Extension de la NMF supervisée pour l'intégration de données omiques

Aurélie MERCADIÉ[1,2], Eléonore GRAVIER[2], Gwendal JOSSE[2], Nathalie VIALANEIX[1] et Céline BROUARD[1]
[1] Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France
{aurelie.mercadie, nathalie.vialaneix, celine.brouard}@inrae.fr
[2] Pierre Fabre Dermo-Cosmétique, F-31300, Toulouse, France
{eleonore.gravier, gwendal.josse}@pierre-fabre.com

Résumé

Le développement des approches haut débit en biologie permet dorénavant la production massive de données dites « omiques », et ce pour des contextes applicatifs variés. Généralement acquis sur les mêmes échantillons, chacun de ces tableaux de données omiques illustre une partie seulement d'un système biologique complexe. L'intégration de ces données permet donc d'étudier ce système en globalité, et de mettre en lumière les relations existantes entre les divers acteurs moléculaires. Cette communication introduit une nouvelle méthode d'intégration découvrant des relations entre données « omiques » qui caractérisent des profils typiques de groupes distincts d'individus. Ici, la méthode de Factorisation Matricielle Non-négative (NMF) [1] est étendue à la problématique d'intégration de données dans un cadre supervisé. Pour cela, nous nous basons sur une variante du problème d'optimisation FR-lda, présenté dans l'article [2], les liens entre omiques étant assurés par l'appariement des individus dans la décomposition. En outre, les auteurs utilisent l'algorithme Majoration-Minimisation (MM) menant à des termes de mise à jour multiplicatifs pour résoudre leur problème. Afin d'assurer une parcimonie exacte dans la décomposition, nous proposons de combiner cette approche avec une mise à jour basée sur une approche proximale. L'utilisation de cette extension sur des données simulées a mis en lumière l'utilité de la partie supervisée de ce modèle ainsi que, dans certains cas, l'intérêt de la mise en place du proximal dans la résolution. Sur données faiblement bruitées, la solution fournie par l'approche proximale extrait des signatures moléculaires directement parcimonieuses, discriminant les individus selon leur groupe d'appartenance. En revanche, sur des données fortement bruitées, l'approche MM, qui ne fournit pas de parcimonie exacte, est plus robuste et classe correctement les variables les plus importantes expliquant les groupes.

Remerciements

La thèse d'Aurélie Mercadié est financée par Pierre Fabre Dermo-Cosmétique et l'ANRT dans le cadre du dispositif CIFRE.

Références

[1] D. Lee and S. H. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems (NIPS 2000)*, vol. 13, pp. 556–562, 2001.

[2] J. Leuschner, M. Schmidt, P. Fernsel, D. Lachmund, T. Boskamp, and P. Maass, "Supervised non-negative matrix factorization methods for maldi imaging applications," *Bioinformatics*, vol. 35, pp. 1940–1947, 2019.