

A Mixture Model for Clustering Three-Way Proximity Data

Un Modello di Mistura per Classificare Dati a Tre Vie di Prossimità

Laura Bocci
Dipartimento di Sociologia e
Comunicazione
Università degli Studi di Roma
“La Sapienza”
laura.bocci@uniroma1.it

Donatella Vicari, Maurizio Vichi
Dipartimento di Statistica, Probabilità
e Statistiche Applicate
Università degli Studi di Roma
“La Sapienza”
donatella.vicari@uniroma1.it
maurizio.vichi@uniroma1.it

Riassunto: In questo lavoro viene proposta una nuova metodologia per individuare una partizione sfocata di matrici di prossimità tra oggetti in classi “omogenee” e contemporaneamente, identificare all’interno di ciascuna classe una partizione consenso degli oggetti. Il modello proposto viene stimato mediante un algoritmo di tipo ECM sotto l’ipotesi che i dati siano campionati da una mistura di densità multinormali.

Keywords: Fuzzy partition, consensus partition, finite mixture of Gaussian densities.

1. Introduction

Large data sets organized into a three-way proximity array are generally difficult to comprehend and specific techniques are necessary to extract relevant information. Classification methodologies for three-way dissimilarity data assume, as for the two-way case, a unique common underlying classification structure in the data. However, since the clustering structure often changes when different occasions in the three-way data are considered, the use of a single classification to reconstruct the taxonomic information frequently appears even more unrealistic. To overcome this problem POP (Partitions Of Partitions) has been recently proposed by Gordon & Vichi (1998). They detect a “secondary” partition of the observed three-way data and identify a consensus partition for each class of the secondary partition. Furthermore, Wedel & DeSarbo (1998) presented a methodology concerned with the estimation of ultrametric trees calibrated on subjects’ pairwise proximities using a finite mixture formulation.

This paper follows the same guidelines of research, including the idea to identify simultaneously a fuzzy partition of the set of occasions in classes of similar proximity matrices and, within each class, to fit a 2-ultrametric matrix (i.e., with at most two different off-diagonal values) that represents the best consensus partition of the objects. Wedel & DeSarbo (1998) noted in their model a large number of parameters to be estimated that causes a very high computational complexity and suggest a parsimonious representation of the within class covariance matrix using in practice an isotropic formulation. We propose a data reduction approach that induces a specific covariance structure inside each class of the fuzzy partition. The model is formalized by a constrained maximum likelihood mixed-integer problem and a numerical method, based on ECM strategy, is also described.

2. Notation

For the convenience of the reader, the terminology used in this paper is listed here:

n, r	number of objects and occasions, respectively;
$I = \{o_1, \dots, o_n\}$	set of n objects to be classified;
$\mathbf{J} = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n'$	$(n \times n)$ centering operator, where $\mathbf{1}_n$ is a vector of n ones;
$\mathbf{S} = [s_{ijh}]$	$(n \times n \times r)$ array specifying inner products between o_i and o_j , $(i, j=1, \dots, n)$, in the h^{th} occasion; the array $\mathbf{S}$ can be written as a set of r $(n \times n)$ inner product matrices i.e., $\mathbf{S} = [\mathbf{S}_1, \dots, \mathbf{S}_r]$, where $\mathbf{S}_h = [s_{ijh}: i, j=1, \dots, n]$;
c	number of classes of the fuzzy partition of the set of occasions;
$C = \{C_1, \dots, C_c\}$	fuzzy partition of the set of occasions (<i>secondary classification</i>);
g_k	number of classes of the k^{th} consensus partition P_k of I ;
$P = \{P_1, \dots, P_c\}$	set of partitions of I (<i>primary classifications</i>);
$\mathbf{M}_k = [m_{ijk}]$	$(n \times n)$ double-centered inner product 2-ultrametric matrix (i.e., with at most two different off-diagonal values) specifying the ultrametric distance between o_i and o_j ($i, j=1, \dots, n$) inside the k^{th} consensus partition P_k of I ;
$\mathbf{U}_k = [u_{ilk}]$	$(n \times g_k)$ binary matrix specifying for each o_i the membership to the l^{th} class of the k^{th} consensus partition P_k of I , i.e. $u_{ilk} = 1$ if o_i belongs to the l^{th} class of P_k , $u_{ilk} = 0$ otherwise;
$\mathbf{s}_h, \mathbf{m}_k$	$(n^2 \times 1)$ column vectors formed by column vectorizing, respectively, matrices $\mathbf{S}_h$ and $\mathbf{M}_k$;
$\underline{\mathbf{S}} = [\mathbf{s}_1, \dots, \mathbf{s}_r]$	$(n^2 \times r)$ matrix whose columns are the r vectors $\mathbf{s}_h$ ($h=1, \dots, r$).

In this paper we will fit a taxonomic structure to $\mathbf{S}$ in order to detect an optimal fuzzy partition of the set of occasions into c classes $C_1, \dots, C_c$, where the h^{th} occasion belongs to class C_k with probability τ_{hk} . The model also identifies a consensus partition P_k of the set I , associated with each class C_k ($k = 1, \dots, c$).

3. The Mixture Model and the Algorithm

The model here proposed will be applied on scalar products between objects, which can be easily computed by double centering each dissimilarity matrix $\mathbf{D}_h$ as $\mathbf{S}_h = -1/2 \mathbf{J} \mathbf{D}_h \mathbf{J}$ or directly from a set of data matrices.

Matrices $\mathbf{S}_h$ will be classified into c classes (*secondary classification*), each represented by a partition of objects (*primary classification*) specified by $\mathbf{M}_k$.

Let $\mathbf{s}_h$ be a n^2 random vector sampled from a population which consists of c classes in proportions $\pi_1, \pi_2, \dots, \pi_c$, and $\mathbf{s}_h$ is assumed to be normally distributed in the k^{th} class with mean vector $\mathbf{m}_k$ and $(n^2 \times n^2)$ covariance matrix Σ_k .

The class-conditional mean vector $\mathbf{m}_k$, identifying the expected consensus partition within the k^{th} class of the fuzzy partition of the set of occasions, is defined as

$$\mathbf{m}_k = \text{vec} \left(\frac{1}{2} \mathbf{b}_k \mathbf{J} \mathbf{U}_k \mathbf{U}_k' \mathbf{J} + \frac{1}{2} \mathbf{a}_k \mathbf{J} [\mathbf{I}_n - \mathbf{U}_k \mathbf{U}_k'] \mathbf{J} \right) \quad (1)$$

where $\mathbf{a}_k$ and $\mathbf{b}_k$ are the two values ($0 < \mathbf{a}_k \leq \mathbf{b}_k$) which, together with the membership matrix $\mathbf{U}_k$, define completely the partition P_k of I (i.e., $\mathbf{M}_k$).

We provide, given the k^{th} class, a local dimensionality reduction method by transforming the observation $\mathbf{s}_h$ as

$$\mathbf{s}_h = \mathbf{A}_k \mathbf{y}_{h|k} \quad (h = 1, \dots, r)$$

where $\mathbf{y}_{h|k}$ is a m -dimensional vector ($m < n^2$) of latent variables (factors) and $\mathbf{A}_k$ is a $(n^2 \times m)$ column orthonormal matrix of factor loadings.

Thus, the density of $\mathbf{y}_{h|k}$ in the k^{th} class is given by

$$f_k(\mathbf{y}_{h|k}; \mathbf{A}_k' \mathbf{m}_k, \mathbf{V}_k) = (2\pi)^{-m/2} |\mathbf{V}_k|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{s}_h - \mathbf{m}_k)' \mathbf{A}_k \mathbf{V}_k^{-1} \mathbf{A}_k' (\mathbf{s}_h - \mathbf{m}_k)\right\} \quad (2)$$

and, unconditionally, the density of $\mathbf{y}_h$ is the finite mixture

$$f(\mathbf{y}_h; \mathbf{A}_1' \mathbf{m}_1, \dots, \mathbf{A}_c' \mathbf{m}_c, \mathbf{V}_1, \dots, \mathbf{V}_c) = \sum_{k=1}^c \pi_k f_k(\mathbf{y}_{h|k}; \mathbf{A}_k' \mathbf{m}_k, \mathbf{V}_k) \quad (3)$$

where $\pi_k \geq 0$ ($k = 1, \dots, c$) and $\sum_{k=1}^c \pi_k = 1$.

It is worth noting that when the rank of Σ_k is m ($m < n^2$), the best rank- m approximation to Σ_k is obtained by using its rank- m eigendecomposition, $\Sigma_k = \mathbf{A}_k \mathbf{V}_k \mathbf{A}_k'$, where $\mathbf{A}_k$ is just the loadings matrix defined above and $\mathbf{V}_k$ is the $(m \times m)$ diagonal covariance matrix of $\mathbf{y}_{h|k}$.

Thus, mixture model (3) is also useful in modeling high-dimensional data by mixture of normal components, particularly in situations where there are limited data.

Given a sample $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_r$ of independent and identically distributed observations, it can be easily shown that maximizing the log-likelihood of model (3) is equivalent to maximize the complete data log-likelihood (Basford & McLachlan, 1988)

$$\ell^*(\mathfrak{g}) = \sum_{h=1}^r \sum_{k=1}^c z_{hk} \{\log \pi_k + \log f_k(\mathbf{y}_{h|k}; \mathbf{A}_k' \mathbf{m}_k, \mathbf{V}_k)\} \quad (4)$$

subject to

$$\sum_{k=1}^c \pi_k = 1; \pi_k \geq 0; 0 < a_k \leq b_k; \mathbf{U}_k = [u_{ilk} : u_{ilk} \in \{0,1\} \ i = 1, \dots, n \ l = 1, \dots, g_k] \ k = 1, \dots, c$$

where $\mathfrak{g}$ is the set of parameters of the overall mixture model and where z_{hk} ($h=1, \dots, r$ and $k=1, \dots, c$) is an unobservable component-label variable defined to be 1 or 0 whether or not $\mathbf{s}_h$ did arise from the k^{th} class of the fuzzy partition.

The constrained problem of maximizing (4) can be faced by using the following algorithm. Given initial feasible values for parameters $\mathbf{A}_k, \mathbf{V}_k, a_k, b_k, \mathbf{U}_k$, and π_k ($k=1, \dots, c$), function (4) is maximized with respect to each set of parameters conditionally upon the others. At each step, the estimated values provide monotonically increasing values of the log-likelihood function (4), so that the convergence of the algorithm to at least a local optimum is guaranteed, being (4) bounded from above. This computational procedure can be considered as an ECM type algorithm.

The global optimal solutions of (4) on the $(t+1)$ th iteration for each set of parameters, given the current estimates, are:

$$\tau_{hk}^{(t+1)} = \frac{\pi_k^{(t)} f_k \left(\mathbf{y}_{h|k}; \mathbf{A}_k^{(t)'} \mathbf{m}_k^{(t)}, \mathbf{V}_k^{(t)} \right)}{\sum_{j=1}^c \pi_j^{(t)} f_j \left(\mathbf{y}_{h|j}; \mathbf{A}_j^{(t)'} \mathbf{m}_j^{(t)}, \mathbf{V}_j^{(t)} \right)} \quad ; \quad \pi_k^{(t+1)} = \frac{tr(diag(\boldsymbol{\tau}_k^{(t+1)}))}{r} \quad h=1, \dots, r \quad k=1, \dots, c$$

$$\begin{aligned} \mathbf{a}_k^{(t+1)} &= 2 \left\{ \left[vec \left(\mathbf{J} - \mathbf{J} \mathbf{U}_k^{(t+1)} \mathbf{U}_k^{(t+1)'} \mathbf{J} \right) \right] \boldsymbol{\Sigma}_k^{(t)-1} \left[vec \left(\mathbf{J} - \mathbf{J} \mathbf{U}_k^{(t+1)} \mathbf{U}_k^{(t+1)'} \mathbf{J} \right) \right] \right\}^{-1} \left[vec \left(\mathbf{J} - \mathbf{J} \mathbf{U}_k^{(t+1)} \mathbf{U}_k^{(t+1)'} \mathbf{J} \right) \right]' \times \\ &\quad \boldsymbol{\Sigma}_k^{(t)-1} \left[\underline{\mathbf{S}} - \frac{1}{2} \mathbf{b}_k^{(t)} vec \left(\mathbf{J} \mathbf{U}_k^{(t+1)} \mathbf{U}_k^{(t+1)'} \mathbf{J} \right) \mathbf{1}_r' \right] diag(\boldsymbol{\tau}_k^{(t+1)}) \mathbf{1}_r [\mathbf{1}_r' diag(\boldsymbol{\tau}_k^{(t+1)}) \mathbf{1}_r]^{-1} \\ \mathbf{b}_k^{(t+1)} &= 2 \left\{ \left[vec \left(\mathbf{J} \mathbf{U}_k^{(t+1)} \mathbf{U}_k^{(t+1)'} \mathbf{J} \right) \right] \boldsymbol{\Sigma}_k^{(t)-1} \left[vec \left(\mathbf{J} \mathbf{U}_k^{(t+1)} \mathbf{U}_k^{(t+1)'} \mathbf{J} \right) \right] \right\}^{-1} \left[vec \left(\mathbf{J} \mathbf{U}_k^{(t+1)} \mathbf{U}_k^{(t+1)'} \mathbf{J} \right) \right]' \times \\ &\quad \boldsymbol{\Sigma}_k^{(t)-1} \left[\underline{\mathbf{S}} - \frac{1}{2} \mathbf{a}_k^{(t)} \left[vec \left(\mathbf{J} - \mathbf{J} \mathbf{U}_k^{(t+1)} \mathbf{U}_k^{(t+1)'} \mathbf{J} \right) \mathbf{1}_r' \right] diag(\boldsymbol{\tau}_k^{(t+1)}) \mathbf{1}_r [\mathbf{1}_r' diag(\boldsymbol{\tau}_k^{(t+1)}) \mathbf{1}_r]^{-1} \right] \end{aligned}$$

where $diag(\boldsymbol{\tau}_k^{(t+1)})$ is the diagonal matrix with elements $\tau_{hk}^{(t+1)}$ for $h=1, \dots, r$. The maximization of $\ell^*(\mathfrak{g})$ with respect to $\mathbf{U}_k$, given the current estimate of $\mathbf{A}_k^{(t)}$, $\mathbf{V}_k^{(t)}$, $\mathbf{a}_k^{(t)}$, $\mathbf{b}_k^{(t)}$, $\pi_k^{(t+1)}$, $\boldsymbol{\tau}_{hk}^{(t+1)}$ ($h=1, \dots, r$ and $k=1, \dots, c$) is solved for the different rows of $\mathbf{U}_k^{(t+1)}$ independently by taking $u_{ilk}^{(t+1)} = 1$, if $\ell^*(\mathfrak{g}, [u_{ilk}^{(t+1)}]) = \max \{ \ell^*(\mathfrak{g}, [u_{ilv}^{(t+1)}]) : v = 1, \dots, g_k \}$ and $u_{ilk}^{(t+1)} = 0$ otherwise. Matrix $\mathbf{A}_k^{(t+1)}$ contains the first m eigenvectors of $(\underline{\mathbf{S}} - \mathbf{m}_k^{(t+1)} \mathbf{1}_r') diag(\boldsymbol{\tau}_k^{(t+1)}) (\underline{\mathbf{S}} - \mathbf{m}_k^{(t+1)} \mathbf{1}_r') / tr(diag(\boldsymbol{\tau}_k^{(t+1)}))$ and $\mathbf{V}_k^{(t+1)}$ is the diagonal matrix of the corresponding first m eigenvalues.

4. Discussion

In this paper a model for simultaneous classification of occasions and objects is proposed. The aim is to obtain a secondary classification (fuzzy partition) of the observed three-way data assumed to be sampled by a mixture of multivariate normal densities that represent the classes of the fuzzy partition. The proposed reduction approach is parsimonious and particularly useful in situations where there are high-dimensional limited data. Examples of the methodology will be given in a successive extended version of this paper to demonstrate practical uses of the method.

References

- Basford, K.E., MacLachlan, G.J. (1988) *Mixture Models. Inference and Application to Clustering*, Marcel Dekker Inc., New York and Basel
- Gordon, A.D., Vichi, M. (1998) Partitions of Partitions, *Journal of Classification*, 15, 265-285.
- Wedel, M., DeSarbo, W.S. (1998) Mixture of (Constrained) Ultrametric Trees, *Psychometrika*, 63, 4, 419-443.