Double Cross Validation for Model Based Classification

Romain François* and Florent Langrognet†

February 28, 2006

Keywords: Cross Validation, Classification, Gaussian Mixtures.

Introduction

Gaussian mixture modelling is a powerful framework for classification. The observations $x_1, \ldots, x_n$ are assumed to arise from a mixture of K normal distributed components.

$$f(.) = \sum_{k=1}^{K} p_k \Phi(.|\mu_k, \Sigma_k) \quad , \quad 0 < p_k < 1 \quad , \quad \sum_{k=1}^{K} p_k = 1 \tag{1}$$

where p_k are the mixing proportions, $\mu_k \in \mathbb{R}^d$ the mean vector of the k^{th} component, Σ_k its covariance matrix and $\Phi(.|\mu, \Sigma)$ the normal probability density function with mean vector μ and variance matrix Σ.

Celeux and Govaert (1995) proposed a decomposition of the variance matrices in terms of *volume*, *orientation* and *shape*. That decomposition yields 14 models from the simplest $[\lambda I]$ (same volume, shape and orientation with spherical variance matrices) to the standard QDA model $[\lambda_k C_k]$.

Mixmod

MIXMOD[1] is an open source C++ software for Gaussian mixture modelling with EM-like algorithms. MIXMOD proposes all the 14 models from Celeux and Govaert (1995) and model selection criteria based on penalized likelihood (BIC, ICL, NEC) or quality of prediction via cross validation. MIXMOD is originally interfaced with MATLAB or SCILAB and has been ported to R recently. The R package *mixmod* should be available on CRAN soon.

```
out  <-mixmod(iris[,1:4], nbCluster=3)
plot(out, type="zones") # produces fig 1 : left
rgl(out, contours=TRUE, obs=TRUE) # produces fig 1 : right
```

Double Cross Validation

The cross-validated error rate may be used to select a model between several candidats when quality of prediction matters. However, that error rate is too optimistic as it does not take into account the uncertainty of the selection procedure.

The *double* cross-validated error rate that we propose makes use of the cross validation methodology at two stages in order to estimate the overall error rate of the procedure *"model the observations by a Gaussian mixture with one of the 14 structures"*. It proceeds as follows :

*INRIA Futurs, projet SELECT. Corresponding Author. R. François, Université Paris SUD, Bat. 425, 91405 Orsay Cedex (France). email : Romain.Francois@inria.fr

†UMR 6623 CNRS, Université de Franche-Comté

[1]The MIXMOD program and its documentation are available freely on the internet: http://www-math.univ-fcomte.fr/mixmod/

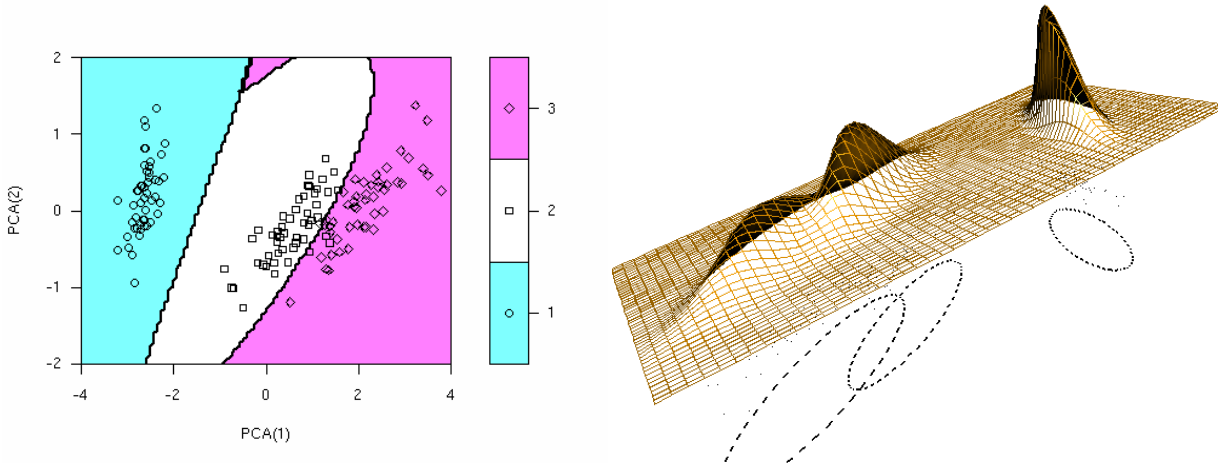

Figure 1: Some MIXMOD graphics. The one on the left uses `filled.contour` to display the classification zones in the two first PCA axes. The graphic on the right uses `rgl` to display the density mixture and the cluster's 95% iso-density ellipses.

- Random split the sample S in V sub-samples $S^{(1)}, \ldots, S^{(V)}$
- For $v = 1, \ldots, V$, do :
  - ★ Merge $V - 1$ subsamples into $S^{(-v)} = S - S^{(v)}$
  - ★ For each candidate model $m \in \mathcal{M}$, compute the discrimination rule $\theta_m^{(-v)}$ and select the best model regarding the cross-validated error-rate :

$$m_v^\star = \underset{m \in \mathcal{M}}{\operatorname{argmin}}\ CV_m^{(-v)}$$

  - ★ Evaluate the error rate t_v of $m_v^\star$ on the test sample $S^{(v)}$
- Average the V error rates $t_1, \ldots, t_V$

Double cross validation : Sketch of the algorithm

The double cross-validated error rate gives a good point estimate of the error rate and a measure of its variability. Moreover, the frequencies of the models inside the V winners $m_1^\star, \ldots, m_V^\star$ may be used to qualify the stability of these models. A model selected frequently has good chances to be well adpated to the problem.

```
R> out  <- mixmod(iris[,1:4], crit="DCV", lab=as.numeric(iris[,5]))
R> out$modelOutput$DCV
[1] 0.0333333
```

References

Birenacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis*, to appear.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–93.

François, R. (2006). *Sélection d'un modèle de mélange pour la classification. Double validation croisée. Application aux données de biopuce.* INRIA Futurs. Mémoire de stage ISUP. `http://addictedtor.free.fr/rapportStage`.