

A Two-Way Model for Simultaneous Reduction and Classification

Un Modello per la Classificazione su Sottospazi di Dimensione Ridotta

Roberto Rocci

Dept. SEGeS, University of Molise
Via De Sanctis, s.n.c. - 86100, Campobasso
E-mail: roberto.rocci@uniroma1.it

Maurizio Vichi

Dept. of Statistics, Probability and Applied
Statistics
University "La Sapienza"
P.le A.Moro, 5 - 00185, Roma.
E-mail: maurizio.vichi@uniroma1.it

Riassunto: In questo lavoro viene presentato un nuovo modello per l'analisi esplorativa di matrici di dati del tipo unità per variabili. L'analisi è condotta operando una classificazione delle unità sotto l'assunzione che il campione estratto sia stato generato da una mistura di Normali multivariate. I parametri della mistura vengono stimati con il metodo della massima verosimiglianza sotto l'ipotesi che i vettori delle medie di ciascuna componente appartengano ad uno stesso sottospazio di dimensione ridotta. Questo permette di identificare dei fattori latenti in grado di spiegare la variabilità tra i gruppi e quindi le variabili manifeste aventi il maggior potere discriminante tra i gruppi. Viene inoltre proposto un algoritmo di tipo ECM per la stima di massima verosimiglianza dei parametri del modello.

Keywords: Classification, Principal components, Mixture models, Maximum likelihood.

1. Introduction

In a cluster analysis context, principal component analysis is frequently used to reduce the number of the variables, then on the main components a cluster analysis algorithm for grouping units into homogeneous classes is applied. This procedure, named "*tandem analysis*", provides a means of projecting multivariate observed units into a lower-dimensional space so as to offer an adequate description of the data in a few components and homogeneous clusters. Though this is by far the most common procedure used in practice, tandem analysis has been criticized by several authors (e.g., Chang, 1983; De Soete & Carroll, 1994). This because principal components account for the maximal amount of the total variance of the data but important information about the successive classification can be lost, even when a large number of components is extracted. To overcome this problem, some authors proposed methodologies for simultaneous classification and reduction of the data (see for example: De Soete & Carroll, 1994; Vichi & Kiers, 2001).

The main purpose of this paper is to formulate an alternative methodology using a mixture model as a form to describe data suspected to consist of relatively distinct groups. It is assumed that the observed data are sampled from a finite mixture of Gaussians, where the components correspond to underlying groups. The mean vectors

of the Gaussian components are constrained to lie in a reduced subspace to identify the latent factors that best explain the between variability of the data.

The plan of the paper is as follows. In section 2 the model is presented. In section 3 an algorithm to compute the maximum likelihood estimates of the model parameters is derived. A brief final discussion follows in section 4.

2. The Mixture Model

Finite mixture densities are a family of functions of the form

$$f(\mathbf{x}_i; \Theta) = \sum_{g=1}^G p_g f_g(\mathbf{x}_i; \vartheta_g). \quad (1)$$

where $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iJ}]'$ is a random vector of J centered variables sampled from a population parametrized by $\Theta = \{\vartheta_1, \vartheta_2, \dots, \vartheta_G, p_1, p_2, \dots, p_G\}$, which consists of G groups, in proportions $p_1, p_2, \dots, p_G$.

The density of $\mathbf{x}_i$ in the g^{th} group is given by

$$\begin{aligned} f_g(\mathbf{x}_i; \vartheta_g) &= f_g(\mathbf{x}_i; \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}) \\ &= (2\pi)^{-\frac{J}{2}} |\boldsymbol{\Sigma}_g|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g)\right\} \end{aligned} \quad (2)$$

where $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are the mean vector and the covariance matrix of the g^{th} group, respectively.

The mean vectors of the groups are assumed to follow a bilinear model. In particular, the mean μ_{jg} of the j^{th} variable in the g^{th} group is related with a reduced set of factors according to the following linear combination of Q ($< J$) latent variables

$$\mu_{jg} = \sum_{q=1}^Q b_{jq} y_{qg}, \quad (3)$$

where y_{qg} represents the score of the q^{th} latent variable in the g^{th} group, while b_{jq} is the loading for variable j on the q^{th} latent variable. Therefore, the mean vector of each group, which lies originally into a J dimensional space, is reproduced by a reduced number Q of latent variables according to the bilinear model (3). Furthermore, inspection of the loadings reveals which latent variables most contribute to reconstruct mean vectors. Equation (3) can be expressed in matrix form, with obvious notation, as

$$\boldsymbol{\mu}_g = \mathbf{B} \mathbf{y}_g. \quad (4)$$

It should be noted that this model is not identified. In fact, the loading matrix $\mathbf{B}$ can be rotated without affecting the model, provided that the latent vectors $\mathbf{y}_g$ are counter-rotated by the inverse transformation. It implies that only the subspace spanned by the columns of $\mathbf{B}$ is identified, as in the exploratory factor analysis model.

3. Algorithm

Assuming that the I observations are independent and identically distributed, the log-likelihood of mixture model (2) can be written as

$$L(\Theta) = \sum_{i=1}^I \lg \left(\sum_{g=1}^G p_g f_{ig} \right), \quad (5)$$

where $f_{ig} = f_g(\mathbf{x}_i; \vartheta_g)$ and ϑ is the set of parameters for the overall mixture model. It can be easily shown (Hathaway, 1986) that maximizing (5) is equivalent to maximize

$$\ell(\Theta) = \sum_{ig} u_{ig} \lg(p_g f_{ig}) - \sum_{ig} u_{ig} \lg(u_{ig}), \quad (6)$$

subject to the constraint $\sum_g u_{ig} = 1$ ($i=1,2,\dots,I$). The algorithm can be formulated as a grouped version of the coordinate ascent method where function (6) is iteratively maximized with respect to a group of parameters conditionally upon the others.

The fundamental steps of the algorithm can be described as follows.

a) Update $\mathbf{U}=[u_{ig}]$: It can be easily shown that (6) obtains a maximum when

$$u_{ig} = \frac{p_g f_{ig}}{\sum_h p_h f_{ih}}, \quad i=1,2,\dots,I; \quad g=1,2,\dots, G, \quad (7)$$

which is the posterior probability that unit i belongs to group g .

b) Update $\mathbf{p}=[p_g]$: Function (6) obtains a maximum with respect to $\mathbf{p}$ when

$$p_g = \frac{1}{I} \sum_i u_{ig}, \quad g=1,2,\dots,G. \quad (8)$$

c) Update $\Sigma=[\Sigma_g]$: function (6) is maximized with respect to Σ_g when

$$\Sigma_g = \frac{1}{I p_g} \sum_i u_{ig} (\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)', \quad g=1,2,\dots,G. \quad (9)$$

d) Update $\mathbf{B}$: we can rewrite the objective function (6) as

$$\ell(\Theta) = \sum_{ig} u_{ig} \lg(f_{ig}) + c = \sum_{ig} u_{ig} \left\{ -\frac{1}{2} [(\mathbf{x}_i - \mathbf{B}\mathbf{y}_g)' \Sigma_g^{-1} (\mathbf{x}_i - \mathbf{B}\mathbf{y}_g)] \right\} + c' \quad (10)$$

where c and c' are additive constants which are independent from $\mathbf{B}$. Differentiating (10) with respect to $\mathbf{B}$ and setting the derivatives to zero, we obtain

$$\text{vec}(\mathbf{B}) = \left\{ \sum_g I_{p_g} [(\mathbf{y}_g \mathbf{y}_g') \otimes \boldsymbol{\Sigma}_g^{-1}] \right\}^{-1} \text{vec} \left(\sum_{ig} u_{ig} \boldsymbol{\Sigma}_g^{-1} \mathbf{x}_i \mathbf{y}_g' \right). \quad (11)$$

e) Update $\mathbf{y}_g$: It can be easily shown that the update of $\mathbf{y}_g$ is

$$\mathbf{y}_g = (\mathbf{B}' \boldsymbol{\Sigma}_g^{-1} \mathbf{B})^{-1} \mathbf{B}' \boldsymbol{\Sigma}_g^{-1} \left(\sum_i u_{ig} \mathbf{x}_i \right) (I_{p_g})^{-1}. \quad (12)$$

On the basis of these fundamental steps, the algorithm can be described as follows:

- Step 0. choose initial values for the parameters;
- Step 1. update in turn $\mathbf{p}$, $\boldsymbol{\Sigma}_g$ ($g=1, 2, \dots, G$), $\mathbf{B}$, $\mathbf{Y}$ and $\mathbf{U}$;
- Step 2. compute (6) for the current parameter values. If the function increase is larger than a fixed threshold, iterate once more according to Step 1. Otherwise, the process has converged.

It is interesting to note that the above computational procedure can be considered as an ECM type algorithm (Meng & Rubin, 1993).

4. Discussion

In this paper a new model for simultaneous partitioning of objects and reduction of variables is proposed. The main assumption is that the observed data are sampled by a mixture of Gaussian densities that represent clusters in the population. The mean vectors are formalized by a parsimonious bilinear model in order to identify the factors that explain the between variability. It is worth noting that the proposed model can be considered as a fuzzy version of the model suggested by De Soete & Carroll (1994), where each group has in addition a specific Mahalanobis metric. It is also interesting to note that a similar mixture approach has been used by Hastie and Tibshirani (1996) in a discriminant analysis context.

References

- Chang, W. (1983): On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32, 267-275.
- De Soete, G., Carroll, J.D. (1994): K-means Clustering in a Low-dimensional Euclidean Space. In: E. Diday et al. (Eds), *New Approaches in Classification and Data Analysis*. Springer, Heidelberg, 212-219.
- Hastie, T., Tibshirani, R. (1996): Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, series B*, 58, 155-176
- Hathaway, R.J. (1986): Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters*, 4, 53-56.
- Meng, X.L., Rubin, D.B. (1993): Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80, 267-278.
- Vichi, M., Kiers, H.A.L. (2001): Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37, 49-64.