Classification de données de composition transformées et applications

Antoine Godichon-Baggioni [1] & Cathy Maugis-Rabusseau [2] & Andrea Rau [3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, 4 Place Jussieu, 75005 Paris, France. antoine.godichon_baggioni@upmc.fr*

[2] *Institut de Mathématiques de Toulouse, INSA Toulouse, 135 Avenue de Rangueil, 31400 Toulouse, France. cathy.maugis@insa-toulouse.fr*

[3] *GABI, INRA, AgroParisTech, Domaine de Vilvert, 78352 Jouy-en-Josas, France. andrea.rau@inra.fr*

Résumé. Bien qu'il y ait de nombreux algorithmes de classification dans la littérature, la question du choix de la meilleur stratégie à adopter lorsque l'on doit traiter des données de composition (i.e de données à valeurs dans le simplex) reste très largement sous explorée. Ce travail est motivé par l'analyse de deux applications basées sur la catégorisation de données de compositions; (1) identifier des groupes de gènes "co-exprimés" à partir de données RNA-seq; et (2) trouver des tendances dans l'utilisation de stations Velib' à Paris. Pour chacune de ces applications, on utilise des transformations appropriées des données, telles que la transformation Centered Log Ratio, et une nouvelle transformation appelée Log Centered Log Ration conjointement avec l'algorithme des K-means.

Mots-clés. Classification, Données de composition, K-means

Abstract. Although there is no shortage of clustering algorithms proposed in the literature, the question of the most relevant strategy for clustering compositional data (i.e., data belonging to the simplex) remains largely unexplored. This work is motivated by the analysis of two applications, both focused on the categorization of compositional profiles: (1) identifying groups of co-expressed genes from high-throughput RNA sequencing data; and (2) finding patterns in the usage of stations over the course of one week in the Velib' bicycle sharing system in Paris, France. For both of these applications, we make use of appropriately chosen data transformations, including the Centered Log Ratio and a novel extension called the Log Centered Log Ratio, in conjunction with the K-means algorithm.

Keywords. Clustering, Compositional data, K-means

1 Introduction

En génomique, on mesure l'expression des gènes à l'aide de technologies telle que les microarrays et le séquençage haut-débit. Dans le cas du séquençage haut-débit, on obtient des données dites RNA-seq, i.e des comptages pour la mesure d'expression des gènes. Les

recherches méthodologiques autour des données RNA-seq sont principalement tournées vers l'analyse différentielle (déterminer des gènes ayant une différence d'expression entre plusieurs conditions expérimentales). Dans le cadre de la co-expression, i.e déterminer des classes de gènes ayant la même évolution d'expression dans différentes conditions expérimentales, peu de méthodes sont actuellement proposées pour les données RNA-seq (Rau et Maugis-Rabusseau, 2017). Comme on s'intéresse à l'évolution des comptages et non à leurs valeurs (i.e on veut classer des gènes avec des comptages à des échelles très différentes), on se ramène à des proportions d'expression ("profils"), et on obtient donc des données à valeur dans le simplex

$$\mathcal{S}^d := \left\{ x = (x_1, ..., x_d) \in \mathbf{R}^d | \sum_{i=1}^{d} x_i = 1, x_i > 0, \forall i \right\}.$$

On va alors chercher à classer les gènes en fonction de leur profil, et on souhaite obtenir des classes précises pour les profils de gènes qui s'expriment très fortement dans une ou plusieurs conditions expérimentales (que l'on nommera profils "typés"), car cela implique que ces gènes ont probablement un rôle biologique commun.

Les données de compositions (i.e à valeurs dans le simplex) sont également étudiées dans de nombreux domaines tels que la géologie (Buccianti et al. 2006), l'économie (Longford et Pittau, 2006) et la génomique (Friedman et Alm, 2012). L'analyse statistique de ces données s'est beaucoup concentrée sur l'obtention de distances permettant de prendre en compte la spécificité des données tout en donnant une structure euclidienne au simplex. Ces distances sont souvent basées sur des transformations dont les plus usuelles sont Centered Log Ratio (CLR), Additive Log Ratio (ALR), ou Isometric Log Ratio (ILR) (Egozcue et al, 2003).

De nombreuses méthodes de classification existent dans la littérature, que l'on peut séparer en deux partie: les méthodes basées sur des modèles tel que les modèles de mélange (McLachlan et Peel, 2004) et les méthodes basées sur des distances telles que la classification hiérarchique (Ward, 1963), ou les K-means (MacQueen et al, 1967). Cependant, peu de méthodes sont adaptées aux données de composition. Dans un travail récent, (Rau et Maugis-Rabusseau, 2017) propose d'utiliser les modèles de mélanges gaussiens après transformation des données (logit,...), transformation qui permet de supprimer la dépendance linéaire des données. Cependant, ce type de méthode n'arrive pas à isoler des classes très précises pour les profils "typés".

Dans ce travail, notre objectif est double: (1) introduire des méthodes de classifications pour les données de composition basées sur des distances, et en particulier l'algorithme de K-means: et (2) utiliser et introduire de nouvelles transformations des données de composition permettant d'obtenir des classes très précises pour les profils "typés". Cette méthode est appliquée pour l'identification de groupes de gènes co-exprimés, et également pour trouver des tendances dans l'utilisation de stations Velib'. Toutes les méthodes présentées sont disponibles dans le package R `coseq`.

2 Méthodes et transformations

2.1 Algorithme de K-means

Soient $X_1, \ldots, X_n$ des points de $\mathbf{R}^d$ à regrouper en K classes. On considère la norme euclidienne usuelle $\|.\|_2$. Soit $\mathcal{C}^{(K)} = \{C_k, k = 1, \ldots, K\}$ une partition en K classes, et μ_k la moyenne de la classe C_k:

$$\mu_k := \frac{1}{|C_k|} \sum_{i \in C_k} X_i,$$

où $|C_k|$ est le nombre d'éléments de la classe k. L'objectif de l'algorithme de K-means est de minimiser la somme des erreurs au carré (SSE), définie pour chaque classification $\mathcal{C}^{(K)}$ par

$$\mathrm{SSE}\left(\mathcal{C}^{(K)}\right) := \sum_{k=1}^{K} \sum_{i \in C_k} \|X_i - \mu_k\|_2^2 ,$$

avec $i \in C_k$ if $\|X_i - \mu_k\|_2 = \min_{k'=1,\ldots,K} \|X_i - \mu_{k'}\|_2$. Dans ce qui suit, on considère l'algorithme de K-means sur les données pouvant être préalablement transformées.

2.2 Transformations pour les données de composition:

Soit $h : \mathcal{S}^d \longrightarrow \mathbf{R}^{d-1}$ un difféomorphisme. La fonction h permet de donner une structure euclidienne au simplex (Pawlowsky-Glahn et Egozcue, 2001). Les transformations les plus usuelles pour les données de composition sont le CLR, l'ALR et l'ILR (Aitchison, 1982). On se concentrera ici sur la transformation $\mathrm{clr} : \mathcal{S}^d \longrightarrow \mathbf{R}^d$, définie pour tout $x \in \mathcal{S}^d$ par

$$\mathrm{clr}(x) := \left(\ln\left(\frac{x_1}{g(x)}\right), \ldots, \ln\left(\frac{x_d}{g(x)}\right) \right), \tag{1}$$

où $g(x)$ est la moyenne géométrique de x. Dans ce cas, les données transformées n'appartiennent pas à $\mathbf{R}^{d-1}$ mais à l'hyperplan de $\mathbf{R}^d$ de vecteur normal $(1, \ldots, 1)$. On cherche alors à minimiser

$$\mathrm{SSE}_{\mathrm{clr}}\left(\mathcal{C}^{(K)}\right) := \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathrm{clr}\,(X_i) - \mu_{k,\mathrm{clr}}\|_2^2 ,$$

où $\mu_{k,\mathrm{clr}}$ est la moyenne des données (CLR)-transformées appartenant à la classe C_K:

$$\mu_{k,\mathrm{clr}} := \frac{1}{|C_k|} \sum_{i \in C_k} \mathrm{clr}\,(X_i) .$$

Log Centered Log Ratio transformation: Pour des données à valeurs dans des espaces de dimension modérée, quand de nombreuses coordonnées ont des proportions très

proches de 0, la transformation CLR a tendance à être assez sensible au fluctuations proche de ces zéros. Cela peut générer des effets particulièrement indésirables sur la classification. Pour prendre en compte ce phénomène et pour donner plus d'importance aux coordonnées avec des grandes valeurs, on propose une extension de la transformation CLR pour les données de composition appelée Log Centered Log Ratio (logCLR). Pour tout $x \in \mathcal{S}^d$, elle est définie par $\text{logCLR}(x) := (\text{logCLR}(x_1), \ldots, \text{logCLR}(x_d))$, avec pour tout j,

$$\text{logCLR}(x_j) := \begin{cases} -\left[\ln\left(1 - \ln\left[x_j/g(x)\right]\right)\right]^2 & \text{if } x_j/g(x) \leq 1, \\ \left(\ln\left[x_j/g(x)\right]\right)^2 & \text{otherwise,} \end{cases}$$

et $g(x)$ est la moyenne géométrique de x. D'ajouter un terme log quand $\frac{x_j}{g(x)} \leq 1$ permet de donner moins d'importance aux données avec des coordonnées proche de 0, tandis que le terme de carré permet de concentrer les profiles proche du centre du simplex $\left(\frac{1}{d}, \ldots, \frac{1}{d}\right)$ autour de 0. L'algorithme des K-means via cette transformation revient à minimiser

$$\text{SSE}_{\text{logCLR}}\left(\mathcal{C}^{(K)}\right) := \sum_{k=1}^{K} \sum_{i \in C_k} \left\| \text{logCLR}(X_i) - \mu_{k,\text{logCLR}} \right\|_2^2,$$

où $\mu_{k,\text{logCLR}}$ est la moyenne des données logCLR-transformées appartenant à la classe C_K:

$$\mu_{k,\text{logCLR}} := \frac{1}{|C_k|} \sum_{i \in C_k} \text{logCLR}(X_i).$$

3 Applications

On s'intéresse ici à la classification de données RNA-seq pour des souris. Plus précisément, on étudie le transcriptome d'embryons de souris dans la zone ventriculaire (VZ), la zone sous ventriculaire (SVZ) et la zone corticale (CP) (voir Fietz et al (2012) et Godichon-Baggioni et al (2019) pour plus de détéails). On dispose de $n = 8969$ gènes, et on va donc chercher à les classer en fonction de la façon dont ils se sont exprimés dans les différentes conditions. On dispose donc de $n = 8969$ profils à valeurs dans $\mathcal{S}^3$, et on utilise l'algorithme des K-means sur les profils et sur les données logCLR-transformées. On voit bien, Figure 1, que l'on obtient des classes beaucoup plus précises pour les profils typés lorsque l'on utilise la transformation logCLR (classes 9,6 et 12 par exemple).

Cette méthode a également été utilisée pour la classification de données "Velib" en fonction de leur fréquentation en semaine. Plus précisément, on s'est intéressé aux données disponible dans le Package R `funFEM` (Bouveyron et al, 2015) correspondant aux comptages de bornes disponibles à chaque heure dans $n = 1213$ stations durant une semaine. Afin d'étudier les habitudes des utilisateurs de chaque station (et non leur nombre), on a ramené les comptages à un tableau de proportion (i.e à des données de composition), puis appliqué les différentes méthodes proposées afin de dégager des groupes de stations très typées (voir par exemple la classe 3 dans la Figure 2, correspondant à des stations proches de zones touristiques).

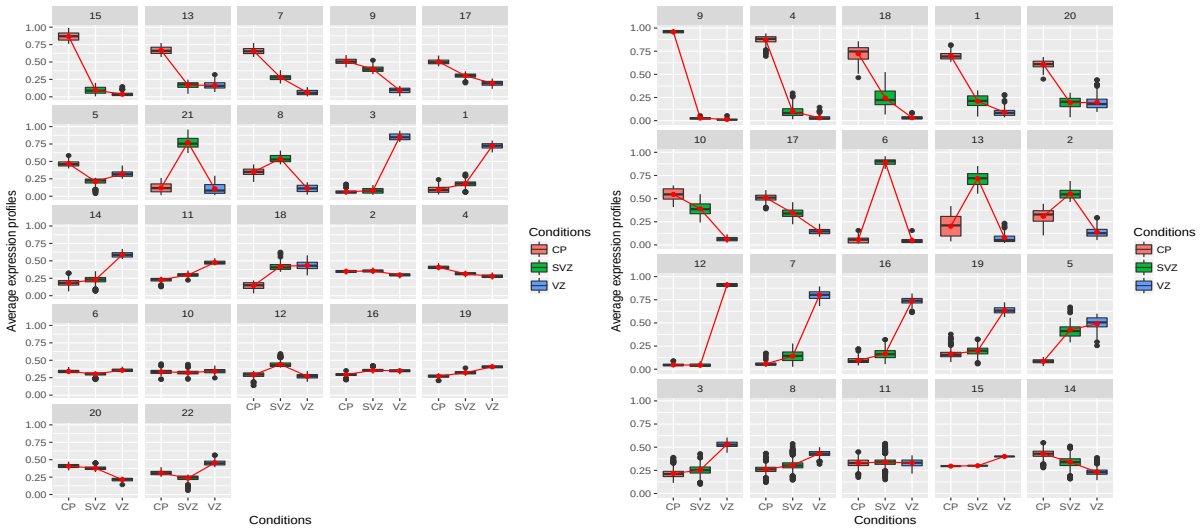

Figure 1: Classifications obtenues avec les K-means sur les profils (à gauche) et via la transformation logCLR (à droite) pour les souris.

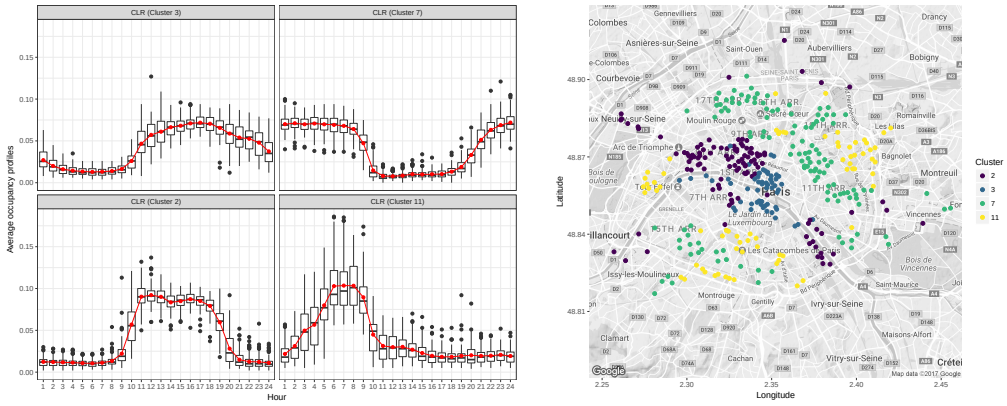

Figure 2: Exemples de classes obtenues pour les stations "velib" (à gauche) et leur localisation (à droite).

4 Conclusion

Dans ce travail, on a décrit une stratégie pour la classification de données de composition à l'aide de l'algorithme des K-means et de transformations. Le choix d'une transformation adaptée, en pratique, dépend beaucoup du type de classes de profiles auxquelles on s'intéresse; les recommandations sur la stratégie à adopter pour classer des données de composition peut donc être résumé comme suit: 1) si une classification équilibrée est désirée, l'algorithme des K-means doit être appliqué sur les données non transformées; 2) pour mettre l'accent sur des profils très typés, la transformation LogCLR doit être utilisée;

3) si on s'intéresse aux profils avec de faibles fluctuations autours des coordonnées proches de 0, la transformation CLR doit être privilégiée.

Bibliographie

Aitchison, J. (1982). *The statistical analysis of compositional data.* Journal of the Royal Statistical Society. Series B (Methodological), pages 139–177.

Bouveyron, C., Côme, E., and Jacques, J. (2015). *The discriminative functional mixture model for a comparative analysis of bike sharing systems.* The Annals of Applied Statistics, 9(4):1726–1760.

Buccianti, A., Tassi, F., and Vaselli, O. (2006). *Compositional changes in a fumarolic field, Vulcano Island, Italy: a statistical case study.* Geological Society, London, Special Publications, 264(1):67–77.

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). *Isometric logratio transformations for compositional data analysis.* Mathematical Geology, 35(3):279–300.

Fietz, Simone A and Lachmann, Robert and Brandl, Holger and Kircher, Martin and Samusik, Nikolay and Schröder, Roland and Lakshmanaperumal, Naharajan and Henry, Ian and Vogt, Johannes and Riehn, Axel and others (2012). *Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal.* Proceedings of the National Academy of Sciences.

Friedman, J. and Alm, E. J. (2012). *Inferring correlation networks from genomic survey data.* PLoS Computational Biology, 8(9):e1002687.

Godichon-Baggioni, A. Maugis-Rabusseau, C. and Rau, A. (2019). *Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data.* Journal of Applied Statistics.

Longford, N. T. and Pittau, M. G. (2006). *Stability of household income in European countries in the 1990s.* Computational statistics & data analysis, 51(2):1364–1383.

MacQueen, J. et al. (1967). *Some methods for classification and analysis of multivariate observations.* Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA.

McLachlan, G. and Peel, D. (2004). *Finite mixture models.* John Wiley & Sons.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). *Geometric approach to statistical analysis on the simplex.* Stochastic Environmental Research and Risk Assessment, 15(5):384–398.

Rau, A. and Maugis-Rabusseau, C. (2017). *Transformation and model choice for RNA-seq co-expression analysis.* Briefings in Bioinformatics, page bbw128.

Ward Jr, J. H. (1963). *Hierarchical grouping to optimize an objective function.* Journal of the American statistical association, 58(301):236–244.