Un critère de sélection de modèle pour la classification non supervisée de données annotées: applications à l'analyse de données d'expression de gènes RNA-seq

Mélina Gallopin[1,2] & Andrea Rau[2,3] & Florence Jaffrézic[2,3] & Gilles Celeux[1,4]

[1] *Université Paris-Sud 11, Laboratoire de Mathématiques, 91405 Orsay, France.*
melina.gallopin@math.u-psud.fr
[2] *INRA, UMR1313 Génétique animale, 78352 Jouy-en-Josas, France.*
andrea.rau@jouy.inra.fr
[3] *AgroParisTech, UMR1313 Génétique animale, 75231 Paris, France.*
florence.jaffrezic@jouy.inra.fr
[4] *Inria Saclay, Île-de-France, 91405 Orsay, France. gilles.celeux@inria.fr*

Résumé. En classification non supervisée, il est souvent utile de pouvoir interpréter a posteriori les classes formées à l'aide d'informations externes. Les modèles de mélange fournissent un cadre probabiliste adapté. Nous proposons de prendre en compte des annotations externes disponibles pour une partie des observations dans l'étape de sélection de modèle, généralement effectuée à l'aide des critères BIC (Bayesian Information Criterion) ou ICL (Integrated Completed Likelihood criterion). Notre critère de sélection de modèle est basé sur une approximation de la log-vraisemblance complétée du modèle sachant les annotations. Il inclut un terme d'entropie mesurant le lien existant entre la classification inférée et les annotations externes. Dans le cas des données d'expression de gènes, ces annotations externes sont fournies par la liste potentiellement incomplète des propriétés fonctionnelles des gènes, aussi appelées termes GO (Gene Ontology). Le critère de sélection proposé conduit à former des classes de gènes plus faciles à interpréter biologiquement. L'intérêt de cette stratégie de sélection de modèle est illustré pour des modèles de mélange gaussien et de lois de Poisson sur des données simulées et sur des données réelles d'expression de gènes RNA-seq.

Mots-clés. Modèle de mélange, sélection de modèle, clustering non-supervisé, données d'expression de gènes RNA-seq, annotation fonctionnelle du génome.

Abstract. In cluster analysis, it is often of interest to interpret the obtained clustering with respect to some external information. Mixture models provide a probabilistic framework for this purpose. We propose a model selection strategy taking into account information provided by external annotation available for some or all observations in the model selection step, usually performed using the Bayesian Information Criterion (BIC) or the Integrated Completed Likelihood criterion (ICL). We derive a model selection criterion based on an approximation of the completed log-likelihood of the model given the

annotations. This criterion includes an entropy term measuring the link between a clustering and the annotation information. In the case of gene expression data, such external information is a (potentially incomplete) list of functional properties attributed to each gene, known as Gene Ontology terms. The proposed model selection strategy leads to a choice of models for which the clusters are more easily interpretable from a biological point of view. We illustrate the interest of this model selection strategy for Gaussian and Poisson mixture models using simulated and RNA-seq gene expression data.

Keywords. Mixture model, model-based clustering, unsupervised clustering, model selection, RNA-seq data, genome functionnal annotation.

1 Introduction

Soit $\mathbf{y}$ la matrice de données de dimension $n \times q$ où n est le nombre d'objets à classer et q le nombre de variables. On suppose que la matrice $\mathbf{y}$ est la réalisation d'un mélange de K variables aléatoires distinctes de paramètres $\Theta_K = (p_1, \ldots, p_{K-1}, \mathbf{a}_1, \ldots, \mathbf{a}_K)$ où $(\mathbf{a}_1, \ldots, \mathbf{a}_K)$ sont les paramètres de chaque composante et $(p_1, \ldots, p_K)$ sont les proportions du modèle de mélange avec $\sum_{k=1}^{K} p_k = 1$. Pour un objet d'indice $i \in (1, \ldots, n)$, on a:

$$f(y_i|K, \Theta_K) = \sum_{k=1}^{K} p_k f(\mathbf{y}_i|\mathbf{a}_k).$$

Soit $\mathbf{z}$ la matrice des données manquantes de dimension $n \times K$ indiquant la classe à laquelle appartient chaque objet: $z_{ik} = 1$ si l'objet i appartient à la classe k, 0 sinon. La procédure de classification revient à estimer ces variables cachées $\mathbf{z}$ à l'aide d'un algorithme EM (Dempster et al., 1977).

Une étape importante dans la procédure de classification est le choix du modèle et du nombre de composantes K. Une stratégie consiste à choisir le modèle et le nombre K maximisant la vraisemblance intégrée suivante, où $\pi(\theta|K)$ est une distribution a priori sur les paramètres du modèle de mélange à K composantes:

$$f(\mathbf{y}|K) = \int_{\theta_K} f(\mathbf{y}|K, \theta)\pi(\theta|K)d\theta.$$

Le critère de sélection de modèle BIC (Bayesian information criterion) proposé par Schwarz (1978) est une approximation de cette log-vraisemblance intégrée, pour n assez grand.

Une autre stratégie consiste à maximiser la vraisemblance complétée intégrée suivante, où $\pi(\theta|K)$ est une distribution a priori sur les paramètres du modèle de mélange à K composantes:

$$f(\mathbf{y}, \mathbf{z}|K) = \int_{\theta_K} f(\mathbf{y}, \mathbf{z}|K, \theta)\pi(\theta|K)d\theta.$$

Cette stratégie vise à regrouper plusieurs composantes du mélange pour former des classes plus cohérentes en terme de classification. Le critère de sélection de modèle ICL, proposé par Biernacki, Celeux et Govaert (2000), est une approximation de cette log-vraisemblance complétée intégrée. D'autres adaptations des critères de sélection existent, notamment le critère SICL proposé par Baudry et al. (2012).

On propose de prendre en compte des informations externes sur la nature des objets à classer dans cette étape de sélection de modèle.

2 Un critère de sélection de modèle pour la classification d'objets annotés

On considère $\mathbf{u}$ la matrice de données binaires de taille $n \times M$ contenant les informations externes disponibles (M annotations différentes) pour les n objets. Soit un objet d'indice i et une annotation d'indice m. On note $u_{im} = 1$ si l'objet i est annoté par m. Si $u_{im} = 0$, on ne peut pas savoir s'il s'agit d'une donnée manquante ou si la valeur nulle est réellement informative. Pour cette raison, on travaille uniquement à partir des annotations positives.

Le critère de sélection de modèle proposé consiste à maximiser la vraisemblance complétée sachant les annotations positives observées, où $\pi(\theta|K)$ est une distribution a priori sur les paramètres du modèle de mélange à K composantes:

$$f(\mathbf{y}, \mathbf{z}|\mathbf{u} = 1, K) = \int_{\Theta_K} f(\mathbf{y}, \mathbf{z}|\mathbf{u} = 1, K, \theta)\pi(\theta|K)d\theta.$$

On note n_{km} le nombre d'objets annotés par m dans la classe k et $n_{.m}$ le nombre d'objets annotés par m au total. On note également $\hat{\theta}_K$ l'estimateur du maximum de vraisemblance, ν_K le nombre de paramètres libres dans le modèle, $\tilde{\mathbf{z}} = MAP(\hat{\theta})$ et $\hat{\mathbf{a}}^\star = \arg\max_{\mathbf{a}} \log f(\mathbf{y}|\tilde{\mathbf{z}}, K, \mathbf{a})$. Une approximation de la log-vraisemblance complétée intégrée sachant les annotations, noté MIL pour *Marked Integrated Likelihood*, s'écrit de la manière suivante:

$$\text{MIL}(K) = \log f(\mathbf{y}|\tilde{\mathbf{z}}, K, \hat{\mathbf{a}}^\star) - \frac{\nu_K}{2}\log(n) + \sum_{m=1}^{M}\sum_{k=1}^{K} n_{km}\log(\frac{n_{km}}{n_{.m}}).$$

Pour une annotation donnée m, le terme $\sum_{k=1}^{K} n_{km}\log(\frac{n_{km}}{n_{.m}})$ est nul si tous les objets annotés par m sont dans une seule et même classe. L'annotation m est alors spécifique à cette classe et l'interprétation des classes est facilitée.

3 Illustration du critère sur un jeu de données simulé

On illustre l'utilisation de ce critère sur un jeu de données. On simule 600 observations issues d'un mélange de quatre lois gaussiennes de dimension deux.

On construit une matrice des annotations externes de taille 600×5. Dans un premier temps, les cinq annotations sont informatives: tous les objets annotés positivement (correspondant aux croix sur la figure 1) appartiennent aux classes 1 et 2. Dans cette configuration, le critère de sélection de modèle MIL privilégie ainsi une classification à trois classes, regroupant les classes 1 et 2.

On construit ensuite une matrice des annotations externes non informative: tous les objets annotés positivement (correspondant aux croix sur la figure 2) sont répartis de manière aléatoire dans les quatre classes. Le critère MIL sélectionne une classification quatre classes ce qui correspond bien au nombre de classes simulées, et au nombre de classes sélectionné par les critères classiques BIC et ICL.

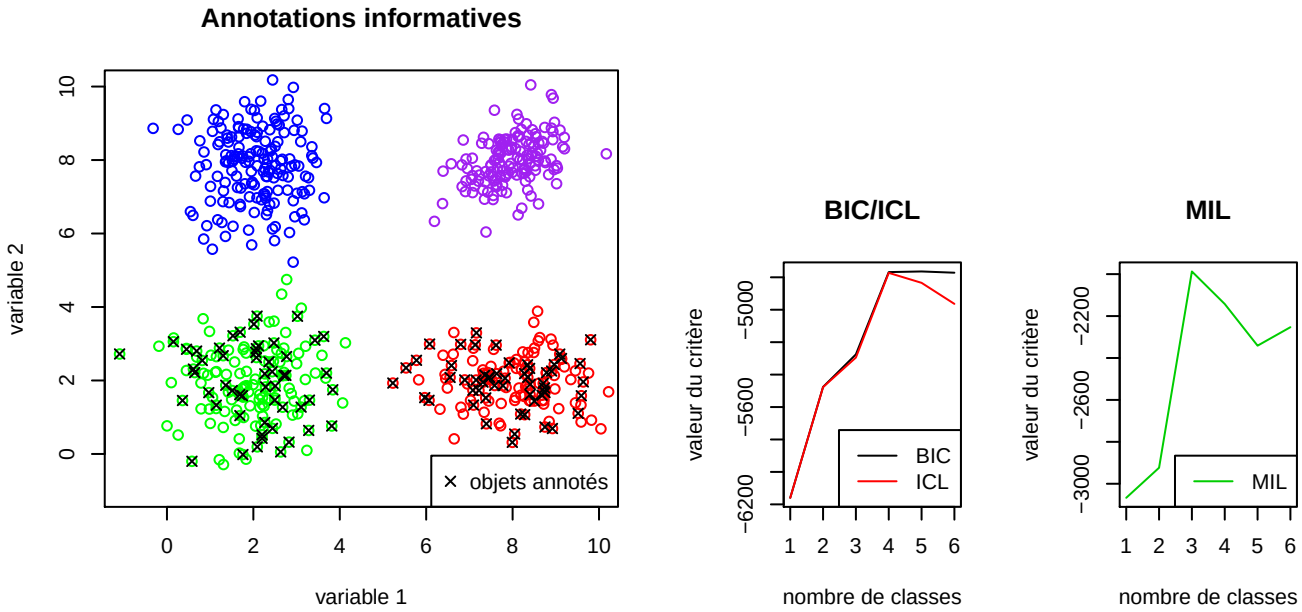

Figure 1:
Sélection de modèle dans le cas d'annotations externes pertinentes pour la classification réalisée.

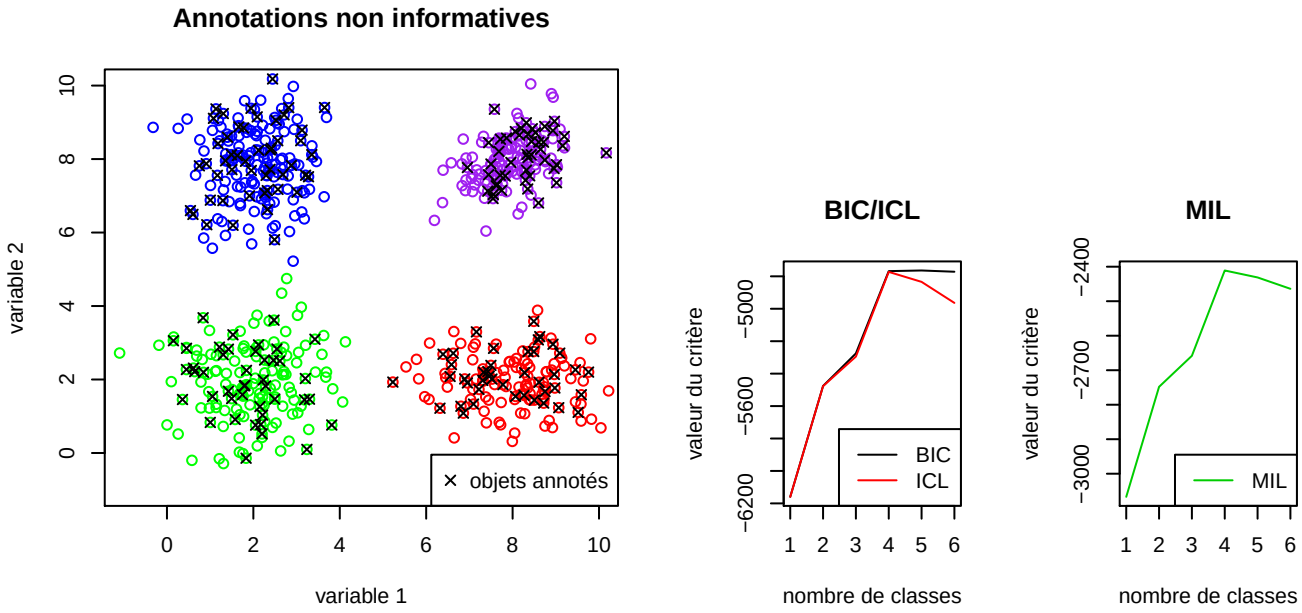

Figure 2: Sélection de modèle dans le cas d'annotations externes non pertinentes pour la classification réalisée.

4 Discussion

D'un point de vue biologique, l'intérêt de ce critère réside dans le fait que les différentes classifications possibles sont construites uniquement à partir des données d'expression de gènes. Les annotations externes contribuent seulement à sélectionner la meilleure classification possible. Si certains gènes ont été mal annotés ou si les annotations utilisées ne sont pas pertinentes pour le problème biologique étudié, l'intégration de ces annotations externes ne va pas détériorer la classification. Si les annotations sont pertinentes, les modules de gènes obtenus seront plus riches et plus pertinents pour la compréhension du processus biologique sous-jacent.

Dans cet exposé, nous présenterons d'abord les détails de l'approximation effectuée dans l'écriture du critère, puis les résultats de simulation pour des mélanges gaussiens et de lois de Poisson. Nous montrerons l'intérêt de ce critère dans l'étude du fonctionnement de l'intestin grêle chez le porc, à partir d'un jeu de données d'expression de gènes RNA-seq issu du département de Génétique Animale de l'INRA.

Bibliographie

[1] McLachlan, G. et Krishman, T. (1997), The EM algorithm and Extensions, *Wiley*, New York.
[2] Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, 6, 461-464.
[3] Biernacki, C., Celeux, G., Govaert,G. (2000), Assessing a mixture model for clustering with the Integrated Classification Likelihood, *IEEE Transaction on PAMI*, 22, 719-725.
[4] Baudry, J-P., Cardoso, M., Celeux,G. , Amorim, M-J. , Sousa Ferreira, A. (2012), Enhancing the selection of a model-based clustering with external qualitative variables, *Inria Research Report*, n8124.