3000 Leuven, Belgium
(e-mail: dirk.depril@psy.kuleuven.ac.be
iven.vanmechelen@psy.kuleuven.ac.be)

Abstract. Given a multiway data set, in several contexts it may be desirable to obtain an overlapping clustering of one of the modes implied by the data. For this purpose a one-mode additive clustering model has been proposed, which implies a decomposition of the data into a binary membership matrix and a real-valued centroid matrix. To fit this model to a data set, a least squares loss function can be minimized. This can be achieved by means of a sequential fitting algorithm (SEFIT) as developed by Mirkin. In this presentation we will propose a new algorithm for the same model, based on an alternating least squares approach.

Keywords: Additive Clustering, Approximation Clustering, ALS algorithm.

1 Introduction

N-way N-mode data often show up in statistical practice. The simplest instance is a two-way two-mode data set. In this paper we will focus on the latter type of data, but everything can easily be extended to the N-way case.

For two-mode two-way data sets a one-mode additive clustering model has been described by several authors, including [Mirkin, 1996]. The aim of the associated data analysis is to fit this model to a data set under study (either in a least squares or in a maximum likelihood classification sense). For this purpose [Mirkin, 1990] proposed a sequential fitting (SEFIT) algorithm. However, at this moment not much information is available about its performance; moreover, as will be discussed below, this algorithm implies some conceptual problems. As a possible way out, in this paper we will present a new algorithm to estimate the same model.

The remainder of this paper is then structured as follows. In Section 2 we will describe the one-mode additive clustering model and in Section 3 we will explain the aim of the associated data analysis. In Section 4 the SEFIT algorithm will be explained and in Section 5 we will present our new algorithm. In Section 6 we present a few concluding remarks.

2 Model

In one-mode additive clustering the data matrix X is approximated by a model matrix. This model matrix M with entries m_{ij} ($i = 1, \ldots, I$, $j = 1, \ldots, J$) can be decomposed as

$$m_{ij} = \sum_{r=1}^{R} a_{ir} g_{rj}, \tag{1}$$

with R denoting total number of clusters, with a_{ir} taking values 1 or 0 and with g_{rj} real-valued. A is called the *cluster membership matrix* with entries a_{ir} indicating whether entity i belongs to cluster r ($a_{ir} = 1$) or not ($a_{ir} = 0$). One may note that apart from the binary nature, no further restrictions are imposed on the values a_{ir}, implying that the resulting clustering may be an overlapping one. The vector $\mathbf{g}_r = (g_{rj})_{j=1}^{J}$ is called the *centroid* of cluster r and the entire matrix G with entries g_{rj} is called the *centroid matrix*. Equation (1) then means that the ith row of M is obtained by summing up the centroids of the clusters to which row i belongs. Note that (1) can also be written in matrix form as

$$M = AG. \tag{2}$$

In the past, this model has been described in [Mirkin, 1990] and [Mirkin, 1996].

To illustrate the conceptual meaningfulness of the one-mode additive clustering model we may refer to the following hypothetical medical example. Consider a patients by symptom data matrix, the entries of which indicate the extent to which each of a set of patients suffers from each of a set of symptoms. In such a context, symptom strength may be attributed to underlying diseases or syndromes, that correspond to clusters of patients. Given that patients might suffer from more than one syndrome (a phenomenon called syndrome co-morbidity), in such a case an overlapping patient clustering is justified. The measured values of symptom strength can be considered additive combinations of the underlying syndrome profiles formalized by the rows of the centroid matrix G of the additive clustering model (1).

3 Aim of the data analysis

A two-way two-mode data matrix X resulting from a real experiment can always be represented by the model in (1). However, in most cases, a large number of clusters R will be needed for this. Therefore one usually looks for a model with a small value for R that fits the data well in some way.

A first way to do this is a deterministic one. In that case one assumes that $X \approx M$ and the goal of the data analysis is then to find the model M with R clusters that optimally approximates the data X according to some loss function. In this paper, the quality of the approximation will be expressed in terms of a least squares loss function:

$$L^2 = \sum_{ij}(x_{ij} - \sum_{r=1}^{R} a_{ir} g_{rj})^2, \tag{3}$$

which needs to be minimized with respect to the unknown a_{ir} and g_{rj} ($i = 1, \ldots, I$, $r = 1, \ldots, R$, $j = 1, \ldots, J$). Note that, if the matrix A is given, then the optimal G according to (3) is the least squares multiple regression estimator $(A'A)^{-1}A'X$. Note that this implies, since we have only 2^{IR} possible binary matrices A, that the solution space of (3) is finite and that therefore in principle it is possible to find the global minimum enumeratively. However, as computation time is an exponential function of the size of the data matrix, an enumerative search will quickly become infeasible. Therefore, in practice suitable algorithms or heuristics need to be developed to find the global optimum of (3).

A second approach to the data analysis is of a stochastic nature. We now assume that:

$$x_{ij} = \sum_{r=1}^{R} a_{ir} g_{rj} + e_{ij}, \tag{4}$$

where e_{ij} is an error term with $e_{ij} \overset{iid}{\text{sim}} N(0, \sigma^2)$. The goal of the data analysis then is to estimate the a_{ir}, g_{rj} and σ that maximize the log-likelihood:

$$\begin{aligned} \log \ell &= \sum_{ij} \log f(x_{ij}|A, G, \sigma) \\ &= -IJ \log \sqrt{2\pi} - IJ \log \sigma - \frac{\sum_{ij}(x_{ij} - \sum_{r=1}^{R} a_{ir} g_{rj})^2}{2\sigma^2} \end{aligned} \tag{5}$$

This can be characterized as a classification likelihood problem. In the latter type of problem, the binary entries of A are considered fixed parameters that need to be estimated, rather than realisations of latent random variables as in mixture-like models. For the estimation of a_{ir} and g_{rj} we need to minimize the sum in the numerator of the right most term in (5). This means that the stochastic approach for estimating the memberships and centroids is fully equivalent to the deterministic approach as explained above. For the estimation of σ^2 we have

$$\hat{\sigma}^2 = \frac{\sum_{ij}(x_{ij} - \sum_{r=1}^{R} \hat{a}_{ir} \hat{g}_{rj})^2}{IJ}, \tag{6}$$

where $\hat{a}_{ir}$ and $\hat{g}_{rj}$ are the maximum likelihood estimators of a_{ir} and g_{rj} respectively.

4 SEFIT

As explained in the previous section, the minimization of the loss function (3) requires suitable algorithms. In this section we will explain a first such algorithm that has been developed by [Mirkin, 1990] and that is a sequentially fitting (SEFIT) algorithm. In this algorithm the membership matrix A is estimated column-by-column meaning that one sequentially looks for new clusters. Suppose $m-1$ clusters have already been found, the mth cluster is then estimated by making use of the residual data

$$x_{ij}^{m} = x_{ij} - \sum_{r<m} a_{ir} g_{rj} \tag{7}$$

and by minimizing the function

$$\sum_{ij} (x_{ij}^{m} - a_{im} g_{mj})^2. \tag{8}$$

Given the memberships a_{im} $(i = 1, \ldots, I)$ the least squares estimates for the centroid values g_{mj} $(j = 1, \ldots, R)$ are given by

$$g_{mj} = \frac{\sum_{i=1}^{I} a_{im} x_{ij}}{\sum_{i=1}^{I} a_{im}^2}, \tag{9}$$

which is the simple mean of the elements in the cluster.

The estimation of the memberships a_{im} $(i = 1, \ldots, I)$ proceeds as follows. We start with a zero memberships column (i.e., an empty cluster) and sequentially add elements of the first mode to the cluster in a greedy way, that is, add that row that yields the biggest decrease in the loss function (8), and continue until no further decrease is obtained.

A full loop of the algorithm then goes as follows. First estimate the memberships a_{im} $(i = 1, \ldots, I)$ using the residuals x_{ij}^{m} by means of the greedy procedure explained above and next estimate the centroid values g_{jm} $(j = 1, \ldots, R)$ by means of equation (9). Denote $\mathbf{a}_m = (a_{im})$ and $\mathbf{g}_m = (g_{mj})$, calculate the outer product $\mathbf{a}_m \mathbf{g}_m$ and subtract it from x_{ij}^{m} yielding the residuals $x_{ij}^{m+1} = x_{ij}^{m} - \mathbf{a}_m \mathbf{g}_m$. This loop is repeated on x_{ij}^{m+1} and the algorithm terminates if the prespecified number of clusters R is reached.

One may note that this algorithm may only yield a local minimum rather than the global optimum of the loss function. Moreover, SEFIT might also have problems in recovering a true underlying model. We now will illustrate this latter problem with a hypothetical example. Suppose the following true structure underlies the data X:

$$M = AG = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \end{pmatrix} = \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{11} & g_{12} & g_{13} \\ g_{11}+g_{21} & g_{12}+g_{22} & g_{13}+g_{23} \\ g_{11}+g_{21} & g_{12}+g_{22} & g_{13}+g_{23} \\ g_{21} & g_{22} & g_{23} \\ g_{21} & g_{22} & g_{23} \end{pmatrix}. \tag{10}$$

Suppose now that we estimate the first cluster and that the correct membership vector $\mathbf{a}_1 = (1,1,1,1,0,0)'$ has been recovered. Then according to (9) the estimate of the corresponding centroid $\mathbf{g}_1$ reads

$$\mathbf{g}_1 = (g_{11} + g_{21}/2, \quad g_{12} + g_{22}/2, \quad g_{13} + g_{23}/2), \tag{11}$$

which is not equal to the true (g_{11}, g_{12}, g_{13}). Clearly a bias has been introduced due to the overlapping nature of the clusters and all further estimates may be influenced by this wrong estimate since in the next step of the algorithm the centroid will be subtracted from the data.

5 ALS algorithm

Our new approach to find the optimum of the loss function (3) is of the alternating least squares (ALS) type: given a membership matrix A we will look for an optimal G conditional upon the given A; given this G we will subsequently look for a new and conditionally optimal A, and so on.

The easiest part is the search for G given the memberships A since this comes down to an ordinary multivariate least squares regression problem, with a closed form solution:

$$G = (A'A)^{-1}A'X. \tag{12}$$

For the estimation of A we can use a separability property of the loss function (3), see also [Chaturvedi and Carroll, 1994]. This loss function indeed can be rewritten as follows:

$$\begin{aligned} L^2 = & \sum_j (x_{1j} - \sum_{r=1}^{R} a_{1r} g_{rj})^2 \\ & + \dots \\ & + \sum_j (x_{Ij} - \sum_{r=1}^{R} a_{Ir} g_{rj})^2. \end{aligned} \tag{13}$$

The latter means that the contribution of the ith row of A has no influence on the contributions of the other rows. As a consequence, A can be estimated row-by-row, which reduces the work to evaluating $I\,2^R$ possible memberships (instead of the full 2^{IR}).

The alternating process is repeated until there is no more decrease in the loss function. Since in each step the optimal conditional solution is found, we create a decreasing row of positive loss function values. As a consequence, this row has a limit which moreover will be reached after a finite number of iterations since there are only a finite number of possible membership matrices A. The iterative process is to be started with some initial membership matrix A, which can for instance be user specified or randomly drawn. Since in the ALS approach entire matrices are estimated rather than single columns, a bias as implied by the SEFIT strategy is avoided. Nevertheless, the ALS algorithm could yield a local optimum of the loss function (3). The only solution for this inconvenience is to use a large number of starts and retain the best solution.

6 Concluding remark

In this paper we proposed a new algorithm for finding overlapping clusters of one mode of a multiway data set. It involves an alternating least squares approach and might overcome some limitations of Mirkin's original SEFIT algorithm. To justify the latter claim, however, an extensive simulation study in which the performance of both algorithms would be compared, would be needed.

References

[Chaturvedi and Carroll, 1994]A. Chaturvedi and J.D. Carroll. An alternating combinatorial optimization approach to fitting the indclus and generalized indclus models. *Journal of Classification*, pages 155–170, 1994.

[Mirkin, 1990]B. G. Mirkin. A sequential fitting procedure for linear data analysis models. *Journal of Classification*, pages 167–195, 1990.

[Mirkin, 1996]B. G. Mirkin. *Mathematical Classification and Clustering.* Kluwer Academic Publishers, Dordrecht The Netherlands, 1996.