# Appendix to "Combining Mixture Components for Clustering" published in the Journal of Computational and Graphical Statistics *


Jean-Patrick Baudry
Université Paris-Sud XI

Adrian E. Raftery
University of Washington

Gilles Celeux
INRIA

Kenneth Lo
University of Washington

Raphaël Gottardo
University of British Columbia


March 5, 2010


*Jean-Patrick Baudry is Doctorant and Gilles Celeux is Directeur de Recherche, both at INRIA Saclay Île-de-France, Université Paris-Sud, Bâtiment 425, 91405 Orsay Cedex, France. Jean-Patrick Baudry is also with Laboratoire MAP5, Université Paris Descartes and CNRS. Adrian E. Raftery is Blumstein-Jordan Professor of Statistics and Sociology, Box 354322, University of Washington, Seattle, WA 98195-4322. Kenneth Lo is Postdoctoral Senior Fellow, Department of Microbiology, Box 358070, University of Washington, Seattle, WA 98195-8070. Raphaël Gottardo is Research Unit Director in Computational Biology, Institut de recherches cliniques de Montréal (IRCM), 110, avenue des Pins Ouest, Montréal, Canada H2W 1R7.


Raftery's research was supported by NIH-NICHD Grant no. HD-54511, NSF grant no. llS-0534094, and NSF grant no. ATM-0724721. The research of Lo and Gottardo was supported by a discovery grant of the Natural Sciences and Engineering Research Council of Canada and by NIH grant no. R01-EB008400,. The authors are grateful to Naiysin Wang for suggesting the idea of using entropy as the criterion for the mixture component merging procedure, which is fundamental to the paper. They also thank Christian Hennig for helpful discussions.

# 1 ALGORITHM

Choose a family of mixture models: $\{\mathcal{M}_{K_{\min}}, \ldots, \mathcal{M}_{K_{\max}}\}$. Complete Gaussian mixture models are suggested: $\mathcal{M}_K$ contains any mixture with $K$ Gaussian components. Here is the algorithm we work with:

1. Compute MLE(K) for each model using the EM algorithm:

$$\forall K \in \{K_{\min}, \ldots, K_{\max}\}, \quad \hat{\theta}_K = \arg \max_{\theta_K \in \Theta_K} \log p(\mathbf{x} \mid K, \theta_K)$$

2. Compute the BIC solution:

$$\hat{K}^{\text{BIC}} = \underset{K \in \{K_{\min}, \ldots, K_{\max}\}}{\operatorname{argmin}} \left\{ -\log p(\mathbf{x} \mid K, \hat{\theta}_K) + \frac{\nu_K}{2} \log n \right\}$$

3. Compute the density $f_k^K$ of each combined cluster k for each $K$ from $\hat{K}^{\text{BIC}}$ to $K_{\min}$:

$$\forall k \in \{1, \ldots, \hat{K}^{\text{BIC}}\}, \quad f_k^{\hat{K}^{\text{BIC}}}(\cdot) = \hat{p}_k^{\hat{K}^{\text{BIC}}} \phi \left( \cdot \mid \hat{a}_k^{\hat{K}^{\text{BIC}}} \right).$$

   For $K = \hat{K}^{\text{BIC}}, \ldots, (K_{\min} + 1)$:

   - Choose the clusters $l$ and $l'$ to be combined at step $K \to K - 1$ :

$$(l, l') = \underset{(k,k') \in \{1,\ldots,K\}^2,\ k \neq k'}{\operatorname{argmax}} \left\{ - \textstyle\sum_{i=1}^n \left\{ t_{ik}^K \log(t_{ik}^K) + t_{ik'}^K \log(t_{ik'}^K) \right\} \right.$$
$$\left. + \textstyle\sum_{i=1}^n (t_{ik}^K + t_{i\,k'}^K) \log(t_{ik}^K + t_{i\,k'}^K) \right\},$$

     where $t_{ik}^K = \frac{f_k^K(x_i)}{\sum_{j=1}^K f_j^K(x_i)}$ is the conditional probablity of component $k$ given the $K$-cluster combined solution.

   - Define the densities of the combined clusters for the (K-1) cluster solution by combining $l$ and $l'$:

     for $k = 1, \ldots, (l \wedge l' - 1), (l \wedge l' + 1), \ldots, (l \vee l' - 1)$ $\quad \{f_k^{K-1} = f_k^K\}$
     $f_{l \wedge l'}^{K-1} = f_l^K + f_{l'}^K$
     for $k = l \vee l', \ldots, (K - 1)$ $\quad \{f_k^{K-1} = f_{k+1}^K\}$

4. To select the number of clusters through ICL:

$$\hat{K}^{\text{ICL}} = \underset{K \in \{K_{\min}, \ldots, K_{\max}\}}{\operatorname{argmin}} \left\{ -\log p(\mathbf{x} \mid K, \hat{\theta}_K) - \sum_{i=1}^n \sum_{k=1}^K t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) + \frac{\nu_K}{2} \log n \right\},$$

   where $t_{ik}(\hat{\theta}_K) = \frac{\hat{p}_k^K \phi(x_i | \hat{a}_k^K)}{\sum_{j=1}^K \hat{p}_j^K \phi(x_i | \hat{a}_j^K)}$ is the conditional probability of component $k$ given the MLE for the model with $K$ Gaussian components.