Choix de modèles quand la vraisemblance est incalculable

Christine Keribin [1]

[1] *Laboratoire de Mathématiques UMR* 8628*, Université Paris Sud, F*-91405 *Orsay cedex; christine.keribin@math.u-psud.fr*

Résumé. Les critères pénalisés comme le critère BIC sont des méthodes fréquemment utilisées pour la sélection de modèles et qui nécessitent le calcul de la vraisemblance. Malheureusement, il peut arriver que la vraisemblance ne soit pas numériquement calculable, comme c'est le cas par exemple pour le modèle des blocs latents (LBM). LBM est un modèle de mélange pour la classification croisée (co-clustering), permettant la classification non supervisée simultanée des lignes et colonnes de grandes matrices de données. A cause de la structure de dépendance complexe entre les variables d'appartenance à une classe en ligne et en colonne conditionnellement aux observations, il est nécessaire d'opérer des approximations pour calculer l'étape d'estimation de l'algorithme EM, conduisant ainsi à un minorant de la vraisemblance maximisée. Pour la même raison, l'approximation asymptotique usuelle pour définir le critère BIC doit être remise en question. D'un autre côté, le critère de vraisemblance complète intégrée (ICL) peut être calculé de façon exacte pour LBM, mais nécessite d'étudier l'influence d'hyper-paramètres. Les liens entre les deux critères sont analysés et une comparaison avec l'inférence bayésienne est discutée.

Mots-clés. modèle de mélange, classification croisée, modèle des blocs latents, sélection de modèle, BIC, critère ICL

Abstract. Penalised likelihood criteria such as BIC are popular methods for model selection and require to compute the maximised likelihood. Unfortunately, this maximised likelihood can be untractable, as it is the case for the latent block model (LBM). LBM is a mixture model for co-clustering, allowing to perform the simultaneous clustering of rows and columns of large data matrices. Due to the complex dependence between the row and column class membership variables conditionally to the observations, approximations are needed to perform the estimation step of the EM algorithm, leading to a lower bound of the maximised likelihood. For the same reason, the usual asymptotic approximation used to derive BIC is itself questionable. On the other hand, the integrated completed likelihood criterion (ICL) is exactly computed for LBM, but requires to investigate the influence of hyperparameters. Links between both criteria are analyzed and comparison with Bayesian inference is discussed.

Keywords. mixture model, co-clustering, Latent Block Model, model selection, BIC, Integrated Completed Likelihood (ICL)

Bibliographie

[1] Biernacki C., Celeux, G., Govaert G. (2000), Assessing a mixture model for clustering with integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 719–725*
[2] Lebarbier E., Mary-Huard T. (2006), Une introduction au critère BIC : fondements théoriques et interprétation, *Journal de la SFdS, vol. 147, iss. 1, pp. 39-57*
[3] Mattias C., Mariadassou M. (2013), Convergence of the groups posterior distribution in latent or stochastic bloc model, *arXiv:1206.7101v2*
[4] Keribin C., Brault V., Celeux G., Govaert G. (2014), Estimation and Selection for the Latent Block Model on Categorical Data. *Statistics and Computing DOI 10.1007/s11222-014-9472-2*