Fused-ANOVA : une méthode de clustering en grande dimension

Audrey Hulot, Julien Chiquet, Florence Jaffrezic, Guillem Rigaill

Fused-ANOVA : une méthode de clustering en grande dimension

Audrey Hulot[1,2,3] & Julien Chiquet[2] & Florence Jaffrézic[1] & Guillem Rigaill[4,5]

[1] *INRA, UMR GABI, F-78352 Jouy-en-Josas, France, audrey.hulot@inra.fr and florence.jaffrezic@inra.fr* — [2] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France, julien.chiquet@agroparistech.fr* — [3] *INSERM UMR 1173, Université de Versailles Saint-Quentin-en-Yvelines, 78180 France* — [4] *Institute of Plant Sciences Paris-Saclay, UMR 9213/UMR 1403, CNR, INRA, Université Paris-SUD, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité* — [5] *Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), Université d'Evry Val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA, guillem.rigaill@inra.fr*

Résumé. Dans le domaine de la génomique, une première approche pour analyser les données est de réaliser un clustering pour identifier des structures dans les données. Étudier des clusters de gènes co-exprimés permet de trouver des entités pouvant expliquer un processus biologique particulier. Le développement récent des techniques de séquençage haut-débit amène à étudier des jeux de données de grande dimension, sur lesquels les algorithmes classiques de clustering ne sont pas optimaux en termes de temps de calcul et de mémoire utilisée. Nous présentons ici une nouvelle méthode de clustering adaptée à la grande dimension ainsi que quelques résultats sur un jeu de données simulé.

Mots-clés. Grande dimension, Classification non supervisée, Biostatistique, clustering spectral

Abstract. A common first step in the analysis of genomic data is to use a clustering method to identify structures in the data. Grouping genes that have similar patterns in expression can lead to the discovery of interesting entities to explain a certain biological process. The recent developments of high-throughput sequencing techniques lead to high-dimensional datasets, on which the usual clustering algorithms are not performing well in terms of computing time and memory use. We present here a new clustering method that is adapted to high-dimensional data and some results on a simulated dataset.

Keywords. High-dimensional data, Unsupervised classification, Biostatistics, Spectral clustering

1 Introduction

Les méthodes de classification non supervisée sont très utilisées dans le domaine de la génomique pour identifier une structure dans les données, en particulier les méthodes

basées sur une hiérarchie (clustering hiérarchique) ou sur une partition des données (k-means). Cependant, ces méthodes présentent des complexités relativement élevées.

Les algorithmes de classification hiérarchique sont fondés sur le calcul des distances/ dissimilarités entre toutes les variables, menant à une complexité au minimum en $\mathcal{O}(n^2)$. L'algorithme des k-means a une complexité en $\mathcal{O}(ndk)$ si l'on considère un nombre d'itérations borné, où d est la dimension et k le nombre de groupes que l'on cherche.

Dans un contexte de grande dimension ($d \ll n$), comme c'est le cas actuellement en génomique, les méthodes de type clustering hiérarchique sont parfois difficilement applicables. En ce qui concerne les k-means, sa complexité modérée permet toujours son application mais on perd la structure d'arbre de la classification hiérarchique. De plus, l'algorithme est sensible à l'initialisation et nécessite d'être relancé plusieurs fois.

La méthode fused-ANOVA [Chiquet et al., 2017] est une version contrainte de l'ANOVA utilisant une pénalité de type fused-LASSO, dans la même veine que Cas-ANOVA [Bondell and Reich, 2009], mais avec de meilleures propriétés statistiques et computationnelles. Fused-ANOVA permet de réaliser un clustering sur un vecteur univarié de dizaines de millions d'entrées en quelques secondes. Une limitation de la méthode est qu'elle ne traite que de problèmes univariés.

Nous présentons ici une version multidimensionnelle de fused-ANOVA, basée sur l'agrégation des arbres obtenus sur chacune des dimensions. Pour réduire la dimension des données, mais aussi isoler l'information pertinente qu'elles contiennent, nous utilisons la théorie du clustering spectral et la méthode de Nyström en amont du fused-ANOVA multidimensionnel. Les résultats de cette méthode seront illustrés sur un jeu de données simulé.

2 Fused-ANOVA multidimensionnel

On considère dans cette partie x_{ij}, les observations d'une variable aléatoire, où $i \in \{1, ..., n\}$ et $j \in \{1, ..., d\}$. Pour le cadre appliqué de ce travail, on considère que x_{ij} est l'expression du gène i dans l'échantillon j. On note X_i le vecteur des observations du gène i et X la matrice composée des vecteurs $X_1, ..., X_n$. On cherche à résoudre le problème d'optimisation suivant, motivé par la MANOVA et le clustering hiérarchique :

$$\underset{\beta \in \mathbb{R}^{Kd}}{\text{minimiser}} \frac{1}{2} \sum_{i=1}^{n} \left\| X_i - \beta_{\kappa(i)} \right\|_2^2 + \lambda \sum_{k,l:k \neq l} \omega_{kl} \Omega(\beta_k - \beta_l)$$

où Ω désigne une norme, ω_{kl} des poids à fixer et β_k le coefficient du groupe k, $k \in \{1, ...K\}$. κ désigne la fonction d'*a priori* sur les groupes d'individus (ici les gènes). Dans le cas où il n'y a pas d'*a priori*, cette fonction est l'identité, et on retrouve un problème de clustering avec $K = n$. Dans [Hocking et al., 2011, Chi and Lange, 2015], le choix se concentre essentiellement autour de la norme ℓ_2, ce qui permet l'agrégation de toutes les dimensions en un seul clustering, dont le nombre de groupes dépend du paramètre λ.

Fused-ANOVA unidimensionnel. Dans [Chiquet et al., 2017], nous avons étudié plus précisément le cas de la norme ℓ_1 qui rend le problème séparable entre les dimensions, et on se ramène ainsi au traitement de p problèmes univariés indépendants. En particulier, nous avons proposé les poids suivants, dits *exponentially adaptative* :

$$\omega_{kl} = n_k n_l exp\{-\gamma\sqrt(n)|\bar{X}_k - \bar{X}_l|\}, \gamma > 0,$$

Ces poids satisfont les deux propriétés suivantes : *i*) une fois que deux éléments ont fusionné, ils restent fusionnés : ainsi, la structure reconstruite est bien une hiérarchie qu'on parcourt selon les valeurs de λ ; *ii*) deux éléments proches fusionnent rapidement : le paramètre de tuning γ permet de contrôler la vitesse de fusion des éléments.

Nous montrons dans [Chiquet et al., 2017] que la méthode ainsi définie a une complexité en $\mathcal{O}(n\log(n))$ pour chacune des p dimensions, moindre que celles du clustering hiérarchique et de la MANOVA. Cependant, on obtient p arbres de classification qu'il faut maintenant agréger pour obtenir une classification consensus.

Agrégation. La version multidimensionnelle reconstruit l'arbre agrégé en le parcourant de haut en bas, partant des variables groupées dans toutes les dimensions, et en progressant jusqu'à ce que toutes les variables soient dans leur propre groupe. On désigne dans la suite par l'expression *règle de séparation*, le couple (pénalité, division) provenant des p arbres. La méthode d'agrégation est décrite par les étapes suivantes :

1. récupérer les règles de séparation sur toutes les dimensions ;
2. classer ces règles de séparation par ordre décroissant de pénalités ;
3. tant que les variables ne sont pas toutes dégroupées :
   — Si, dans la configuration actuelle de l'arbre, la règle de séparation considérée peut s'opérer, l'appliquer ;
   — Sinon, passer à la règle de séparation suivante ;
4. créer l'arbre agrégé à partir des règles de séparation retenues.

Cette méthode a une complexité en $\mathcal{O}(np\log(n))$. L'objectif d'une complexité inférieure à celle des méthodes classiques est satisfait.

3 Spectral fused-ANOVA

L'information apportée par différentes dimensions peut être brouillée, une simple combinaison d'arbres ne permettant pas de récupérer la structure des données. Nous proposons d'effectuer une étape de réduction de dimension à l'aide de méthodes spectrales avant d'appliquer la méthode de clustering Fused-ANOVA multidimensionnelle.

Clustering Spectral. L'idée du clustering spectral est d'utiliser le spectre d'une matrice de similarité comme nouvelles données pour effectuer le clustering [Ng et al., 2002].

On propose ici d'utiliser une matrice de Kernel gaussien K comme matrice de similarité entre les différents vecteurs d'observations :

$$K_{ij} = \exp(-\left\|X_i - X_j\right\|_2^2 / 2\sigma^2),$$

avec σ un paramètre de dispersion. Le Laplacien normalisé de la matrice K est ensuite calculé comme $L = D^{-1/2}KD^{-1/2}$ où D est la matrice diagonale telle que $D_{ii} = \sum_{j=1}^{n} K_{ij}$. On applique alors une méthode de clustering quelconque (ici, Fused ANOVA) aux vecteurs propres associés aux r plus grandes valeurs propres, ces vecteurs propres étant considérés comme les nouvelles données d'entrée.

Approximation de Nyström. Dans un contexte de grande dimension, calculer et stocker toute la matrice K, de taille $n \times n$, est extrêmement coûteux en temps et en mémoire, de même qu'effectuer une SVD du Laplacien (lui-même de taille $n \times n$).

On utilise l'approximation de Nyström pour les matrices de Gram [Williams and Seeger, 2001, Drineas and Mahoney, 2005] qui permet d'approcher la matrice K, notée $K(n, n)$ dans la suite. En notant $s \subset n$ un sous-échantillon de n, l'approximation se fait ainsi :

$$K(n, n) \approx K(n, s)K(s, s)^{\dagger}K(s, n),$$

où $K(n, s)$ désigne la matrice de kernel limitée aux s colonnes du sous-échantillon choisi, et $K(s, s)^{\dagger}$ la matrice pseudo-inverse de $K(s, s)$.

On calcule ensuite non pas la SVD du Laplacien mais seulement sur $K(s, s) = U\Sigma V^{\intercal}$. On obtient X^{new}, la matrice des nouvelles données, de dimension $n \times r$, avec r les dimensions retenues, associées aux plus grandes valeurs propres :

$$X^{\text{new}} \approx K(n, s)V_{(r)}\Sigma_{(r)}^{-1},$$

où $V_{(r)}$ $(s \times r)$ et $\Sigma_{(r)}$ $(r \times r)$ sont les matrices tronquées issues de la SVD. Le clustering s'effectue sur une version renormalisée de X^{new}.

La complexité de cette méthode est en $\mathcal{O}(s^3np),\ n \ll p$ du fait de la décomposition en valeurs propres. Le fait de prendre un sous-échantillon très petit par rapport à n permet d'avoir un temps de calcul moindre qu'une méthode à complexité quadratique.

4 Exemple d'application

Cette simulation a pour but d'illustrer l'impact de la taille du sous-échantillon de variables sur la classification obtenue. Le jeu de données comprend 100 observations de 50000 variables, simulées selon 15 groupes aux proportions différentes. Chaque groupe a été caractérisé par une loi normale de moyenne $\mu_k,\ k \in \{1, ..., 15\}$, et d'écart-type $\sigma = 0.4$.

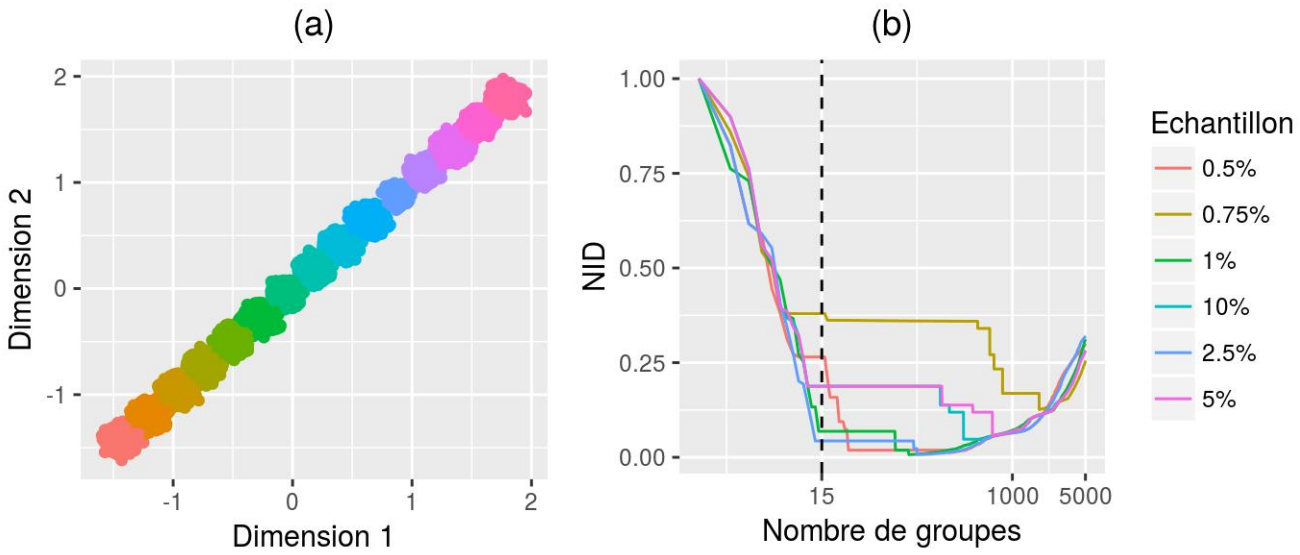

FIGURE 1 – (a) : Données simulées, tracées sur la première et deuxième dimensions. Chaque couleur correspond à un groupe. (b) : Courbes de NID (échelle log) obtenues pour différentes tailles de sous-échantillons.

Le paramètre σ permet de contrôler la séparabilité des classes. La Figure 1 (a) montre les caractéristiques des groupes.

Six tailles de sous-échantillon ont été testées : 0.5%, 0.75%, 1%, 2,5%, 5% et 10% du total des variables. Les sous-échantillons ont été choisis en sélectionnant à chaque fois les variables ayant la plus grande variance. Les résultats des simulations sont représentés sur la Figure 1 (b). Les performances des différentes simulations ont été comparées par la *Normalized Information Distance* [NID, Vinh et al., 2010], indicateur de distance entre deux classifications, basé sur le nombre d'éléments en communs dans les clusters et compris entre 0 et 1. Une distance normalisée à 0 signifie que les deux classifications sont identiques.

La performance obtenue sur le sous-échantillon de 1% des gènes de la table est la meilleure en termes de NID et de nombre de groupes obtenus (102 groupes, NID : 0.007). Le nombre de groupes obtenus semble important, mais seuls 16 clusters contiennent plus de 5 variables. Parmi ces 16 clusters, on retrouve 13 groupes correspondant parfaitement à ceux simulés, et 3 clusters correspondant aux deux derniers groupes. Ce résultat permet de montrer qu'un petit sous-échantillon de gènes permet d'obtenir une bonne classification.

5 Conclusions et discussion

Nous avons proposé une nouvelle méthode de classification, pour une application en grande dimension. L'exemple montre que la méthode obtient une bonne classification pour

un sous-ensemble très petit de la table de départ, ce qui permet d'envisager, à terme, d'appliquer la méthode sur des jeux de données comme ceux obtenus en métagénomique, où le nombre de variables est de l'ordre du million, pour quelques centaines d'individus. Quelques améliorations de la méthode sont encore à envisager.

La question du choix du sous-échantillon de variables à prendre pour appliquer la méthode reste ouverte. Si l'on sait, par les simulations, que l'on peut ne prendre que 1% des variables, nous ne savons pas lesquelles sélectionner. Il n'existe pas, à notre connaissance, de méthode de sélection de variables qui soit adaptée à la grande dimension.

Nous avons montré les résultats obtenus sur une simulation, où la classification était connue. Nous compléterons ces résultats par l'étude de la séparabilité des classes (via le paramètre σ) et de l'homogénéité de la taille des groupes sur la qualité du clustering. Enfin, en vue de l'application à des données réelles, la question du choix du niveau de coupure dans l'arbre, pour définir les clusters sera également explorée.

Références

Howard D Bondell and Brian J Reich. Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, 65(1) :169–177, 2009.

Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4) :994–1013, 2015.

Julien Chiquet, Pierre Gutierrez, and Guillem Rigaill. Fast tree inference with weighted fusion penalties. *Journal of Computational and Graphical Statistics*, 26(1) :205–216, 2017.

Petros Drineas and Michael W Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(Dec) : 2153–2175, 2005.

Toby Dylan Hocking, Armand Joulin, Francis Bach, and Jean-Philippe Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, page 1, 2011.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering : Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct) :2837–2854, 2010.

Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.