Sélection de variables en classification non-supervisée sans estimation de paramètres

Matthieu Marbac[1] & Mohammed Sedki[2]

[1] *Inserm U1181, matthieu.marbac@inserm.fr*
[2] *Université Paris-Sud, Inserm U1181, mohammed.sedki@inserm.fr*

Résumé. Nous présentons le critère MICL (*Maximum Integrated Complete-data Likelihood*) qui est utilisé en classification non supervisée pour la sélection de variables d'un mélange gaussien. Ce critère est basé sur la forme explicite de la vraisemblance complétée intégrée et permet d'effectuer la sélection de modèle préalablement à l'estimation des paramètres. Ainsi, il évite les procédures d'optimisation complexes et chronophages inhérentes aux critères classiques tels que BIC et ICL. Toutefois, ses propriétés restent similaires à celles du critère ICL. L'apport du critère MICL est illustré sur différents jeux de données réelles.

Mots-clés. Classification, critère d'information, mélange gaussien, sélection de modèle.

Abstract. We introduce the MICL criterion (*Maximum Integrated Complete-data Likelihood*) which allows to perform the variable selection for a cluster analysis done by a Gaussian mixture. This criterion, based on the closed form of the integrated complete-data likelihood, carries out the variable selection before to achieve the parameter estimation. Thus, it avoids complex and time consuming procedures of optimization required by the BIC and the ICL criteria. However, the properties of the MICL criterion are similar to those of the ICL criterion. The benefit of the MICL criterion is shown on different challeging real datasets.

Keywords. Clustering, Gaussian mixture, information criteria, model selection.

1 Introduction

La sélection de variables en classification non supervisée a un double objectif : améliorer la *qualité de l'inférence* et favoriser l'*interprétation des classes* en mettant en évidence le sous-ensemble de variables discriminantes.

Cet objectif peut-être atteint par des méthodes de *régularisation*. Parmi celles-ci, la plus populaire consiste en une version des K-means pénalisée en norme ℓ_1 pour imposer la sparsité (Witten and Tibshirani, 2010). Si cette approche permet de gérer des données de grandes dimensions, elle est tributaire de la grille de pénalités utilisée. De plus, elle ne répond que partiellement aux problèmes du choix du nombre de classes.

Dans un cadre probabiliste, la distribution des données peut se modéliser par un mélange de distributions gaussiennes ayant des matrices de covariance diagonales. La sélection de variables est alors un problème complexe de choix de modèle du fait du très grand nombre de modèles candidats (Raftery and Dean, 2006). Dans ce contexte, l'objectif est de déterminer le modèle maximisant un critère d'information tel que BIC (Schwarz, 1978) ou ICL (Biernacki et al., 2000). L'estimation de ce modèle est effectuée par des méthodes dites pas-à-pas (*stepwise*) qui sont sous-optimales. De plus, comme les critères BIC et ICL nécessitent l'estimateur du maximum de vraisemblance, celui-ci doit être estimé pour chaque comparaison de modèles, ce qui implique des temps de calculs conséquents.

Nous proposons d'utiliser un nouveau critère d'information, nommé MICL (Maximum Integrated Complete-data Likelihood), qui est basé sur la vraisemblance complétée intégrée. Il possède donc une forme explicite pour un mélange gaussien. Le critère MICL a pour avantage de ne pas nécessiter d'estimateur de paramètres, cependant il utilise un estimateur de la partition. De plus, le modèle maximisant ce critère est obtenu par un algorithme d'optimisation alternée. Cette approche permet alors de réduire considérablement les temps de calculs lorsque le nombre d'individus est modéré (inférieur à 10^4). De plus, elle permet d'éviter les problèmes de sous-optimalité inhérents aux méthodes pas-à-pas.

Cet article est organisé comme suit. La partie 2 rappelle brièvement le contexte de sélection de variables pour un mélange gaussien. La partie 3 introduit le nouveau critère MICL et présente la procédure permettant de déterminer le modèle maximisant ce critère. Dans la partie 4, la méthode proposée est comparée à deux approches de référence sur données réelles. Une discussion est menée dans la partie 5.

2 Sélection de variables pour un mélange gaussien

2.1 Mélange gaussien avec variables non discriminantes

Les données sont composées de n observations $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ où le vecteur $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ est décrit par d variables continues. On considère que les observations sont issues d'un modèle de mélange gaussien à g composantes supposant l'indépendance conditionnelle entre les variables. Ainsi, la densité du modèle s'écrit

$$f(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \prod_{j=1}^{d} \phi(x_{ij}|\mu_{kj}, \sigma_{kj}^2), \tag{1}$$

où $\boldsymbol{m}$ spécifie le modèle, où $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$ regroupe l'ensemble des paramètres, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ étant le vecteur des proportions défini sur le simplex de taille g, $\boldsymbol{\mu} = (\mu_{kj}; k = 1, \ldots, g; j = 1, \ldots, d)$, $\boldsymbol{\sigma} = (\sigma_{kj}; k = 1, \ldots, g; j = 1, \ldots, d)$, et où $\phi(.|\mu_{kj}, \sigma_{kj}^2)$ est la densité d'une gaussienne univariée de moyenne μ_{kj} et de variance σ_{kj}^2.

Une variable est non pertinente pour la classification si sa distribution marginale est égale pour toutes les classes. Ainsi, en introduisant ω_j tel que $\omega_j = 1$ si la variable n'est pas pertinente pour la classification et $\omega_j = 0$ sinon, on a les égalités suivantes

$$\forall j \in \{j' : \omega_{j'} = 1\},\ \mu_{1j} = \ldots = \mu_{gj} \text{ et } \sigma_{1j} = \ldots = \sigma_{gj}. \tag{2}$$

Par conséquent, le modèle $\boldsymbol{m} = (g, \boldsymbol{\omega})$ est défini par le nombre de composantes g et le vecteur binaire $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ indiquant le rôle des variables dans la classification.

2.2 La vraisemblance complétée intégrée

Une partition des individus est définie par le vecteur $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ où $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$ indique la classe de l'individu i, *i.e.* $z_{ik} = 1$ si $\boldsymbol{x}_i$ est issu de la composante k et $z_{ik} = 0$ sinon. En classification non supervisée, $\mathbf{z}$ est une variable manquante. Ainsi la vraisemblance calculée sur l'ensemble des données (observées et manquantes) est appelée vraisemblance complétée. Celle-ci s'écrit

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{g} \left(\pi_k \prod_{j=1}^{d} \phi(x_{ij}|\mu_{kj}, \sigma_{kj}^2)\right)^{z_{ik}}. \tag{3}$$

La vraisemblance intégrée se définie alors comme

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \int_{\boldsymbol{\Theta}_{\boldsymbol{m}}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) d\boldsymbol{\theta}. \tag{4}$$

où $p(\boldsymbol{\theta}|\boldsymbol{m})$ est la distribution *a priori* des paramètres. Celle-ci s'écrit comme

$$p(\boldsymbol{\theta}|\boldsymbol{m}) = p(\boldsymbol{\pi}|\boldsymbol{m}) \prod_{j=1}^{d} p(\boldsymbol{\sigma}_{\bullet j}^2, \boldsymbol{\mu}_{\bullet j}|\boldsymbol{m}), \tag{5}$$

où $\boldsymbol{\sigma}_{\bullet j}^2 = (\sigma_{kj}^2; k = 1, \ldots, g)$, $\boldsymbol{\mu}_{\bullet j}^2 = (\mu_{kj}^2; k = 1, \ldots, g)$, et où

$$p(\boldsymbol{\sigma}_{\bullet j}^2, \boldsymbol{\mu}_{\bullet j}|\boldsymbol{m}) = \Big(\prod_{k=1}^{g} p(\sigma_{kj}^2|\boldsymbol{m}) p(\mu_{kj}|\boldsymbol{m}, \sigma_{kj}^2)\Big)^{1-\omega_j} \Big(p(\sigma_{1j}^2|\boldsymbol{m}) p(\mu_{1j}|\boldsymbol{m}, \sigma_{1j}^2)\Big)^{\omega_j}. \tag{6}$$

Des lois conjuguées sont utilisées pour les priors : $\boldsymbol{\pi}|\boldsymbol{m}$ suit une distribution de Dirichlet $\mathcal{D}_g(\frac{1}{2}, \ldots, \frac{1}{2})$, $\sigma_{kj}^2|\boldsymbol{m}$ suit une distribution Inverse-Gamma $\mathcal{IG}(\alpha_j/2, \beta_j^2/2)$ et $\mu_{kj}|\boldsymbol{m}, \sigma_{kj}^2$ suit une distribution gaussienne $\mathcal{N}(\lambda_j, \sigma_{kj}^2/\delta_j)$, où $(\alpha_j, \beta_j, \lambda_j, \delta_j)$ sont des hyper-paramètres. Ainsi la vraisemblance intégrée a la forme explicite suivante

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \frac{\Gamma(\frac{g}{2})}{\Gamma(\frac{1}{2})^g} \frac{\prod_{k=1}^{g} \Gamma(n_k + \frac{1}{2})}{\Gamma(n + \frac{g}{2})} \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}), \tag{7}$$

où $\mathbf{x}_{\bullet j} = (x_{ij}; i = 1, \ldots, n)$, $n_k = \sum_{i=1}^{n} z_{ik}$. En particulier,

$$p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}) = \begin{cases} \left(\frac{1}{\pi}\right)^{n/2} \frac{\Gamma\left(\frac{n+\alpha_j}{2}\right)}{\Gamma\left(\frac{\alpha_j}{2}\right)} \left(\frac{\beta_j^{\alpha_j}}{s_j^{\alpha_j+n}}\right) \sqrt{\frac{\delta_j}{n+\delta_j}} & \text{si } \omega_j = 1 \\ \prod_{k=1}^{g} \left(\frac{1}{\pi}\right)^{n_k/2} \frac{\Gamma\left(\frac{n_k+\alpha_j}{2}\right)}{\Gamma\left(\frac{\alpha_j}{2}\right)} \left(\frac{\beta_j^{\alpha_j}}{s_{jk}^{\alpha_j+n_k}}\right) \sqrt{\frac{\delta_j}{n_k+\delta_j}} & \text{si } \omega_j = 0, \end{cases} \tag{8}$$

avec $s_j^2 = \beta_j^2 + \sum_{i=1}^{n}(x_{ij} - \bar{\mathrm{x}}_j)^2 + \frac{(\lambda_j - \bar{\mathrm{x}}_j)^2}{(\delta_j^{-1}+(n+\delta_j)^{-1})}$, $\bar{\mathrm{x}}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $s_{jk}^2 = \beta_j^2 + \sum_{i=1}^{n} z_{ik}(x_{ij} - \bar{\mathrm{x}}_{jk})^2 + \frac{(\lambda_j - \bar{\mathrm{x}}_{jk})^2}{(\delta_j^{-1}+(n_k+\delta_j)^{-1})}$ et $\bar{\mathrm{x}}_{jk} = \frac{1}{n_k}\sum_{i=1}^{n} z_{ik}x_{ij}$. Notons que pour j tel que $\omega_j = 1$, on a $p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}) = p(\mathbf{x}_{\bullet j}|g, \omega_j)$ car la partition n'a pas d'impact sur la valeur de l'intégrale.

3 Sélection de modèle par le critère MICL

Nous proposons d'effectuer la sélection de modèle par le critère MICL (Maximum Integrated Complete-data Likelihood). Ce critère correspond à la plus grande valeur de la vraisemblance intégrée complétée parmi l'ensemble des partitions. Ainsi, il s'écrit

$$\mathrm{MICL}(\boldsymbol{m}) = \ln p(\mathbf{x}, \mathbf{z}^\star_{\boldsymbol{m}}|\boldsymbol{m}) \text{ avec } \mathbf{z}^\star_{\boldsymbol{m}} = \arg\max_{\mathbf{z}} \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}). \tag{9}$$

En supposant les partitions $\mathbf{z}$ équiprobables conditionnellement à $\mathbf{x}$, on a $\mathrm{MICL}(\boldsymbol{m}) \propto \ln p(\boldsymbol{m}|\mathbf{x}, \mathbf{z}^\star_{\boldsymbol{m}})$. Le critère MICL est similaire au critère ICL et il hérite de ses propriétés principales comme sa robustesse et ses conditions de consistance (Marbac, M. and Sedki, M., 2015). De plus, à la différence des critères ICL et BIC, il ne nécessite pas l'estimateur du maximum de vraisemblance et il bénéficie du fait que $\mathbf{z}^\star_{\boldsymbol{m}}$ soit numériquement accessible par l'algorithme détaillé dans cette partie.

On considère que les modèles ont au plus $g_{\max}$ composantes et on note $\mathcal{M}$ l'espace des modèles en compétition. On cherche alors à obtenir le modèle $\boldsymbol{m}^\star$ qui maximise le critère MICL

$$\boldsymbol{m}^\star = \arg\max_{\boldsymbol{m}\in\mathcal{M}} \mathrm{MICL}(\boldsymbol{m}) \text{ où } \mathcal{M} = \left\{\boldsymbol{m} = (g, \boldsymbol{\omega}) : \ 1 \leq g \leq g_{\max} \text{ et } \boldsymbol{\omega} \in \{0,1\}^d\right\}. \tag{10}$$

Nous notons $\mathcal{M}_g$ la restriction de $\mathcal{M}$ au sous-ensemble des modèles à g classes et $\boldsymbol{m}^\star_g$ le modèle maximisant le critère MICL parmi $\mathcal{M}$. Ainsi,

$$\boldsymbol{m}^\star_g = \arg\max_{\boldsymbol{m}\in\mathcal{M}_g} \mathrm{MICL}(\boldsymbol{m}) \text{ avec } \mathcal{M}_g = \{(g, \boldsymbol{\omega}) : \boldsymbol{\omega} \in \{0,1\}^d\}. \tag{11}$$

Par conséquent, l'estimation de $\boldsymbol{m}^\star_g$ pour $g = 1, \ldots, g_{\max}$ permet de déterminer $\boldsymbol{m}^\star$ car

$$\boldsymbol{m}^\star = \arg\max_{g=1,\ldots,g_{\max}} \mathrm{MICL}(\boldsymbol{m}^\star_g). \tag{12}$$

Pour $g = 1, \ldots, g_{\max}$, $\boldsymbol{m}_g^\star$ est estimé par un algorithme d'optimisation alternée. Partant d'un point initial $(\mathbf{z}^{[0]}, \boldsymbol{m}^{[0]})$ avec $\boldsymbol{m}^{[0]} \in \mathcal{M}_g$, l'algorithme alterne entre deux optimisations de la vraisemblance complétée intégrée : la maximisation en $\mathbf{z}$ conditionnellement à $(\mathbf{x}, \boldsymbol{m})$ et la maximisation en $\boldsymbol{m}$ conditionnellement à $(\mathbf{x}, \mathbf{z})$. Ainsi, son itération $[r]$ s'écrit :
Étape partition : estimer $\mathbf{z}^{[r]}$ tel que

$$\ln p(\mathbf{x}, \mathbf{z}^{[r]} | \boldsymbol{m}^{[r]}) \geq \ln p(\mathbf{x}, \mathbf{z}^{[r-1]} | \boldsymbol{m}^{[r]}).$$

Étape modèle : estimer $\boldsymbol{m}^{[r+1]} = \arg\max_{\boldsymbol{m} \in \mathcal{M}_g} \ln p(\mathbf{x}, \mathbf{z}^{[r]} | \boldsymbol{m})$, d'où

$$\boldsymbol{m}^{[r+1]} = (g, \boldsymbol{\omega}^{[r+1]}) \text{ avec } \omega_j^{[r+1]} = \underset{\omega_j \in \{0,1\}}{\arg\max}\, p(\mathbf{x}_{\bullet j} | g, \omega_j, \mathbf{z}^{[r]}).$$

L'étape partition s'effectue par une méthode itérative où chaque itération optimise l'affectation de classe d'un individu tiré aléatoirement. L'algorithme général converge en un optimum local de $\ln p(\mathbf{x}, \mathbf{z} | \boldsymbol{m})$. Il est donc nécessaire d'effectuer plusieurs initialisations pour s'assurer de la convergence vers $\boldsymbol{m}_g^\star$. Cependant, les expériences numériques montrent que le nombre d'optima locaux est suffisamment faible pour permettre un bon comportement de l'approche proposée.

4 Applications

Nous comparons notre approche (notée MS) à l'approche de régularisation de Witten and Tibshirani (2010) (notée WT) et à l'approche probabiliste de Raftery and Dean (2006) (notée RD). Les résultats de MS ont été obtenus par le package *VarSelLCM*[1] avec les options par défaut. Les résultats de WT ont été obtenus par le package *sparcl* avec l'option *wbound=seq(1.1,25,len=30)*, tandis que ceux de RD ont été obtenus par le package *clustvarsel* avec l'algorithme *headlong* dans la direction *forward*. Le tableau 1 présente les résultats obtenus par les trois méthodes. Notons que WT a été utilisé en considérant uniquement le bon nombre de classes pour palier à l'absence de critères d'information. Sur ces applications, WT sélectionne moins de variables que les autres méthodes mais cela dégrade fortement la qualité de la partition (ARI).

Concernant les méthodes probabilistes, RD surestime le nombre de classes tandis que MS le retrouve sur ces trois applications. De plus, MS est généralement plus rapide que RD. Sur *coffee*, MS obtient une meilleure partition en sélectionnant moins de variables. Sur *SRBCT*, la partition retournée par MS est moins bonne que celle de RD. Cependant, MS obtient une meilleure adéquation aux données au sens du BIC, bien qu'il n'ait pas pour objectif de maximiser ce critère. Cela s'explique par la sous-optimalité de la procédure pas-à-pas utilisée par RD. Enfin, sur *wine* et *golub*, en sélectionnant plus de variables, MS obtient une meilleure partition et une meilleure adéquation aux données.

1. Téléchargeable à l'url `https://r-forge.r-project.org/R/?group_id=2011`

Données	$d/n/g$	Méthode	$\hat{g}$	BIC	MICL	NVD	ARI	Tps
coffee	12/43/2	WT	·	·	·	2	0.37	0.33
		RD	3	-521	-899	6	0.38	0.72
		MS	2	-532	-871	4	1.00	0.11
wine	13/178/3	WT	·	·	·	2	0.54	0.66
		RD	4	-3628	-4127	6	0.65	1.39
		MS	3	-3619	-3934	8	0.77	1.58
SRBCT	2308/83/4	WT	·	·	·	3	0.00	18.56
		RD	6	-123818	-138501	34	0.30	386.71
		MS	4	-111458	-126975	931	0.04	71.81
golub	3051/83/2	WT	·	·	·	2	0.37	15.80
		RD	4	-95176	-108074	8	0.00	276.16
		MS	2	-92145	-106358	466	0.79	9.00

TABLE 1 – Résultats des méthodes : nombre de classes estimé ($\hat{g}$), BIC, MICL, nombre de variables discriminantes (NVD), Adjusted Rand Index (ARI) et temps en seconde (Tps).

5 Discussion

MICL permet de répondre à une problématique de choix de modèle en contournant l'estimation des paramètres par l'utilisation d'un estimateur de partition. Lors des applications, cette méthode a montré son intérêt vis-à-vis du modèle estimé (nombre de classes, adéquation aux données, nombre de variables sélectionnées), de la partition associée (ARI) et du temps de calcul. L'extension au cas d'un mélange de lois de la famille exponentielle est à l'étude, ainsi que la modélisation de variables redondantes.

Références

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7) :719–725.

Marbac, M. and Sedki, M. (2015). Variable selection for model-based clustering using the integrated complete-data likelihood. *Preprint, arXiv :1501.06314*.

Raftery, A. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473) :168–178.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2) :461–464.

Witten, D. and Tibshirani, R. (2010). A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association*, 105(490) :713–726.