MIXMOD : a software for model-based classification with continuous and categorical data

Christophe Biernacki, Gilles Celeux, Gérard Govaert, Florent Langrognet

MIXMOD: a software for model-based classification with continuous and categorical data

C. Biernacki, G. Celeux, G. Govaert and F. Langrognet

Abstract The MIXMOD (MIXture MODeling) program fits mixture models to a given data set for the purposes of density estimation, clustering or discriminant analysis. A large variety of algorithms to estimate the mixture parameters are proposed (EM, Classification EM, Stochastic EM), and it is possible to combine these to yield different strategies for obtaining a sensible maximum for the likelihood (or complete-data likelihood) function. MIXMOD is currently intended to be used for multivariate Gaussian mixtures and also for latent class models, respectively devoted to continuous and categorical data. In both situations, numerous meaninful and parsimonious models are proposed. Moreover, different information criteria for choosing a parsimonious model (the number of mixture components, for instance) are included, their suitability depending on the particular perspective (cluster analysis or discriminant analysis). Written in C++, MIXMOD is interfaced with SCILAB and MATLAB. The program, the statistical documentation and the user guide are available on the internet at the following address:
`http://www-math.univ-fcomte.fr/mixmod/index.php`

C. Biernacki
Université Lille 1 & CNRS, Villeneuve d'Ascq (France), e-mail: biernack@math.univ-lille1.fr

G. Celeux
INRIA, Orsay e-mail: Gilles.Celeux@inria.fr

G. Govaert
UTC & CNRS, Compigne (France) e-mail: Gerard.Govaert@utc.fr

F. Langrognet
CNRS, Besançon (France) e-mail: Florent.Langrognet@univ-fcomte.fr

1 Overview of MIXMOD

Because of their high flexibility, finite mixture distributions are become a very popular approach to model a wide variety of random phenomena. In particular, it is recognized as being a powerful tool for density estimation, clustering and discriminant analysis. Consequently, fields which are potentially concerned by the mixture modelling approach are extremely varied, including astronomy, biology, genetics, economics, etc. Thus, softwares implementing the most recent evolutions in model-based cluster and discriminant analysis are welcome for various categories of people as researchers, engineers and teachers. mixmod is a software having for goal to meet these particular needs.

MIXMOD is publicly available under the GPL license and is distributed for different platforms (Linux, Unix, Windows). It is an object-oriented package built around C++ language but it is interfaced with the widely used mathe- matical softwares Matlab and Scilab. It was developed jointly by INRIA & CNRS, by several laboratories of mathematics (university of Besançon and of Lille 1), and by the Heudiasyc laboratory of Compiègne.

In its present version, MIXMOD proposes multivariate Gaussian mixture models for continuous data and also multivariate latent class models for categorical data. The main features of the present version of the software are the following:

- three levels of use from the beginner to the expert;
- fourteen geometrically meaningful Gaussian mixture models from different variance matrices parameterizations;
- five multinomial meaningful models;
- estimation of mixture parameters with EM and EM-like algorithms, pro- vided with different initialization strategies and possibility of combining such algorithms;
- possibility of partial labeling of individuals (semi-supervised situation);
- criteria to select a model which depends on the cluster or the discriminant analysis purpose;
- numerous displays including densities, iso-densities, discriminant rules, ob- servations, labels, etc. in canonical or PCA (Principal Component Analysis) axes and for several dimensions (1D, 2D and 3D).

2 MIXMOD for continuous data

Cluster analysis is concerned with discovering a group structure in a n by d matrix $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where $\mathbf{x}_i$ is an individual of $\Re^d$. Consequently, the structure to be discovered by clustering is typically a partition of $\mathbf{x}$ into K groups defined by the labels $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$, with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$, $z_{ik} = 1$ or 0, according to the fact that $\mathbf{x}_i$ belongs to the kth class or not. In the Gaussian mixture model, each $\mathbf{x}_i$ is assumed to arise independently from a mixture with density

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k h(\mathbf{x}_i|\mu_k, \Sigma_k) \quad (1)$$

where p_k is the mixing proportion ($0 < p_k < 1$ for all $k = 1, ..., K$ and $p_1 + ... + p_K = 1$) of the kth component and $h(\cdot|\mu_k, \Sigma_k)$ denotes the d-dimensional Gaussian density with mean μ_k and variance matrix Σ_k. The vector of mixture parameters is also denoted by $\theta = (p_1, \ldots, p_{K-1}, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$.

In MIXMOD, a partition of the data can directly be derived from the maximum likelihood estimates $\hat{\theta}$ of the mixture parameters obtained, for instance, by the EM [10] or the SEM [6] algorithm, by assigning each $\mathbf{x}_i$ to the component providing the largest conditional probability that $\mathbf{x}_i$ arises from it using a MAP (Maximum A Posteriori) principle:

$$\hat{z}_{ik} = \begin{cases} 1 \text{ if } k = \arg\max_{\ell=1\ldots,K} t_\ell(\mathbf{x}_i|\hat{\theta}) \\ 0 \text{ if not} \end{cases} \text{ where } t_k(\mathbf{x}_i|\hat{\theta}) = \frac{\hat{p}_k h(\mathbf{x}_i|\hat{\mu}_k, \hat{\Sigma}_k)}{\sum_\ell \hat{p}_\ell h(\mathbf{x}_i|\hat{\mu}_\ell, \hat{\Sigma}_\ell)}. \quad (2)$$

Alternatively, an estimate $\hat{\theta}$ can be retained as being the maximum *completed* likelihood estimate obtained, for instance, by the CEM algorithm [8]. Many strategies for using and combining these algorithms are available in MIXMOD for helping to improve the optimisation process [4].

Following Celeux and Govaert [9], the software integrates also a parameterization of the variance matrices of the mixture components consisting of expressing the variance matrix Σ_k in terms of its eigenvalue decomposition $\Sigma_k = \lambda_k D_k A_k D_k'$ where $\lambda_k = |\Sigma_k|^{1/d}$, D_k is the matrix of eigenvectors of Σ_k and A_k is a diagonal matrix, such that $|A_k| = 1$, with the normalized eigenvalues of Σ_k on the diagonal in a decreasing order. The parameter λ_k determines the *volume* of the kth cluster, D_k its *orientation* and A_k its *shape*. By allowing some but not all of these quantities to vary between clusters, we obtain parsimonious and easily interpreted models which are appropriate to describe various clustering situations, including standard methods as *k*means for instance.

It is of high interest to automatically select one of these Gaussian models and/or the number K of mixture components. However, choosing a sensible mixture model is highly dependent of the modelling purpose. Three criteria are available in an unsupervised setting: BIC [13], ICL [3] and NEC [2]. In a density estimation perspective, BIC must be preferred. But in a cluster analysis perspective, ICL and NEC can provide more parsimonious answers. Nevertheless, NEC is essentially devoted to choose the number of mixture components K, rather that the model parameterization.

When the labels $\mathbf{z}$ are known, discriminant analysis is concerned. In this situation, the aim is to estimate the group $\mathbf{z}_{n+1}$ of any new individual $\mathbf{x}_{n+1}$ of $\Re^d$ with unknown label. In MIXMOD, the n couples $(\mathbf{x}_i, \mathbf{z}_i)$,..., $(\mathbf{x}_n, \mathbf{z}_n)$ are supposed to be n i.i.d. realizations of the following joint distribution:

$$f(\mathbf{x}_i, \mathbf{z}_i|\theta) = \prod_{k=1}^{K} p_k^{z_{ik}} [h(\mathbf{x}_i|\mu_k, \Sigma_k)]^{z_{ik}}. \quad (3)$$

An estimate $\hat{\theta}$ of θ is obtained by the maximum likelihood method (from complete data $(\mathbf{x}, \mathbf{z})$ and, then, the MAP principle is involved to class any new individual $\mathbf{x}_{n+1}$. In the software, two criteria are proposed in this supervised setting for selecting one of the previous Gaussain models: BIC and cross-validation.

3 MIXMOD for categorical data

We consider now that data are n objects described by d categorical variables, with respective number of categories $m_1, \ldots, m_d$. The data can be represented by n binary vectors $\mathbf{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$ $(i = 1, \ldots, n)$ where $x_i^{jh} = 1$ if the object i belongs to the category h of the variable j and 0 otherwise. Denoting $m = \sum_{j=1}^{d} m_j$ the total number of categories, the data are defined by the matrix $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with n rows and m columns. Binary data can be seen as a particular case of categorical data with d dichotomous variables, i.e. $m_j = 2$ for any $j = 1, \ldots, d$.

The latent class model assumes that the d ordinal variables are independent given the latent variable. Formulated in mixture terms [11], each $\mathbf{x}_i$ arises independently from a mixture of multivariate multinomial distributions defined by

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k h(\mathbf{x}_i|\alpha_k) \text{ where } \quad h(\mathbf{x}_i|\alpha_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}} \quad (4)$$

with $\alpha_k = (\alpha_k^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$. We recognize the product of d conditionally independent multinomial distributions of parameters α_k^j. In this situation, the mixture parameter is denoted by $\theta = (p_1, \ldots, p_{K-1}, \alpha_1, \ldots, \alpha_K)$. This model may present problems of identifiability [12] but most situations of interest are identified. sc mixmod proposes four parsimonious models declined from the previous one by following extension of the parameterization of Bernoulli distributions used by [7] for clustering and also by [1] for kernel discriminant analysis.

All models, algorithms an criteria presented previously in the Gaussian situation are also implanted in MIXMOD for the latent class model. Both the clustering and the supervised classification purposes can be involved by the user in this context.

4 An extension coming soon for high dimensional data

In order to deal with high dimensional data, Mixture of Factor Analyzers have been considered by several authors including [5]. In MIXMOD, a family of eight Gaussian

mixture models introduced by [5] are being implemented for discriminant analysis in high dimensional spaces.

References

1. Aitchison, J. and Aitken, C.G. (1976). Multivariate binaery discrimination by the kernel method, *Biometrika*, **63**, 413–420.
2. Biernacki, C. Celeux, G. et Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of components arising from a mixture, *Pattern Recognition letters*, **20**, 267–272.
3. Biernacki, C. Celeux, G. et Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 7, 719–725.
4. Biernacki, C., Celeux, G., Govaert, G. and Langrognet, F. (2006). Model-bases cluster and discriminant analysis with MIXMOD software. *Computational Statistics & Data Analysis*, **51**, 2, 587–600.
5. Bouveyron, C., Girard, G. and Schmid, C. (2007). High dimensional discriminant analysis, *Communications in Statistics: Theory and Methods*, **36**, 2607–2623.
6. Celeux, G. et Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Statis. Quaterly*, **2**, 73–82.
7. Celeux, G. et Govaert, G. (1991). Clustering Criteria for Discrete Data and Latent Class Models, *Journal of Classification*, **8**, 157–176.
8. Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, **14**, 315–332.
9. Celeux, G. et Govaert, G. (1995). Gaussian Parsimonious Clustering Models. *Pattern Recognition*, **28**, 781–793.
10. Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B*, **39**, 1–38.
11. Everitt, B. (1984). *An Introduction to Latent Variable Models*. London, Chapman and Hall.
12. Goodman, L.A. (1974). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models, *Biometrika*, **61**, 215–231.
13. Schwarz, G. (1978). Estimating the Dimension of a Model, *Annals of Statistics*, **6**, 461–464.