Sélection de mélanges de GLM pour la classification

Olivier Lopez [1] & Xavier Milhaud[2]

[1] *ENSAE-CREST, 3 Avenue Pierre Larousse, 92245 Malakoff Cedex, France; olivier.lopez@ensae.fr*

[2] *ENSAE-CREST, 3 Avenue Pierre Larousse, 92245 Malakoff Cedex, France; xavier.milhaud@ensae.fr*

Résumé. La classification par mélanges finis de modèles linéaires généralisés est un sujet d'actualité qui a été traité dans plusieurs articles récents, par exemple Hennig et Liao (2013) et Hannah et al. (2011). En pratique, l'étape de sélection de modèle est souvent réalisée via l'utilisation de critères de sélection pénalisés connus tels que les critères AIC ou BIC. Les simulations montrent cependant que ces critères ont tendance à surestimer la dimension du modèle sous-jacent, ce qui conduit naturellement à étudier un nouveau critère. Ce critère, ICL^*, a été introduit par Baudry (2009) et sa définition est basée sur un nouveau contraste qui comporte un terme entropique: grâce à des inégalités de concentration, nous prouvons les propriétés de convergence du M-estimateur associé. La consistance du critère de classification ICL^* s'en déduit sous certaines hypothèses classiques liées au terme de pénalité du critère. Une étude par simulations permet de confirmer les résultats théoriques obtenus tout en confortant l'intérêt de la méthode dans une optique de classification.

Mots-clés. sélection de modèle, GLM, vraisemblance classifiante conditionnelle.

Abstract. Model-based clustering from finite mixtures of generalized linear models has recently been of interest in many papers (Hennig & Liao (2013), Hannah et al. (2011)). In practice, the model selection step is usually performed by using AIC or BIC penalized criteria. Though, simulations show that they tend to overestimate the actual dimension of the model. These evidence led us to consider a new criterion close to ICL, firstly introduced in Baudry (2009). Its definition requires to introduce a contrast embedding an entropic term: using concentration inequalities, we derive key properties about the convergence of the associated M-estimator. The consistency of the corresponding classification criterion then follows depending on some classical requirements on the penalty term. Finally a simulation study enables to corroborate our theoretical results, and shows the effectiveness of the method in a clustering perspective.

Keywords. Model Selection, GLM, Conditional Classification Likelihood.

1 Introduction

Depuis une trentaine d'années, l'utilisation des modèles mélanges de modèles linéaires généralisés (GLM) s'est considérablement développée après la parution de Dempster et al. (1977). La popularité de cette classe de modèle provient en grande partie de leur capacité à intégrer des facteurs de risque impactant un phénomène d'intérêt, tout en ayant la capacité de traiter des données fortement hétérogènes. D'autre part, la sélection de modèle est un problème statistique important étudié depuis longtemps dans la littérature, y compris dans le contexte des modèles mélanges (Gassiat et Van Handen (2013)). Cette problématique est d'autant plus cruciale qu'un grand nombre d'applications ont aujourd'hui pour but de décrire de manière explicite la structure d'une population: le nombre de composantes du mélange sous-jacent est donc un point-clef, puisque chaque composante représente idéalement une sous-population à laquelle appartiennent certains individus. Parmi les techniques de sélection, les critères d'information pénalisés restent les plus couramment utilisés. Pourtant, il n'existe généralement pas de résultat théorique sur leur consistance avec des mélanges classiques. D'autres approches comme les tests de ratio de vraisemblance (Azais et al. (2009), Gassiat (2002)) ont alors été développées, mais ces tests sont bien souvent difficiles à utiliser dans la pratique. Dans le contexte des mélanges, McLachlan et Peel (2000) et Fraley et Raftery (1998) s'accordent à dire que le critère BIC donne de meilleurs résultats que l'AIC puisqu'il cherche à minimiser la divergence de Kullback-Leibler à la distribution théorique. Cependant le problème de surestimation de la dimension du modèle sélectionné via ces critères est bien connue, même si la convergence du BIC pour estimer l'ordre d'un mélange gaussien a par ailleurs été démontrée dans Keribin (1999). Cette tendance à la surestimation s'accentue encore davantage lorsque le modèle est mal spécifié. Ces constatations nous ont donc amenés à étudier le critère ICL^*, un dérivé d'ICL (Biernacki (2000)), introduit par Baudry (2009) dont les travaux ont prouvé la consistance pour des mélanges gaussiens. Cependant, aucune propriété semblable n'existe dans le cadre de mélanges de GLM. ICL^* est particulièrement adapté à la classification puisqu'il est basé sur l'utilisation d'un contraste comportant un terme entropique: l'entropie joue un rôle déterminant en pénalisant la vraisemblance du modèle par une quantité liée à la confiance en la classification des individus a posteriori. Plus cette confiance est faible, plus la pénalisation incluse au contraste est importante. L'objectif est ici d'étudier et de démontrer la consistance de ce critère pour la sélection de mélanges de GLM, en vérifiant des propriétés de convergence du M-estimateur associé.

2 L'estimateur du maximum de vraisemblance classifiante conditionnelle

Nous utilisons dans la suite des modèles mélanges discrets: pour une observation $y_j \in R^d$, la densité du mélange vaut

$$\forall n_g \in N^*,\ \forall \psi_g \in (\Pi_g \times \Theta^{n_g}),\quad f(y_j;\psi_g) = \sum_{i=1}^{n_g} \pi_i\, f_i(y_j;\theta_i) \tag{1}$$

o les $f_i(y_j;\theta_i)$ sont les composantes du mélange, et les $\pi_i \in [0,1]$ sont les poids tels que $\sum_{i=1}^{n_g} \pi_i = 1$. On note Π_g l'ensemble des g-tuples $(\pi_1,...,\pi_{n_g})$ qui satisfont cette dernière condition. ψ_g est l'ensemble de paramètres du modèle. L'ensemble des mélanges à n_g composantes que nous considérons est donné par

$$M_g \quad := \quad \Big\{ f(.;\psi_g) = \sum_{i=1}^{n_g} \pi_i\, f_i(.;\theta_i) \quad | \quad \psi_g = (\pi_1,...,\pi_{n_g},\theta_1,...,\theta_{n_g}) \in \Psi_g \Big\}.$$

Pour matérialiser le concept d'optimisation de la vraisemblance classifiante conditionnelle (notée L_{cc} dans la suite), nous définissons au préalable ce contraste qui apparait naturellement lors de l'application de l'algorithme EM sur données complètes (Y_j, Z_j), o Z_j est le "label" des observations. Plus précisément Z_{ij} vaut 1 si l'observation j appartient à la composante i, et 0 sinon. La vraisemblance classifiante conditionnelle s'exprime alors comme une fonction de la vraisemblance des données observées:

$$\forall \psi_g \in \Psi_g, \qquad \ln L_{cc}(\psi_g;\mathbf{Y}) = \ln L(\psi_g;\mathbf{Y}) - Ent(\psi_g;\mathbf{Y}), \tag{2}$$

avec $\mathbf{Y} = (Y_1,...,Y_n)$ est un vecteur de n observations i.i.d., et le terme qui lie les deux log-vraisemblances est proche de ce que l'on appelle communément l'entropie:

$$\forall \psi_g \in \Psi_g, \quad \forall\, Y_j \in R^d, \qquad Ent(\psi_g;Y_j) = -\sum_{i=1}^{n_g} \tau_i(Y_j;\psi_g)\, \ln \tau_i(Y_j;\psi_g).$$

Ici, $\tau_i(Y_j;\psi_g)$ est la probabilité *a posteriori* que l'observation j appartienne à la composante i. L'entropie est maximale en cas d'équiprobabilité ($\tau_1(Y_j;\psi_g) = ... = \tau_{n_g}(Y_j;\psi_g)$), et minimale lorsqu'une des probabilités a posteriori vaut 1. Comme le montre (2), ce terme peut être vu comme une pénalisation de la vraisemblance observée: plus l'erreur potentielle de classification a posteriori des observations via la règle de Bayes est grande, plus la pénalisation est importante (et vice versa). La limite de l'entropie vaut 0 lorsqu'il existe i tel que $\tau_i(Y_j;\psi_g)$ tende vers 0 ou 1. Mais cette entropie n'est pas dérivable en 0, ce qui implique certaines restrictions sur l'espace des paramètres à considérer afin de garantir la convergence du M-estimateur associé à ce contraste. L'estimateur du maximum de vraisemblance classifiante conditionnelle (noté $ML_{cc}E$ dans la suite) est défini par

$$\hat{\psi}_g^{ML_{cc}E} = \arg\max_{\psi_g \in \Psi_g} \frac{1}{n} \sum_{j=1}^{n} \ln L_{cc}(\psi_g;\, y_j), \tag{3}$$

La loi des grands nombres garantit normalement que cet estimateur soit une bonne approximation de $\psi_g^b = \arg\max_{\psi_g \in \Psi_g} E_{f^0}[\ln L_{cc}(\psi_g, \mathbf{Y})]$, o f_0 est la densité théorique (inconnue) de $\mathbf{Y}$. Le $ML_{cc}E$ n'a pas pour objectif de retrouver la distribution théorique, même dans le cas d'un modèle bien spécifié. Son but est de réaliser le meilleur compromis entre qualité d'ajustement aux données et confiance en la classification (a posteriori) des observations à chacune des composantes du mélange. En étudiant le comportement des déviations du processus empirique $\phi(\psi_g; y) = \ln L_{cc}(\psi_g; y) - \ln L_{cc}(\psi_g^b; y)$, nous déterminons des bornes exponentielles pour la convergence de $\hat{\psi}_g^{ML_{cc}E}$ vers ψ_g^b. Nous en déduisons également des vitesses de convergence dans le cas asymptotique.

3 Un nouveau critère de sélection: le critère ICL^*

Pour éviter le problème de surestimation de la dimension du modèle et de son nombre de composantes, Biernacki (2000) définit le critère ICL. Initialement le critère ICL était défini sur la même base que le critère BIC (en particulier la pénalité était celle du BIC), sauf que le contraste considéré était la vraisemblance des données complètes (faisant ainsi apparaitre un terme proche de l'entropie). Le terme entropique est alors considéré comme faisant partie intégrante de la pénalité liée au critère de sélection, mais cette forme de pénalité ne permet pas d'obtenir des résultats de consistance du critère. C'est alors que Baudry (2009) propose de redéfinir le critère ICL en intégrant la partie entropique de cette pénalité dans le contraste, définissant la vraisemblance L_{cc}. Ainsi la pénalité du critère ICL^* modifié est de nouveau identique à celle du BIC, bien que le contraste ainsi que l'estimateur considérés dans ce critère soient clairement différents. Il s'agit donc de sélectionner parmi une collection de m modèles celui qui vérifie

$$M_{ICL^*} = \arg \min_{M_g \in \{M_1, \ldots, M_m\}} \Big(- \ln L_{cc}(\hat{\psi}_g^{ML_{cc}E}) + \frac{K_g}{2} \ln n \Big).$$

Nous montrons dans notre étude la consistance d'un tel critère de sélection dans le contexte des mélanges de GLM. Sélectionner le nombre de composantes d'un mélange par cette procédure conduit à sélectionner le nombre théorique de composantes. Des simulations permettent de confirmer les résultats théoriques du comportement de l'estimateur $ML_{cc}E$ ainsi que du critère ICL^*. Ce critère se révèle effectivement intéressant dans l'optique de classification, et permet bien souvent d'éviter la surestimation de la dimension du modèle mélange de GLM.

Bibliographie

[1] Hennig, C. et Liao, T.F. (2013), How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification, *Journal of the Royal Statistical Society: Series C*, 62, 3, 309–369.

[2] Hannah, L.A. et Blei, D.M. et Powell, W.B. (2011), Dirichlet process mixtures of generalized linear models, *Journal of Machine Learning Research*, 1, 1–33.
[3] Gassiat, E. et Van Handen, R. (2013), Consistent order estimation and minimal penalties, *IEEE Trans. Info. th*, 59, 2, 1115–1128.
[4] Baudry, J.P. (2009), Sélection de modèle pour la classification non supervisée. Choix du nombre de classes., *Thèse de doctorat*, Université Paris Sud XI.
[5] Dempster, A.P. and Laird N.M. and Rubin D.B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, 39, 1–38.
[6] Azais, J.-M. and Gassiat, E. and Mercadier, C. (2009), The likelihood ratio test for general mixture models with possibly structural parameters, *ESAIM P&S*, 13, 301–327.
[7] Gassiat, E. (2002), Likelihood ratio inequalities with applications to various mixtures, *Annales de l'Institut Henri Poincaré*, 38, 897–906.
[8] Fraley, C. and Raftery, A.E. (1998), How many clusters? Which clustering method? answer via model-based cluster analysis, *The Computer Journal*, 41, 8, 578–588.
[9] McLachlan, G. and Peel, D. (2000), Finite Mixture Models, *Wiley Series In Probability and Statistics*, New York.
[10] Keribin, C. (1999), Tests de modèles par maximum de vraisemblance, *Thèse de doctorat*, Université Paris Sud XI.
[11] Biernacki, C. (2000), Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Transactions on PAMI*, 22, 719–725.