Le modèle des blocs latents, une méthode régularisée pour la classification en grande dimension

Christine Keribin [1,2] & Christophe Biernacki [3]

[1] *Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France,* christine.keribin@math.u-psud.fr
[2] *INRIA Saclay - Île de France, Équipe* Select
[3] *Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille*

Résumé. Les modèles de mélange sont connus pour être un outil efficace de classification non supervisée quand la dimension des observations est faible, mais échouent en grande dimension à cause d'un manque de parcimonie. Certaines tentatives pour prendre en compte la redondance ou la pertinence des variables se heurtent à des problèmes de complexité explosive. Nous recommandons d'utiliser le modèle des blocs latents, un modèle probabiliste de classification croisée simultanée des individus et des variables, pour classifier des individus en grande dimension. Nous illustrons de façon empirique le compromis biais-variance de la stratégie de classification croisée dans des scénarii en grande dimension comportant des caractéristiques de redondance et de non pertinence et nous montrons son effet régularisateur sur la classification simple.

Mots-clés. classification, classification croisée, modèle des blocs latents, grande dimension

Abstract. Standard model-based clustering is known to be very efficient for low dimensional data sets, but it fails for properly addressing high dimension (HD) ones, where it suffers from both statistical and computational drawbacks. In order to counterbalance this curse of dimensionality, some proposals have been made to take into account redundancy and features utility, but related models are not suitable for too many variables. We advocate that the latent bloc model, a probabilistic model for co-clustering, is of particular interest to perform HD clustering of individuals even if it is not its primary function. We illustrate in an empirical manner the trade-off bias-variance of the co-clustering strategy in scenarii involving HD fundaments (correlated variables, irrelevant variables) and show the ability of co-clustering to outperform simple mixture row-clustering.

Keywords. clustering, co-clustering, latent block model, high dimension

1 Clustering par classification simple

La classification non supervisée ou *clustering* permet de définir une partition en K groupes ou *clusters* d'un ensemble de n observations $(x_1, \ldots, x_n) \in \mathcal{X}^n$ d'un espace $\mathcal{X}$ de dimension

d. Le clustering peut être utilisé pour explorer les données, les résumer ou améliorer la flexibilité d'étapes ultérieures de prédiction. Il compte de nombreuses applications en marketing, biologie, imagerie ou fouille de texte par exemple.

Parmi les méthodes de clustering, l'approche par modèle de mélange formule la recherche des clusters comme un problème d'estimation. Le modèle de mélange paramétrique (Biernacki (2017)) est une méthode probabiliste qui considère les observations x_i comme issues de n variables aléatoires indépendantes et de même loi de densité

$$f(.;\theta) = \sum_{k=1}^{K} \pi_k \varphi(.;\alpha_k); \quad \sum_{i=1}^{K} \pi_k = 1, \pi_k \geq 0$$

où les paramètres $\pi_1, \ldots, \pi_K$, représentent les *poids* du mélange, et $\varphi(.;\alpha_k)$ la densité de la k-ième *composante* du mélange de paramètre α_k. Les densités des composantes sont en général prises dans une même famille paramétrique connue, et les lois des composantes ne différent que par la valeur de leur paramètre α_k. Il s'agit d'estimer le paramètre $\theta = (\pi_1, \ldots, \pi_K, \alpha_1, \ldots, \alpha_K)$, et pour chaque observation x_i la composante ou classe dont telle est issue. On définit ainsi les labels $z_i \in \{1, \ldots, K\}$ d'appartenance aux classes, variables latentes inconnues indépendantes et de même loi multinomiale $\mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K))$, et la matrice de classification $\mathbf{z} = (z_{ik})$ où $z_{ik} = 1$ si et seulement si la i-ème observation se trouve dans la classe k. La règle de classification de Bayes affecte à chaque observation le groupe qui maximise les probabilités conditionnelles $\tau_{ik} = \mathbb{P}(z_{ik} = 1|x_i;\theta)$. L'algorithme EM de Dempster et al (1977) permet de calculer l'estimateur du maximum de vraisemblance $\widehat{\theta}$ de θ et les probabilités conditionnelles estimées $\widehat{\tau}_{ik} = \mathbb{P}(z_{ik} = 1|x_i;\widehat{\theta})$. Chaque observation est affectée à une classe suivant la règle du maximum *a posteriori* : $\widehat{z}_i = \arg\max_k \widehat{\tau}_{ik}$.

Si cette méthode probabiliste est reconnue pour être très efficace pour les données de dimension d faible, elle se heurte à des problèmes computationnels et statistiques en grande dimension. Dans le cas des variables gaussiennes, Maugis (2009) a proposé l'algorithme SelvarClust, permettant de classer les variables en trois groupes (variables informatives, non informatives, et linéairement dépendantes) tandis que Fop et al (2018) en font deux groupes (variables informatives, non informatives) pour l'analyse de classes latentes. Mais ces méthodes échouent en grande dimension à cause d'un manque de parcimonie et d'une combinatoire explosive.

2 Clustering par classification croisée

Le modèle des blocs latents, permettant de déterminer une partition simultanée des lignes et des colonnes d'une matrice, possède une modélisation très parcimonieuse. Cette propriété en fait un candidat naturellement régularisé pour traiter le clustering en grande

dimension, même si ce n'est pas sa fonction initiale. Le clustering des colonnes peut être vu comme une stratégie de contrôle drastique de la variance de l'estimation, mais elle engendre du biais, et il est intéressant d'en étudier l'effet sur la classification.

Le modèle des blocs latents (LBM, Govaert et Nadif (2013)), est une extension du modèle de mélange à la classification croisée. Il postule :

(i) l'existence d'une partition latente des observations $\mathbf{x} = (x_{ij})$ sous forme d'un produit cartésien d'une classification des lignes $\mathbf{z} = (z_{ik})$ en K clusters par une classification des colonnes $\mathbf{w} = (w_{j\ell})$ en L clusters,

(ii) l'indépendance des variables latentes $\mathbf{z}$ et $\mathbf{w}$ qui sont respectivement iid de loi multinomiale $z_i \sim \mathcal{M}(1, \boldsymbol{\pi})$ et $w_j \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_L))$, avec $\rho_\ell = \mathbb{P}(w_{j\ell} = 1)$ pour $\ell = 1, \ldots, L$,

(iii) l'indépendance, conditionnelle aux labels $(\mathbf{z}, \mathbf{w})$, des cellules x_{ij}, variables aléatoires de loi de densité $\varphi(.; \alpha_{k\ell})$ dépendant uniquement du bloc (k, ℓ) d'appartenance.

Sous ces conditions, la vraisemblance s'écrit

$$p(\mathbf{x}; \theta) = \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik} w_{j\ell}}$$

où $\mathcal{Z} \times \mathcal{W}$ représente l'ensemble de toutes les partitions croisées de $n \times d$ cellules en $K \times L$ blocs. Le calcul de la vraisemblance ou de son logarithme nécessite la somme de $K^n L^d$ termes non factorisables, ce qui n'est pas réalisable en temps raisonnable avec les moyens informatiques existants, même pour un faible nombre d'observations et de blocs. Différentes stratégies d'estimation existent (Keribin et al (2015)) et permettent d'accéder à l'estimation des probabilités conditionnelles des labels en ligne et en colonne. Les labels inconnus $(\mathbf{z}, \mathbf{w})$, et donc en particulier une partition des individus, sont calculés par la règle du maximum *a posteriori*.

Nous comparons de façon empirique la classification des individus par mélange simple et par classification croisée et discutons sur des scenarii illustratifs l'effet régularisateur et bénéfique du co-clustering, surpassant les performances d'une classification par mélange simple.

3 Comparaison

Soit ν la dimension du paramètre α_k de la densité conditionnelle $\varphi(.; \alpha_k)$. La dimension du modèle de mélange simple est $(K-1) + Kd\nu$ sous l'hypothèse d'indépendance conditionnelle, et augmente linéairement avec la dimension d. Ainsi, la qualité de l'estimation

du paramètre θ décroit avec le nombre de variables d à nombre d'observations n fixé et pâtit donc de l'augmentation importante de la variance d'estimation (Biernacki et Maugis (2017)). Cependant, les composantes peuvent être de plus en plus séparées, en particulier si les variables ajoutées sont informatives, la dimension pouvant cette fois apporter un effet bénéfique. Un LBM à $K \times L$ blocs est quant à lui de dimension $(K-1)+(L-1)+KL\nu$, indépendant du nombre de variables d, ce qui illustre son extrême parcimonie par rapport au modèle de mélange. Ceci reste vrai quand L varie raisonnablement en fonction de d, par exemple, $L = o(d)$.

Nous illustrons empiriquement la comparaison de l'erreur de classification de ces deux méthodes, en fonction de deux caractères importants (redondance, pertinence) des variables dans le cas d'un mélange de deux gaussiennes d-variées, de même variance Σ :

$$x_i|_{z_i=k} \sim \mathcal{N}_d(\mu_k, \Sigma); k \in \{1, 2\};\ \pi_1 = \pi_2 = 0.5.$$

Dans ce cas, l'erreur de Bayes vaut $\mathbb{E}(\mathbf{z} \neq \mathbf{z}^B) = F^*(-\frac{1}{2}||\mu_1 - \mu_2||_{\Sigma^{-1}})$ où F^* représente la fonction de répartition de la loi gaussienne centrée réduite univariée. Nous considérons les scénarii suivants :

C1 (basique) : toutes les variables sont informatives, et indépendantes conditionnellement au groupe: $\mu_1 = 0_d$, $\mu_2 = 1_d$, $\Sigma = \mathrm{I}_d$. L'erreur de Bayes $\mathbb{E}(\mathbf{z} \neq \mathbf{z}^B) = F^*(-\sqrt{d}/2)$ tend vers 0 avec d.

C2 (perte de pertinence) : les variables sont de moins en moins informatives : $\mu_1 = 0_d$, $\mu_2 = (1, \frac{1}{2^2}, \ldots, \frac{1}{d^2})$, $\Sigma = \mathrm{I}_d$. L'erreur de Bayes décroit plus lentement, et tend vers $F^*(-\sqrt{\pi^4/90}/2) \neq 0$.

C3 (corrélation): les variables sont dépendantes conditionnellement au groupe, de même coefficient de corrélation $c > 0$, et elles sont toutes informatives : $\mu_1 = 0_d$, $\mu_2 = 1_d$, $\Sigma = \Sigma(c)$. L'erreur de Bayes décroit vers la limite $F^*(-1/(2\sqrt{c}))$, non nulle quand la corrélation est non nulle.

C4 (redondance exacte) : un jeu de données de taille $n/2$ est tiré suivant C1, puis dupliqué. Même comportement de l'erreur de Bayes qu'en C1.

C5 (variables non identiquement distribuées) : $\mu_1 \sim \mathcal{N}_d(0, \mathrm{I}_d)$, $\mu_2 \sim \mathcal{N}_d(1_{\sqrt{d}}, 0_{d-\sqrt{d}}, \mathrm{I}_d)$. Même comportement de l'erreur de Bayes qu'en C1.

C6 (variables corrélées et de moins en moins pertinentes) : C2 avec $\Sigma = \Sigma(c)$. Même comportement de l'erreur de Bayes qu'en C3.

Quatre méthodes de classification sont opérées: *MixDiag* utilise un mélange de deux lois gaussiennes de d variables indépendantes (`Gaussian_pk_Lk_Bk` dans la terminologie Lebret et al (2015)), *MixCor* un mélange de deux lois gaussiennes de d variables de

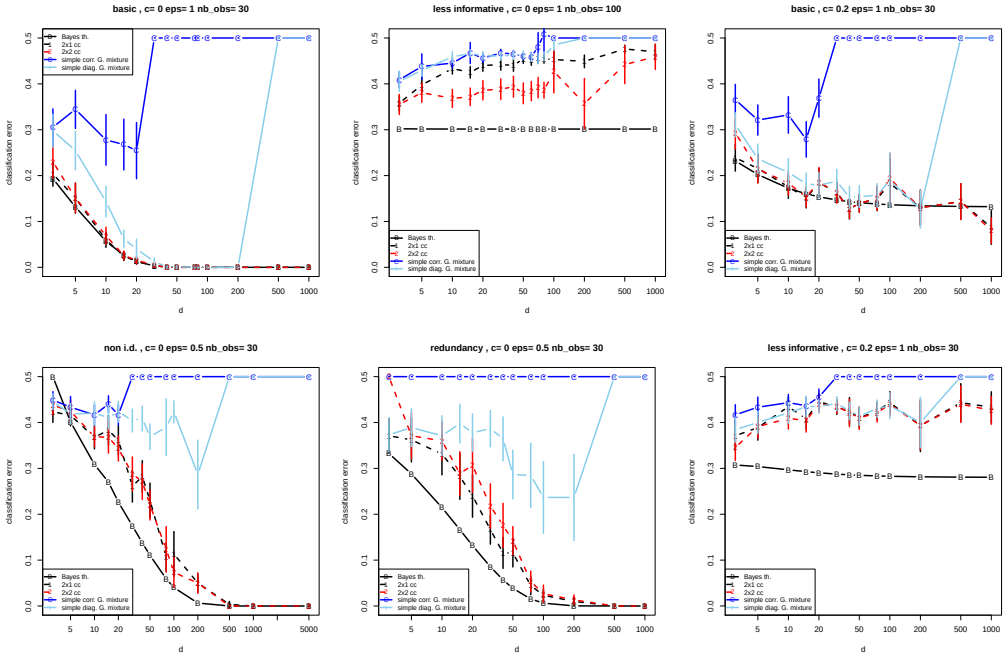

Figure 1: Influence du nombre de variables sur l'erreur de classification pour les différents scénarii. Première ligne C1 (gauche), C2 (centre), C3 (droite); deuxième ligne C4 (gauche), C5 (centre), C6 (droite). Sont représentées l'erreur de Bayes (trait plein noir), l'erreur empirique avec la méthode *MixDiag* (trait plein cyan), avec *MixCor* (trait plein bleu sombre), avec *cc21* (pointillé noir), avec *cc22* (pointillé rouge). Moyenne faite sur 30 échantillons de 2×15 observations, les barres verticales représentent l'intervalle de confiance de niveau 95%

covariance quelconque (`Gaussian_pk_Lk_C`), *cc21* un co-clustering à $(K = 2 \times L = 1)$ blocs, *cc22* un co-clustering à $(K = 2 \times L = 2)$ blocs. L'erreur de classification des scenarii C1 à C6, moyennée sur $B = 30$ échantillons de taille $2 \times 15 = 30$, est représentée pour ces différentes méthodes en figure 1.

Dans les scénarii proposés, le clustering *MixCor* (trait plein bleu sombre) est toujours sans biais, mais sa variance augmente rapidement avec d, et cette méthode n'est jamais compétitive. Le clustering *MixDiag* (trait plein cyan) est non baisé, sauf pour les scenarii C3 et C6. Deux comportements sont intéressants à noter : l'erreur de classification suit l'erreur de Bayes dans scenario C3 (variables corrélées) alors que l'estimation est biaisée : dans ce cas, le biais ne met pas en péril la séparation. En revanche, quand il y a perte de pertinence des variables (C2 et C6), l'erreur de classification ne suit plus celle de Bayes, mais augmente avec la dimension, révélant à nouveau un effet variance. Si *MixDiag* tient son rang vis à vis du co-clustering pour des dimensions modérées, cette méthode diverge

quand le nombre de variables devient trop important, ce qui la rend alors inutilisable.

La méthode de classification par co-clustering *cc22* (trait pointillé rouge) est biaisée sauf pour les scenarii C1 et C5, tandis que *cc21* (trait plein cyan) est toujours biaisée. Ces méthodes s'avèrent très robustes. L'erreur de classification converge souvent vers l'erreur de Bayes (C1, C3, C4, C5), que la méthode soit biaisée ou non. Dans les scenarii C2 et C6, le biais est trop important, et il faudrait augmenter la taille de la partition des variables.

Prolongements vers une analyse théorique Ces résultats prometteurs permettent d'illustrer le co-clustering comme un outil de régularisation pour effectuer un clustering en grande dimension. Il propose un modèle extrêmement parcimonieux, qui, bien que généralement biaisé, assure souvent une excellente performance, et surpasse celle d'un mélange simple. Il a par exemple la propriété native de regrouper les variables exactement redondantes, et peut définir une classe de variables non informatives. Ces résultats préliminaires sont à prolonger en étudiant les propriétés et performances de la classification par co-clustering de façon théorique, en étendant également l'étude à plus de deux groupes d'individus. Pour ce faire, il sera nécessaire de préciser la notion de redondance et de pertinence d'une variable vis à vis du processus de classification.

Bibliographie

Biernacki C. (2017) *Mixture models* dans *Choix de modèles et agrégation*, Editeurs J-J. Droesbecke. G. Saporta C. Thomas-Agnan, Technip

Biernacki C. et Maugis C. (2017) *High-dimensional clustering* dans *Choix de modèles et agrégation*, Editeurs J-J. Droesbecke. G. Saporta C. Thomas-Agnan, Technip

Dempster, A.P.; Laird, N.M. et Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1), 1–38.

Fop, M., et Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, 12, 18-65.

Govaert, G., et Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons.

Keribin, C., Brault, V., Celeux, G., et Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6), 1201-1216.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., et Govaert, G. (2014). *Rmixmod: the R package of the model-based unsupervised, supervised and semi-supervised classification mixmod library*. Journal of Statistical Software, 67(6), 241-270.

Maugis, C., Celeux, G., et Martin-Magniette, M.-L. (2009) Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, vol. 53, no 11, p. 3872-3882.