Sélection de variables en classification non-supervisée de données catégorielles

Matthieu Marbac[1] & Mohammed Sedki[2]

[1] *Inria Lille, matthieu.marbac-lourdelle@inria.fr*
[2] *Université Paris-Sud, Inserm U1181, mohammed.sedki@u-psud.fr*

Résumé. Nous présentons deux méthodes de sélection de variables, pour un clustering de données catégorielles en grande dimension fait par le modèle des classes latentes. La première approche s'effectue par le critère BIC, maximisé par une version modifiée de l'algorithme EM. Ainsi, la sélection de modèle et l'estimation des paramètres sont faites simultanément. Afin de palier aux problèmes de BIC (propriété de convergence asymptotique et non prise en compte de l'objectif de clustering), le critère MICL peut être utilisé. Ce critère, basé sur la forme explicite de la vraisemblance complétée intégrée, permet d'effectuer la sélection de modèle préalablement à l'estimation des paramètres. La maximisation de MICL est faite par un algorithme d'optimisation alternée. L'intérêt de la procédure est illustré sur des données de génétique des populations composées de 1235 observations décrites par 160470 SNPs.

Mots-clés. Classification, modéle de mélange, sélection de modèle.

Abstract. We present two methods for selecting variables in a model-based clustering of high-dimensional categorical data, done by the latent class model. The first way, for selecting variables, consists in selecting the BIC by a modified version of the EM algorithm. Thus, model selection and parameter inference are done simultaneously. To circumvent the issues of the BIC (aymptotic property of convergence and no modelling of the clustering aim), the MICL can be used. This information criterion, based on the closed-form of the intergrated complete-data likelihood, permits to achieve model selection before parameter inference. The maximization of the MICL is carried out via an algorithm of alternate optimization. The interest of the procedure is illustrated on a dataset of population genetics describing 1235 observations of 160470 SNPs.

Keywords. Clustering, mixture model, 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.

Le modèle des classes latentes (mélange de distributions multinomiale) est l'outil classique pour classifier des données catégorielles (Goodman, 1974). Récement, White et al.

(2016) ont proposé une approche bayésienne pour sélectionner les variables dans ce modèle. Cependant, cette approche est trop chronophage pour considèrer des données en grandes dimensions. Dans un premier temps, nous montrons que la recherche de modèle, au sens de BIC (Schwarz, 1978), peut se faire simultanément à l'estimation des paramètres, par l'utilisation d'un algorithme EM modifié (Green, P.J., 1990).

Dans un cadre de clustering, il peut être interessant d'utiliser des critères de choix de modèle basés sur la vraisemblance complétée intégrée. En effet, ces critères sont non-asymptotiques et permettent de prendre en compte l'objectif de clustering (Biernacki, C. and Celeux, G. and Govaert, G., 2010). On montre ici comment faire la sélection de modèle au sens du critère MICL (Marbac and Sedki, 2016), qui possède une forme explicite lorsque des priors conjugués sont utilisés. Il 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 présente le contexte de sélection de variables pour le modèle des classes latentes. La partie 3 montre comment effectuer la sélection de variables par les critères BIC et MICL. La partie 4 illustre l'intérêt de la sélection de variables sur des données génomiques. Une discussion est menée dans la partie 5.

2 Sélection de variables pour les classes latentes

Les n observations à classer $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ sont décrites par d variables catégorielles, où la variable j possède m_j modalités. L'observation i est décrite par le vecteur $\boldsymbol{x}_i = (x_{ijh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$ avec $x_{ijh} = 1$ si l'observation i prend la modalité h pour la variable j, et $x_{ijh} = 0$ sinon. Le modèle des classes latentes considère que les observations sont indépendantes et générées par un modèle de mélange à g composantes, défini par la fonction de distribution de probabilité (fdp):

$$f(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \prod_{j=1}^{d} f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj}) \text{ avec } f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj}) = \prod_{h=1}^{m_j} (\alpha_{kjh})^{x_{ijh}}, \tag{1}$$

où $\boldsymbol{m}$ spécifie le modèle, où $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ regroupe l'ensemble des paramètres, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ étant le vecteur des proportions défini sur le simplex de taille g, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kj}; k = 1, \ldots, g; j = 1, \ldots, d)$, $\boldsymbol{\alpha}_{kj} = (\alpha_{kjh}; h = 1, \ldots, m_j)$, α_{kjh} étant la probabilité d'observer la modalité h pour la variable j dans la composante k. Une variable est non pertinente pour la classification si sa distribution marginale est égale pour toutes les classes. Le vecteur binaire $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ indique le rôle des variables dans la classification

avec $\omega_j = 1$ si la variable j est pertinente pour la classification et $\omega_j = 0$ sinon. Ainsi,

$$\forall j \in \{j' : \omega_{j'} = 0\},\ \forall h \in \{1, \ldots, m_j\},\ \alpha_{1jh} = \ldots = \alpha_{gjh}. \tag{2}$$

Un modèle $\boldsymbol{m} = (g, \boldsymbol{\omega})$ est alors défini par le nombre de composantes et le rôle des variables. La vraisemblance observée s'écrit

$$p(\mathbf{x}|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta}). \tag{3}$$

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. La vraisemblance complétée s'écrit alors

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{g} \left(\pi_k \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_{kjh})^{x_{ijh}}\right)^{z_{ik}}. \tag{4}$$

3 Sélection de modèle

3.1 Sélection de modèle par le critère BIC

Considérant une borne $g_{\max}$ sur le nombre maximal de composantes, l'espace des modèles en compétition est déterminé par $\mathcal{M} = \left\{\boldsymbol{m} = (g, \boldsymbol{\omega}) :\ g \in \{1, \ldots, g_{\max}\} \text{ et } \boldsymbol{\omega} \in \{0, 1\}^d\right\}$. On cherche le modèle $\hat{\boldsymbol{m}}$ qui maximise le critère BIC, ainsi

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m} \in \mathcal{M}}{\arg\max}\, \mathrm{BIC}(\boldsymbol{m}) \text{ avec } \mathrm{BIC}(\boldsymbol{m}) = \ln p(\mathbf{x}|\boldsymbol{m}, \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) - \frac{\nu_{\boldsymbol{m}}}{2} \ln n, \tag{5}$$

où $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$ est l'estimateur du maximum de vraisemblance associé au modèle $\boldsymbol{m}$ et où $\nu_{\boldsymbol{m}} = (g-1) + \sum_{j=1}^{d} (m_j - 1) \times ((g-1)\boldsymbol{\omega}_j + 1)$ est le nombre de paramètres libres de ce modèle. Pour estimer $\hat{\boldsymbol{m}}$, on recherche, pour $g = 1, \ldots, g_{\max}$, le modèle $\hat{\boldsymbol{m}}_g$ à g composantes qui maximise BIC.

L'estimation de $\hat{\boldsymbol{m}}_g$ se fait par un algorithme EM modifié (Green, P.J., 1990) permettant de maximiser la vraisemblance pénalisée (la pénalité étant $\frac{\nu_{\boldsymbol{m}}}{2} \ln n$). Cet algorithme, initalisé en $(\boldsymbol{m}^{[0]}, \boldsymbol{\theta}^{[0]})$ avec $\boldsymbol{m}^{[0]} = (g, \boldsymbol{\omega}^{[0]})$, s'écrit à l'étape $[r]$:
Étape E Calcul de la partition floue

$$t_{ik}^{[r]} := \frac{\pi_k^{[r-1]} \prod_{j=1}^{d} f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj}^{[r-1]})}{\sum_{\ell=1}^{g} \pi_\ell^{[r-1]} \prod_{j=1}^{d} f_{\ell j}(x_{ij}|\boldsymbol{\alpha}_{\ell j}^{[r-1]})},$$

Étape M Maximisation de l'espérance de la vraisemblance complétée pénalisée sur $(\boldsymbol{\omega}, \boldsymbol{\theta})$

$$\omega_j^{[r]} = \begin{cases} 1 & \text{si } \Delta_j^{[r]} > 0 \\ 0 & \text{sinon} \end{cases},\ \pi_k^{[r]} = \frac{n_k^{[r]}}{n} \text{ et } \boldsymbol{\alpha}_{kj}^{[r]} = \begin{cases} \boldsymbol{\alpha}_{kj}^{\star[r]} & \text{si } \omega_j^{[r]} = 1 \\ \tilde{\boldsymbol{\alpha}}_j & \text{sinon} \end{cases},$$

où $\Delta_j = \sum_{k=1}^{g}\sum_{i=1}^{n} t_{ik}^{[r]}\big(\ln f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj}^{\star[r]}) - \ln f_{1j}(x_{ij}|\tilde{\boldsymbol{\alpha}}_j)\big) - (g-1)\times(m_j-1)\frac{\ln n}{2}$, $n_k^{[r]} = \sum_{i=1}^{n} t_{ik}^{[r]}$, $\tilde{\boldsymbol{\alpha}}_j = (\tilde{\alpha}_{jh}; h = 1,\ldots,m_j)$ avec $\tilde{\alpha}_{jh} = \frac{1}{n}\sum_{i=1}^{n} x_{ijh}$ et $\boldsymbol{\alpha}_{kj}^{\star[r]} = (\alpha_{kjh}^{\star[r]}; h = 1,\ldots,m_j)$ avec $\alpha_{kjh}^{\star[r]} = \frac{1}{n_k^{[r]}}\sum_{i=1}^{n} t_{ik}^{[r]} x_{ijh}$.

3.2 Sélection de modèle par le critère MICL

La vraisemblance intégrée se définit alors comme

$$p(\mathbf{x},\mathbf{z}|\boldsymbol{m}) = \int p(\mathbf{x},\mathbf{z}|\boldsymbol{m},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{m})d\boldsymbol{\theta}, \tag{6}$$

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

$$p(\boldsymbol{\theta}|\boldsymbol{m}) = p(\boldsymbol{\pi}|\boldsymbol{m})\prod_{j=1}^{d} p(\boldsymbol{\alpha}_{\bullet j}|\boldsymbol{m}) \text{ et } p(\boldsymbol{\alpha}_{\bullet j}|\boldsymbol{m}) = \Big(\prod_{k=1}^{g} p(\boldsymbol{\alpha}_{kj}|\boldsymbol{m})\Big)^{\omega_j}\Big(p(\boldsymbol{\alpha}_{1j}|\boldsymbol{m})\Big)^{1-\omega_j}, \tag{7}$$

où $\boldsymbol{\alpha}_{\bullet j} = (\alpha_{kj}; k = 1,\ldots,g)$. Des lois conjuguées non informative de Jeffrey (Robert, 2007) sont utilisées, ainsi $\boldsymbol{\pi}|\boldsymbol{m}$ et $\boldsymbol{\alpha}_{kj}|\boldsymbol{m}$ suivent respectivement des distributions de Dirichlet $\mathcal{D}_g(\frac{1}{2},\ldots,\frac{1}{2})$ et $\mathcal{D}_{m_j}(\frac{1}{2},\ldots,\frac{1}{2})$. Ainsi la vraisemblance intégrée a la forme explicite suivante

$$p(\mathbf{x},\mathbf{z}|\boldsymbol{m}) = \frac{\Gamma(\frac{g}{2})}{\Gamma(\frac{1}{2})^g}\frac{\prod_{k=1}^{g}\Gamma(n_k+\frac{1}{2})}{\Gamma(n+\frac{g}{2})}\prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|g,\omega_j,\mathbf{z}), \tag{8}$$

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

$$p(\mathbf{x}_{\bullet j}|g,\omega_j,\mathbf{z}) = \begin{cases} \Big(\frac{\Gamma(\frac{m_j}{2})}{\Gamma(\frac{1}{2})^{m_j}}\Big)^g \prod_{k=1}^{g}\frac{\prod_{h=1}^{m_j}\Gamma\big(\sum_{i=1}^{n} z_{ik}x_{ijh}+\frac{1}{2}\big)}{\Gamma\big(n_k+\frac{m_j}{2}\big)} & \text{if } \omega_j = 1 \\ \frac{\Gamma(\frac{m_j}{2})}{\Gamma(\frac{1}{2})^{m_j}}\frac{\prod_{h=1}^{m_j}\Gamma\big(\sum_{i=1}^{n} x_{ijh}+\frac{1}{2}\big)}{\Gamma\big(n+\frac{m_j}{2}\big)} & \text{if } \omega_j = 0 \end{cases}. \tag{9}$$

On propose d'effectuer la sélection de modèle par le critère MICL (Maximum Integrated Complete-data Likelihood). Ce critère correspond à la plus grande valeur, parmi l'ensemble des partitions, de la vraisemblance intégrée complétée (Marbac and Sedki, 2016). Ainsi, il s'écrit

$$\text{MICL}(\boldsymbol{m}) = \ln p(\mathbf{x},\mathbf{z}^\star_{\boldsymbol{m}}|\boldsymbol{m}) \text{ avec } \mathbf{z}^\star_{\boldsymbol{m}} = \arg\max_{\mathbf{z}} \ln p(\mathbf{x},\mathbf{z}|\boldsymbol{m}). \tag{10}$$

Le critère MICL est similaire à ICL et il hérite de ses propriétés principales (robustesse et ses conditions de consistance). De plus, à la différence de ICL et BIC, MICL 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.

Pour obtenir le modèle $\boldsymbol{m}^\star$ qui maximise le critère MICL dans $\mathcal{M}$, on cherche le modèle $\boldsymbol{m}_g^\star$ à g composantes qui maximise MICL à nombre de classes fixé, pour $g = 1, \ldots, g_{\max}$. Ce dernier est estimé par un algorithme d'optimisation alternée. Partant d'un point initial $(\mathbf{z}^{[0]}, \boldsymbol{m}^{[0]})$ avec $\boldsymbol{m}^{[0]} = (g, \boldsymbol{\omega}^{[0]})$, l'algorithme alterne entre deux étapes: la maximisation en $\mathbf{z}$ conditionnellement à $(\mathbf{x}, \boldsymbol{m})$ et la maximisation en $\boldsymbol{\omega}$ 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: maximiser $\ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{m})$ sur $\boldsymbol{\omega}$, 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 $\boldsymbol{m}_g^\star$.

4 Applications

Pour tenter d'expliquer la forte mortalité observée dans plusieurs populations de Pygmées, des données de polymorphismes SNP de $n = 1235$ individus ont été récoltées. Ainsi chaque individu est décrit par $d = 160470$ SNP (variables catégorielles à 3 modalités). Les individus peuvent être divisés en 2 grandes populations: les *agriculteurs Bantous* (population sédentaire composée de 31 sous-populations) et les *chasseurs-cueilleurs* (population nomade composée de 6 sous-populations). Cependant, à la demande des généticiens des populations, cette information n'a pas été incluse dans le modèle mais sera utilisée pour le valider.

Cependant, puisque BIC est un critère asymptotique, son utilisation est hautement critiquable lorsque $n << d$. Cela illustre l'intérêt des critères de choix de modèle exacts. Ainsi, l'estimation de modèle a été faite par MICL. Ce critère a sélectionné $g^\star = 2$ composantes et 58954 variables (soit 37%). De plus, les variables sélectionnées ont été triées par ordre décroissant de pouvoir discriminant, en utilisant le rapport $\frac{p(\mathbf{x}_{\bullet j}|g,\omega_j=1,\mathbf{z}_{\boldsymbol{m}}^\star)}{p(\mathbf{x}_{\bullet j}|g,\omega_j=0,\mathbf{z}_{\boldsymbol{m}}^\star)}$. Le Tableau 1 présente la matrice de confusion entre la partition estimée et les 2 grandes populations. On constate une cohérence de la partition estimée vis à vis des populations de Pygmees. On remarque que le modèle affecte à la Classe 2 (composée de tous les agriculteurs Bantous) 36 chasseurs-cueilleurs. Ces individus font partis de deux des six sous-populations des chasseurs-cueilleurs. Ils ont donc un patrimoine génétique plus proche des agriculteurs Bantous.

	chasseurs cueilleurs	agriculteurs Bantous
Classe 1	196	0
Classe 2	36	1003

Table 1: Table de confusion entre la partition estimée et les populations de Pygmées

5 Discussion

La sélection de variables en clustering permet d'améliorer la qualité des estimateurs et de faciliter la partition. Pour sélectionner les variables d'un modèle des classes latentes, BIC peut être maximisé directement par un algorithme EM modifié. Ainsi, on évite algorithmes sous-optimaux et chronophages basés sur des comparaisons de modèles deux à deux (*e.g.,* backward, forward...). Cependant, puisque BIC est un critère asymptotique, son utilisation est hautement critiquable lorsque $n << d$. MICL permet de répondre à cette problématique de choix de modèle, en contournant l'estimation des paramètres par l'utilisation d'un estimateur de partition. De plus ce critère permet de prendre en compte l'objectif de clustering. L'extension au cas d'un mélange de lois de la famille exponentielle est à l'étude, ainsi que la modélisation de variables redondantes.

References

Biernacki, C. and Celeux, G. and Govaert, G. (2010). Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11):2991–3002.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Green, P.J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):443–452.

Marbac, M. and Sedki, M. (2016). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, to appear.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.

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

White, A., Wyse, J., and Murphy, T. B. (2016). Bayesian variable selection for latent class analysis using a collapsed gibbs sampler. *Statistics and Computing*, 26(1-2):511–527.