

Identification of multiple clusterings using Gaussian mixtures

Individuazione di partizioni multiple con modelli basati su misture gaussiane

Andrea Pastore and Stefano Tonellato

Abstract In this paper we present a generalisation of the algorithm by Raftery and Dean [3], which is tailored for the selection of variables in Gaussian mixture models. Our proposal allows for the detection of multiple independent clusterings. A simulation study is provided in order to assess the performance of the algorithm for different sample sizes and different numbers of variables in the data set.

Abstract *In questo lavoro si presenta una generalizzazione dell'algoritmo introdotto da Raftery e Dean [3], mirato alla selezione di variabili in misture gaussiane nel contesto della classificazione. La nostra proposta consente di individuare molteplici classificazioni indipendenti. Si illustrano i risultati ottenuti su alcune simulazioni, allo scopo di valutare empiricamente il comportamento dell'algoritmo al variare delle dimensioni del campione e del numero di variabili nel data set.*

Key words: Unsupervised classification, Gaussian mixtures, BIC

1 Introduction

Traditionally, in cluster analysis, a unique partition of a set of objects is searched for. When dealing with high dimensional data, however, different grouping structures may be identified when looking at the data under different perspectives, and features relevant to one clustering may differ from the ones relevant under an alternative view. An excellent overview of the methods proposed for detecting multiple clusterings in the machine learning and in the data mining literature is provided by [2], who also introduce a nonparametric Bayesian model that can discover multiple clusterings in the data. Within a model-based framework, Galimberti and Soffritti [1] approach the identification of multiple partitions as a model comparison problem. In this paper, we shall indicate how a generalisation of the variable selection

Dipartimento di Economia, Università Ca' Foscari Venezia, Fondamenta S. Giobbe - Cannaregio 873, 30125 Venezia, Italy, e-mail: pastore@unive.it, stone@unive.it

method proposed by Raftery and Dean [3] can lead to the identification of multiple clusterings. We shall consider a model for the multidimensional variable Y characterised by the following structure:

$$p(Y|\theta) = p(Y^{\bar{C}_T}|Y^{C_T}, \theta_{\bar{C}_T}) \prod_{t=1}^T p_t(Y^{R_t}|\theta_t), \quad (1)$$

where Y^{R_t} , $t = 1, \dots, T$, are subsets of variables giving rise to T clusterings, and their distributions are independent finite Gaussian mixtures, $p_t(Y^{R_t})$; $Y^{C_T} = \cup_{t=1}^T Y^{R_t}$ represents the subset of all clustering variables, $Y^{\bar{C}_T} = Y \setminus Y^{C_T}$ represents the subset of Y which provides no information about clustering, and $\theta = [\theta_{\bar{C}_T}, \theta_1, \dots, \theta_T]$ indicates the vector of unknown parameters. Equation (1) implies that

$$p(Y^{C_T}|\theta) = \prod_{t=1}^T p_t(Y^{R_t}|\theta_t),$$

i.e. the clustering variables giving rise to the T distinct mixture models are independent. Each of the T mixtures can be used to partition the data into clusters by applying an approximation of the Bayes rule of classification based on the fitted parametric model [3].

2 Model selection.

2.1 Raftery and Dean method.

Raftery and Dean [3] introduced a variable selection algorithm that we now illustrate. Assume we have a dataset Y which, at any step of the variable selection algorithm, is partitioned in three subsets of variables, $Y^{(1)}$ (the subset of already selected variables), $Y^{(2)}$ (the variable being considered for inclusion into or exclusion from the set of clustering variables), and $Y^{(3)}$ (the set of the remaining variables). The decision of including $Y^{(2)}$ into (excluding $Y^{(2)}$ from) the set of clustering variables, is taken by comparing the following models:

$$M_1 : p(Y|\theta_{M_1}, M_1) = p(Y^{(3)}|Y^{(1)}, Y^{(2)}, \theta_{13})p(Y^{(2)}|Y^{(1)}, \theta_{12})p(Y^{(1)}|\theta_{11}) \quad (2)$$

$$M_2 : p(Y|\theta_{M_2}, M_2) = p(Y^{(3)}|Y^{(1)}, Y^{(2)}, \theta_{13})p(Y^{(1)}, Y^{(2)}|\theta_{21}) \quad (3)$$

where $p(Y^{(1)}|\theta_{11})$ and $p(Y^{(1)}, Y^{(2)}|\theta_{21})$ in (2) and (3) are finite Gaussian mixtures. The decision of including $Y^{(2)}$ into (excluding $Y^{(2)}$ from) $Y^{(1)}$, is taken whenever $B(M_1, M_2) < 0$ ($B(M_1, M_2) > 0$), where

$$B(M_1, M_2) = BIC(M_1) - BIC(M_2) \quad (4)$$

and $BIC(M)$, the Bayes Information Criterion for model M , is defined as

$$BIC(M) = 2 \times \log(\text{maximised likelihood}) - (\text{no. of parameters}) \times \log(n).$$

When Y is assumed to be distributed as a multivariate mixture with Gaussian components, the contribution of $p(Y^{(3)}|Y^{(1)}, Y^{(2)}, \theta_{13})$ and $p(Y^{(2)}|Y^{(1)}, \theta_{12})$ in (2) to $BIC(M_1)$ can be easily computed by regressing $Y^{(3)}$ on $Y^{(1)}$ and $Y^{(2)}$, and $Y^{(2)}$ on $Y^{(1)}$, whereas the contribution of $p(Y^{(1)}|\theta_{11})$ can be computed by estimating a mixture model for $Y^{(1)}$. Furthermore, it is assumed that the term $p(Y^{(3)}|Y^{(2)}, Y^{(1)})$ does not change when we move from M_1 to M_2 . Therefore, in order to compare $BIC(M_1)$ with $BIC(M_2)$, only the estimation of a finite mixture model for $(Y^{(1)}, Y^{(2)})$ is needed. The algorithm is given by the following steps, as described in [3]:

1. Select the first clustering variable to be the one that has the most evidence of univariate clustering.
2. Select the second clustering variable to be the one that shows the most evidence of bivariate clustering including the first variable selected.
3. Propose the next clustering variable to be the one that shows the most evidence of multivariate clustering including the previous variables selected. Accept this variable as a clustering variable if the evidence favors this over its not being a clustering variable.
4. Propose the variable for removal from the current set of selected clustering variables to be the one for which the evidence of multivariate clustering including all variables selected versus multivariate clustering only on the other variables selected and not on the proposed variable is weakest. Remove this variable from the set of clustering variables if the evidence for clustering is weaker than that for not clustering.
5. Iterate steps 3 and 4 until two consecutive steps have been rejected, then stop.

2.2 An algorithm for the identification of multiple clusterings

The aim of this section is to introduce an algorithm suitable for the identification of a model like (1). Assume we have identified $t - 1$ subsets of Y providing information about $t - 1$ independent clusterings. Now, we wonder whether some variables in $Y^{\bar{C}_{t-1}}$, independent of Y^{R_s} , $s = 1, \dots, t - 1$, can still provide any information about an undetected clustering. Let $Y_t^{(1)}$ be a subset of $Y^{\bar{C}_{t-1}}$ providing such information, $Y_t^{(2)}$ a variable in $Y^{\bar{C}_{t-1}}$ candidate for inclusion into (or exclusion from) $Y_t^{(1)}$, and $Y_t^{(3)}$ the set of the remaining variables in $Y^{\bar{C}_{t-1}}$. The decision of including (excluding) $Y_t^{(2)}$ into (from) $Y_t^{(1)}$, is taken by comparing the following models:

$$\begin{aligned} M_{t,1}: p(Y|\theta_{M_{t,1}}, M_{t,1}) &= p(Y_t^{(3)}|Y_t^{(2)}, Y_t^{(1)}, Y^{C_{t-1}}, \theta_{t,13}) \\ &\quad p(Y_t^{(2)}|Y_t^{(1)}, Y^{C_{t-1}}, \theta_{t,12}) \\ &\quad p_t(Y_t^{(1)}|\theta_{t,11})p(Y^{C_{t-1}}|\theta_{C_{t-1}}), \end{aligned} \quad (5)$$

$$\begin{aligned} M_{t,2}: p(Y|\theta_{M_{t,2}}) &= p(Y_t^{(3)}|Y_t^{(2)}, Y_t^{(1)}, Y^{C_{t-1}}, \theta_{t,13}) \\ &\quad p_t(Y_t^{(2)}, Y_t^{(1)}|\theta_{t,21})p(Y^{C_{t-1}}|\theta_{C_{t-1}}), \end{aligned} \quad (6)$$

where $p_t(Y_t^{(1)}|\theta_{t,11})$ and $p_t(Y_t^{(2)}, Y_t^{(1)}|\theta_{t,21})$ in (5) and (6) are the densities of two finite Gaussian mixtures. Notice that, under $M_{t,2}$, $Y_t^{(2)}$ is a clustering variable and $Y_t^{(1)}$ and $Y_t^{(2)}$ are independent of $Y^{C_{t-1}}$. The algorithm we propose works as follows:

I. If $t = 1$ run Raftery and Dean algorithm, otherwise

1. Select the first clustering variable in the t -th clustering, $Y_t^{(1)}$ to be the one that minimizes $BIC(M_1^*) - BIC(M_2^*)$, with

$$\begin{aligned} M_{t,1}^* : p(Y|\theta_{M_{t,1}^*}) &= p(Y_t^{(3)}|Y_t^{(1)}, Y^{C_{t-1}}, \theta_{t,13})p(Y^{(1)}|Y^{C_{t-1}}, \theta_{t,11}) \\ &\quad p(Y^{C_{t-1}}|\theta_{C_{t-1}}), \\ M_{t,2}^* : p(Y|\theta_{M_{t,2}^*}) &= p(Y_t^{(3)}|Y_t^{(1)}, Y^{C_{t-1}}, \theta_{t,13})p(Y_t^{(1)}|\theta_{t,21}) \\ &\quad p(Y^{C_{t-1}}|\theta_{C_{t-1}}). \end{aligned}$$

2. Propose the next clustering variable to be the one that shows the most evidence of multivariate clustering including the previous variables selected. Accept this variable as a clustering variable if the evidence favors (6) over (5).
3. Propose the variable for removal from the current set of selected clustering variables to be the one for which the evidence of multivariate clustering including all variables selected versus multivariate clustering only on the other variables selected and not on the proposed variable is weakest. Remove this variable from the set of clustering variables if the evidence for (6) is weaker than that for (5).
4. Iterate steps 2. and 3. until two consecutive steps have been rejected, then stop.

II. Iterate step I. until no further grouping structure can be found in the data.

2.3 Identification of the number of clusterings

After running the algorithm, we identify a number, say T^* , of different clusterings, produced by the following T^* models:

$$\begin{aligned} M_1 : p(Y|\theta_{M_1}) &= p(Y^{\bar{C}_1}|Y^{C_1}, \theta_{\bar{C}_1})P(Y^{C_1}|\theta_{C_1}) \\ M_T : p(Y|\theta_{M_T}) &= p(Y_T^{\bar{C}_T}|Y^{C_T}, \theta_{\bar{C}_T})P(Y^{C_T}|\theta_{C_T}), \quad T = 2, \dots, T^* \end{aligned}$$

At each step, the algorithm estimates one of such models, and, as a byproduct, computes the value of its BIC , say $BIC(M_T)$. By selecting the model maximising the BIC criterion, M_{T_0} , we identify the *optimal* number of clusterings, T_0 .

It is worth to notice that, when $T_0 = 1$, M_{T_0} coincides with the (single clustering) model identified by Raftery and Dean's algorithm; when $T_0 > 1$ and $Y^{\bar{C}_{T_0}}$ is

empty, i.e. when all the variables in Y provide some information about clustering, M_{T_0} reduces to:

$$p(Y|\theta_{M_{T_0}}) = \prod_{t=1}^{T_0} p(Y^{R_t}|\theta_{C_{T_0}}),$$

which is the model considered in [1].

3 Simulation study.

We shall present the results of two experiments based on the simulation of samples from

$$Y = [Y'_1, \dots, Y'_T, U'_{11}, \dots, U'_{1S}, U'_{K1}, \dots, U'_{KS}, W_1, \dots, W_J]' \quad (7)$$

where

- $Y_t = [Y_{1t}, Y_{2t}]' \sim \sum_{g=1}^{G_t} \pi_{gt} \phi(x|\mu_{gt}, \Sigma_{gt})$, where $t = 1, \dots, T$, $G_t > 1$, $\pi_{gt} = G_t^{-1}$, the functions $\phi_{gt}(x|\mu_{gt}, \Sigma_{gt})$ are bivariate Gaussian densities, with expectation μ_{gt} and variance matrix Σ_{gt} such that $\mu_{gt} = (g-1)m_t$ and Σ_{gt} is equal to one of the following matrices:

$$S_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 1.5 & -0.75 \\ -0.75 & 1.5 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad S_4 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Furthermore, Y_t and Y_s are independent for $t \neq s$.

- $U_{ks} = Y_{1k} + \varepsilon_s$, $\varepsilon_s \stackrel{iid}{\sim} N(0, 1)$, $s = 1, \dots, S$, $k = 1, \dots, K \leq T$.
- $W_j \stackrel{iid}{\sim} N(1.5, 4)$, $j = 1, \dots, J$, with W_j independent of Y_t and ε_s for all j, s, t .

3.1 Experiment 1

We fixed $T = 3$, $G_1 = 2$, $G_2 = 3$, $G_3 = 4$, $m_1 = 5$, $m_2 = 4$, $m_3 = 3$, $\Sigma_{11} = \Sigma_{12} = \Sigma_{13} = S_1$, $\Sigma_{21} = \Sigma_{22} = \Sigma_{23} = S_2$, $\Sigma_{32} = \Sigma_{33} = S_3$, and $\Sigma_{34} = S_4$, $K = 1$ and for different values of S , J and n we generated 1000 independent samples of size n from (7). Table 1 shows the results of the simulation: we can notice that our algorithm very often selects the variables correctly for moderately large sample sizes, whereas the dimension of the data set does not affect sensibly its performance.

3.2 Experiment 2

In this experiment we fixed $T = 2$, $G_1 = G_2 = 2$, $m_1 = 5$, $m_2 = 3$, $\Sigma_{1t} = S_1$, $\Sigma_{2t} = S_2$, and $K = 1$. For different values of S and J , we generated 1000 independent samples

n	J	S	N. of correct variable selections	N. of correct variable selections for		
				mixture 1	mixture 2	mixture 3
100	2	1	708	985	988	723
150	2	1	914	984	985	928
200	2	1	956	974	973	978
200	9	1	946	966	964	982
200	19	1	940	970	968	969
200	10	10	905	968	970	935
200	49	1	925	961	961	966
200	25	25	886	970	973	917

Table 1 Results on 1000 independent simulations for each value of n and J in experiment 1.

of size n from (7), ran our algorithm on each sample and computed the adjusted Rand index (ARI) between the true simulated clustering and the ones we estimated. Table 2 shows the results of the simulation: the identification of the variables is as satisfactory as in the previous experiment and the average ARI index between the true and the estimated partition is very close to 1.

n	J	S	N. of correct variable selections	N. of correct variable selections for		Average ARI	
				mixture 1	mixture 2	1 st clustering	2 nd clustering
100	2	1	982	991	987	0.980	0.861
150	2	1	992	993	993	0.989	0.889
200	2	1	997	997	999	0.994	0.902
200	9	1	993	995	997	0.997	0.904
200	19	1	985	996	987	0.995	0.892
200	10	10	993	996	996	0.995	0.901
200	49	1	978	997	979	0.995	0.890
200	25	25	973	992	979	0.995	0.892

Table 2 Results on 1000 independent simulations in experiment 2.

References

1. Galimberti, G., Soffritti, G.: Model-based methods to identify multiple cluster structures in a data set. *Computational Statistics & Data Analysis*. **52**(1), 520 – 536 (2007).
2. Niu, D., Dy, J.G., and Ghahramani, Z.: A nonparametric bayesian model for multiple clustering with overlapping feature views. *Journal of Machine Learning Research*. **22**, 814–822 (2012).
3. Raftery, A.E., Dean, N.: Variable selection for model-based clustering. *Journal of the American Statistical Association*. **101**(473), 168–178 (2006).
4. Maugis, C., Celeus, G., Martin-Magnette, M.: Variable selection for Clustering with Gaussian Mixture Models. *Biometrics*. **65**, 701–709 (2009).