PIntMF : Une méthode de factorisation matricielle pénalisée pour l'intégration de données multi-omiques

Morgane Pierre-Jean, Florence Mauger, Jean-François Deleuze, Edith Le Floch

PIntMF : Une méthode de factorisation matricielle pénalisée pour l'intégration de données multi-omiques

Morgane Pierre-Jean [1], Florence Mauger, [1] Jean-François Deleuze [1] et Edith Le Floch [1]

[1] *Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, 91057, Evry, France. morgane.pierre-jean@cng.fr*

Résumé. La génération de données multi-omiques est en pleine expansion avec l'amélioration des technologies à haut débit. L'intégration au sein d'une seule analyse de plusieurs sources d'information du génome pourrait permettre une meilleure compréhension des maladies ou des systèmes biologiques. Nous proposons ici une méthode non-supervisée de factorisation matricielle pénalisée multi-blocs pour intégrer des données multi-omiques. Cette méthode a pour but d'identifier de nouveaux groupes d'individus au sein d'une même maladie ainsi que d'identifier les variables pertinentes conduisant à cette classification. Nous avons appliqué cette méthode sur des données simulées pour comparer ses performances à des méthodes intégratives non-supervisées existantes et sur des données réelles. Cette nouvelle méthode permet de bien classer les individus et d'identifier les variables liées aux groupes avec plus de précision. Sur les données réelles, la méthode permet d'établir une nouvelle classification qui a un lien avec la survie des patients.

Mots-clés. Méthode non-supervisée, factorisation matricielle, multi-omique, classification

Abstract. The generation of multi-omics data is growing with the improvement of high-throughput technologies. The integration in the same analysis of several levels of the genome could allow a better understanding of diseases or biological systems. Here, we propose a non-supervised penalized matrix factorization method to integrate multi-omics data which aims to identify new groups of individuals within the same disease, as well as relevant markers leading to this classification. We applied this method to simulated data and compared its performances with existing integrative unsupervised methods. Our method leads to a correct clustering of individuals and identifies relevant biomarkers with more precision. The results on real data highlight a new clustering linked to the patient's survival.

Keywords. Unsupervised method, Matrix factorization, multi-omics, clustering

1 Contexte

Grâce au développement de technologies à haut débit, pour un même échantillon, la génération de plusieurs types de données "omiques" (génomique, transcriptomique, épigénomique, protéomique et métabolomique) se développe de plus en plus. En effet, le TCGA (The Cancer Genome Atlas, Weinstein et al. (2013)) a généré, pour un grand nombre de cancers et pour plusieurs échantillons, à la fois des données d'ADN, d'ARN, de méthylation, et même de protéomique. Depuis la dernière décennie, des méthodes d'intégration non-supervisées ont été développées pour analyser les données multi-omiques (Chauvel et al. (2019); Pierre-Jean et al. (2019); Cantini et al. (2020)). D'un point de vue mathématique, les données omiques peuvent être considérées comme des matrices et les variables pertinentes peuvent être extraites à l'aide de méthodes de factorisation matricielle. L'analyse en composantes principales (ACP, Hotelling (1933)) et la factorisation matricielle non négative (NMF (Non-Negative Matrix factorization), Lee and Seung (1999)) sont deux méthodes courantes de factorisation matricielle sous contraintes permettant de classer les échantillons et de mettre en évidence des variables pertinentes (Burstein et al. (2015)). L'ACP est une méthode puissante de réduction de dimensions et de visualisation des données tandis que la NMF qui n'impose pas l'orthogonalité des variables latentes met en évidence des groupes de variables associées aux groupes. Plus récemment, des extensions multi-blocs de NMF ont été développées pour permettre une analyse intégrative des données multi-omiques (Mo et al. (2013); Chalise et al. (2014); Chen and Zhang (2018)).

Ici, nous proposons un méthode intégrative de factorisation matricielle pénalisée (PIntMF : Penalized Integrative Matrix Factorization) dans le but d'intégrer des données multi-omiques. Nous avons comparé

cette méthode à plusieurs méthodes d'intégration existantes sur des simulations et sur des données réelles du TCGA. Un package `R` implémentant la méthode est également disponible sur github.

2 PIntMF : Penalized Integrative Matrix Factorization

2.1 Modèle

Dans cette section, nous décrivons le modèle PIntMF ainsi que sa résolution. On considère K matrices $\mathbf{X}^1, \dots \mathbf{X}^K$ en entrée du modèle. Chaque matrice $\mathbf{X}^k$ est de taille $n \times J_k$ (où n est le nombre d'individus et J_k le nombre de variables dans chaque bloc k. Nous proposons ici un modèle basé sur la factorisation matricielle de chaque bloc de données k i.e :

$$\mathbf{X}^k \approx \mathbf{W}\mathbf{H}^k \tag{1}$$

où $\mathbf{W}$ désigne une matrice de base commune et $\mathbf{H}^k$ une matrice spécifique à chaque bloc k. $\mathbf{W}$ est de taille $n \times P$ et $\mathbf{H}^k$ est de taille $P \times J_k$. Par conséquent, la variable P représente le nombre de variables latentes dans le modèle.

Nous imposons des constraintes de positivité sur les coefficients de la matrice $\mathbf{W}$ (comme dans un modèle NMF classique). Une contrainte de parcimonie a été ajoutée afin d'améliorer l'interprétation du modèle d'un point de vue de la classification et de diminuer sa complexité. La matrice $\mathbf{W}$ est utilisée pour faire une classification des individus en utilisant les K blocs de données omiques simultanément. La classification finale est une classification hierarchique ascendante classique avec la distance de Ward sur la matrice $\mathbf{W}$.

Sur $\mathbf{H}^k$, une contrainte de parcimonie est ajoutée également pour sélectionner des variables pertinentes.

Nous résumons les contraintes décrites précédemment au problème d'optimisation suivant :

$$\min_{\mathbf{W},\mathbf{H}^1,\dots,\mathbf{H}^k} \sum_{k=1}^{K} \|\mathbf{X}^k - \mathbf{W}\mathbf{H}^k\|_F^2 + \lambda_k \|\mathbf{H}^k\|_1 + \sum_{i=1}^{n} \mu_i \|\mathbf{w}_{i\bullet}\|_1 \quad \text{st.} \quad \mathbf{W} \geq 0 \tag{2}$$

where $\|\mathbf{H}^k\|_1 = \sum_{p=1}^{P} \sum_{j=1}^{J_k} |h_{pj}^k|$.

Le problème d'optimisation décrit ci-dessus (Equation 2) n'est pas jointement-convexe sur $\mathbf{W}, \mathbf{H}_1, \dots, \mathbf{H}_k$, mais est convexe séparément sur chacune des matrices. Par conséquent, le problème peut être résolu alternativement sur $\mathbf{W}$, et $\mathbf{H}^1, \dots, \mathbf{H}^k$ jusqu'à convergence de la fonction $g(\mathbf{X}^1, \dots, \mathbf{X}^K, \mathbf{W}, \mathbf{H}^1, \dots, \mathbf{H}^K) = \sum_{k=1}^{K} \|\mathbf{X}^k - \mathbf{W}\mathbf{H}^k\|_F^2$.

2.2 Résolution de W

Dans cette étape, les $\mathbf{H}^k$ sont fixés et l'équation 2 est résolue sur $\mathbf{W}$.

$$\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{X}^k - \mathbf{W}\mathbf{H}^k\|_F^2 + \sum_{i=1}^{n} \mu_i \|\mathbf{w}_{i\bullet}\|_1 \quad \text{st.} \quad \mathbf{W} \geq 0 \tag{3}$$

Tous les lignes de la matrices $\mathbf{W}$ sont indépendantes quand les $\mathbf{H}^k$ sont fixes. Le problème pour une ligne (individu) i peut être écrit de la façon suivante :

$$\min_{\mathbf{w}_{i\bullet}} \sum_{k=1}^{K} \|\mathbf{x}_{i\bullet}^k - \mathbf{w}_{i\bullet}\mathbf{H}^k\|^2 + \mu_i \|\mathbf{w}_{i\bullet}\|_1 \quad \text{st.} \quad \mathbf{w}_{i\bullet} \geq 0 \tag{4}$$

Le problème d'optimisation décrit par l'équation 4 est un problème Lasso classique avec une contrainte de positivité. Nous avons fixé μ_i à 1 mais il serait intéressant de calibrer la contrainte par validation croisée. Il peut être facilement et rapidement résolu en utilisant le package `glmnet R` développé par Friedman et al. (2010).

2.3 Résolution de $\mathbf{H}^k$

Quand $\mathbf{W}$ est fixé, les $\mathbf{H}^k$ peuvent être résolus indépendamment les uns des autres. Pour faciliter la lecture de cette section, l'indice k a été retiré des équations.

$$\min_{\mathbf{H}} Q(\mathbf{H}) = \min_{\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2 + \lambda \sum_{p=1}^{P} \sum_{j=1}^{J} |h_{pj}| \tag{5}$$

$$\begin{aligned} Q(\mathbf{H}) = &\quad \text{Trace}\left\{(\mathbf{X} - \mathbf{WH})(\mathbf{X} - \mathbf{WH})^T\right\} + \lambda \textstyle\sum_{p=1}^{P} \sum_{j=1}^{J} |h_{pj}| \\ = &\quad vec(\mathbf{X} - \mathbf{WH})^T vec(\mathbf{X} - \mathbf{WH}) + \lambda \textstyle\sum_{p=1}^{P} \sum_{j=1}^{J} |h_{pj}| \end{aligned}$$

$$\text{On note } \mathbf{h} = vec(\mathbf{H}) = \begin{pmatrix} \mathbf{H}_{11} \\ \vdots \\ \mathbf{H}_{P1} \\ \vdots \\ \mathbf{H}_{1J} \\ \vdots \\ \mathbf{H}_{PJ} \end{pmatrix} \text{ et } \mathbf{x} = vec(\mathbf{X}) = \begin{pmatrix} \mathbf{X}_{11} \\ \vdots \\ \mathbf{X}_{n1} \\ \vdots \\ \mathbf{X}_{1J} \\ \vdots \\ \mathbf{X}_{nJ} \end{pmatrix}.$$

$$\begin{aligned} Q(\mathbf{H}) = &\quad (\mathbf{x} - vec(\mathbf{WH}))^T (\mathbf{x} - vec(\mathbf{WH})) + \lambda \|\mathbf{h}\|_1 \\ = &\quad (\mathbf{x} - (\mathbb{I}_n \otimes \mathbf{W}) vec(\mathbf{H}))^T (\mathbf{x} - (\mathbb{I}_n \otimes \mathbf{W}) vec(\mathbf{H})) + \lambda \|\mathbf{h}\|_1 \\ Q(\mathbf{h}) = &\quad (\mathbf{x} - \tilde{\mathbf{W}}\mathbf{h})^T (\mathbf{x} - \tilde{\mathbf{W}}\mathbf{h}) + \lambda \|\mathbf{h}\|_1 \end{aligned}$$

où $\mathbf{h} = vec(\mathbf{H})$, $\mathbf{x} = vec(\mathbf{X})$, $\mathbb{I}_n$ est la matrice identité de taille n et $\tilde{\mathbf{W}} = \mathbb{I}_n \otimes \mathbf{W}$

On peut reformuler le problème de la façon suivante : $Q(\mathbf{h}) = \|\mathbf{x} - \tilde{\mathbf{W}}\mathbf{h}\|^2 + \lambda \|\mathbf{h}\|_1$

Minimiser $Q(\mathbf{h})$ revient donc à résoudre un problème Lasso classique. La valeur de λ sera optimisée par validation croisée pour chaque bloc $k = 1, \ldots, K$.

Comme pour $\mathbf{W}$ on utilise le package `glmnet` pour résoudre ce problème.

2.4 Normalisation de W

Afin d'interpréter la matrice $\mathbf{W}$ comme une matrice de poids, nous avons ajouté une contrainte sur la somme des coefficients de chaque ligne de $\mathbf{W}$. Pour éviter des problèmes de convergence ou de non-identifiabilité, la contrainte sur la somme est ajoutée après avoir estimé la matrice $\mathbf{W}$. On divise chaque ligne par sa somme après chaque étape :

$$\mathbf{w}_{i\bullet} = \frac{\mathbf{w}_{i\bullet}}{\sum_{p=1}^{P} \mathbf{w}_{ip}} \tag{6}$$

2.5 Initialisation de l'algorithme

La méthode de Wang et al. (2014) (Similarity Network Fusion : SNF) a déjà montré de bonnes performances pour la classification en multi-omiques. Nous avons choisi d'initialiser $\mathbf{W}$ avec la classfication avec P groupes de cet algorithme. $\mathcal{C}_p$ désigne le groupe p de SNF. Ainsi,

$$w_{ip} = \begin{cases} 1 & \text{si } i \in \mathcal{C}_p \\ 0 & \text{Sinon} \end{cases} \tag{7}$$

Cette initialisation a l'avantage de prendre en compte simultanément les K blocs dans la classification.

3 Résultats

3.1 Simulations

Nous avons évalué les performances du modèle présenté dans la section précédente sur le même schéma de simulation que dans l'article de Pierre-Jean et al. (2019). Ce schéma simule trois blocs de données hétérogènes (sous des distributions différentes : Binaire, Beta et Gaussienne). Quatre groupes déséquilibrés respectivement composés de 25, 20, 5 et 10 individus ont été simulés dans ces trois blocs. PIntMF a été comparé à 6 méthodes existantes qui donnaient les meilleurs résulats dans Pierre-Jean et al. (2019) à savoir : SNF de Wang et al. (2014), intNMF de Chalise et al. (2014), SGCCA de Tenenhaus et al. (2014), MoCluster de Meng et al. (2015), iClusterPlus de Mo et al. (2013), et CIMLR de Ramazzotti et al. (2018). Les performances des méthodes ont été évaluée à la fois sur la capacité des méthodes à retrouver la classification simulée mais également sur la capacité à retrouver les variables simulées associées à la classification.

Évaluation des performances sur le clustering

Les performances de la classification ont été évaluées en utilisant le Rand Index Ajusté (ou Adjusted Rand Index (ARI), Hubert and Arabie (1985)) sur 4 Benchmarks de simulation. L'ARI est égal à 1 si la partition trouvée est la même que celle simulée et 0 si elle est "complètement" différente.

Sur les 4 benchmarks simulés, PIntMF, MoCluster et SNF ont des meilleures performances que les autres méthodes avec un ARI égal à 1 dans la plupart des cas (Fig. 1a). On peut noter que PIntMF ne fait aucune erreur de classification sur les Benchmarks 1, 2 et 4.

Sélection de variables

La performance sur la sélection de variables a été évaluée en utilisant les taux de faux positifs (TFP) et de vrais positifs (TVP). Nous avons résumé les TFP et TVP en traçant des courbes ROC puis en calculant l'AUC (Area under the curve) associé.

Nous ne pouvons pas évaluer les performances SNF car la méthode est basée sur des matrices de similarité qui ne donnent pas de poids aux variables dans les blocs.

Le calcul de l'AUC (Figure 1b) montre que PIntMF est une méthode qui obtient des performances très satisfaisantes et similaires à MoCluster sur les trois types de blocs. SGCCA et CIMLR ont des performances légèrement inférieures en moyenne sur les 3 types de distribution. intNMF ne semble pas adapté aux données simulées sous la distribution beta et iClusterPlus ne parvient pas à retrouver les variables pertinentes sous la distribution Binaire.

3.2 Analyses sur données réelles

Nous avons également utilisé PIntMF sur des données réelles de glioblastome issues de la base de données du TCGA. Ce jeu de données contient 55 individus, avec des données d'expression (1740 gènes), de nombre de copies d'ADN (1599 gènes) et de méthylation (1515 gènes). En accord avec les différentes catégories des tumeurs, nous avons sélectionné 5 variables latentes dans le modèle PIntMF. La heatmap (Figure 1c) représente pour chacun des échantillons (en colonne) la valeur pour chaque variable latente. Nous avons superposé la classification des types de glioblastome et la classification trouvée par PIntMF. La classification est légèrement différente mais semble cohérente avec celle existante dans les données cliniques. Pour les individus non-classés (NA), on pourrait les assigner, grâce à PIntMF, à des groupes déjà existants. Une analyse de survie avec les groupes trouvés par PIntMF a également été réalisée (Figure 1d) et montre que les groupes ont des taux de survies différents (p-value significative à 5%). En particulier, le groupe bleu a un très mauvais pronostic alors que la survie du groupe orange est beaucoup plus favorable.

4 Conclusions et perspectives

Pour conclure, PIntMF donne de bonnes performances sur les simulations que ce soit pour la classification ou la sélection de variables. Sur les données réelles, il semble que les groupes permettent de catégoriser des patients qui n'ont pas été assignés à une classe de tumeur particulière. La méthode révèle aussi des taux survie différents par groupe.

Contrairement aux autres méthodes qui utilise des pénalités (MoCluster, SGCCA, iClusterPlus), un avantage de la résolution de PIntMF est que les valeurs des pénalités sur les matrices $\mathbf{H}^k$ sont automatiquement ajustées. Ainsi, les utilisateurs ont seulement besoin de définir le nombre de variables latentes.

Nous avons également implémenté un package R (nommé PIntMF) pour reproduire les résultats de cet article. Ce package est disponible sur github.

Enfin, il faudrait encore améliorer l'automatisation du choix du nombre de variables latentes. En effet, ceci rendrait la méthode encore plus facile d'utilisation dans le cadre d'applications en génomique. On pourrait également optimiser la contrainte de parcimonie sur la matrice $\mathbf{W}$.

Références

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10) :1113–1120, 2013.

Cécile Chauvel, Alexei Novoloaca, Pierre Veyre, Frédéric Reynier, and Jérémie Becker. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics*, 2019.

Morgane Pierre-Jean, Jean-Francois Deleuze, Edith Le Floch, and Florence Mauger. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in bioinformatics*, 2019.

Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anais Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for cancer study. *bioRxiv*, 2020.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6) :417, 1933.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788, 1999.

Matthew D Burstein, Anna Tsimelzon, Graham M Poage, Kyle R Covington, Alejandro Contreras, Suzanne AW Fuqua, Michelle I Savage, C Kent Osborne, Susan G Hilsenbeck, Jenny C Chang, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research*, 21(7) : 1688–1698, 2015.

Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11) :4245–4250, 2013.

Prabhakar Chalise, Devin C Koestler, Milan Bimali, Qing Yu, and Brooke L Fridley. Integrative clustering methods for high-dimensional molecular data. *Translational cancer research*, 3(3) :202, 2014.

Jinyu Chen and Shihua Zhang. Discovery of two-level modular organization from matched genomic data via joint matrix tri-factorization. *Nucleic acids research*, 46(12) :5967–5976, 2018.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) :1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

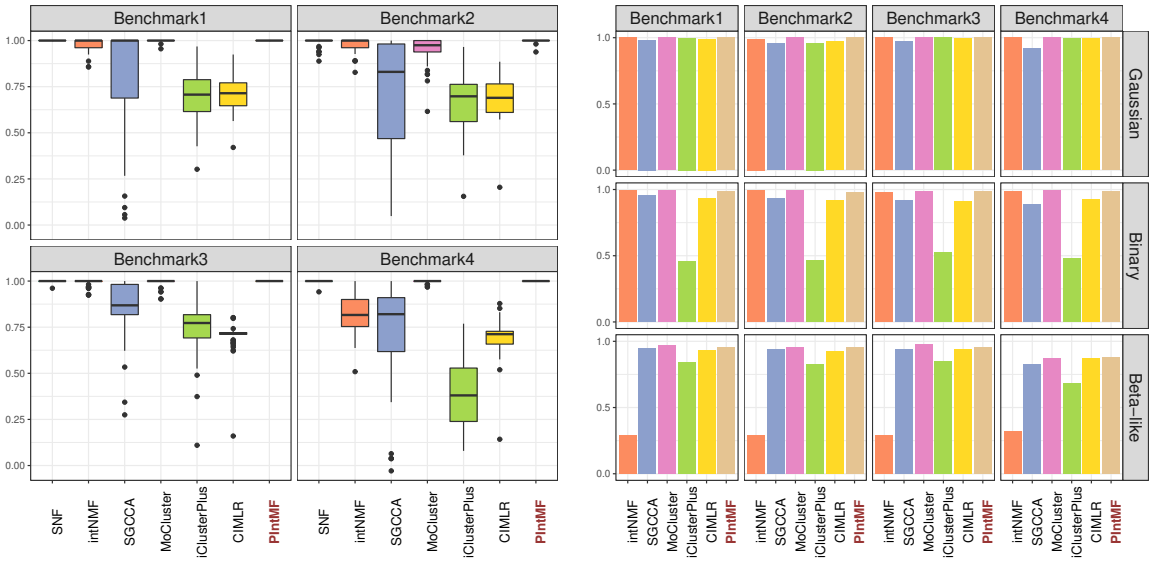
Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11 (3) :333, 2014.

Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3) :569–583, 2014.

Chen Meng, Dominic Helm, Martin Frejno, and Bernhard Kuster. mocluster : Identifying joint patterns across multiple omics data sets. *Journal of proteome research*, 15(3) :755–765, 2015.

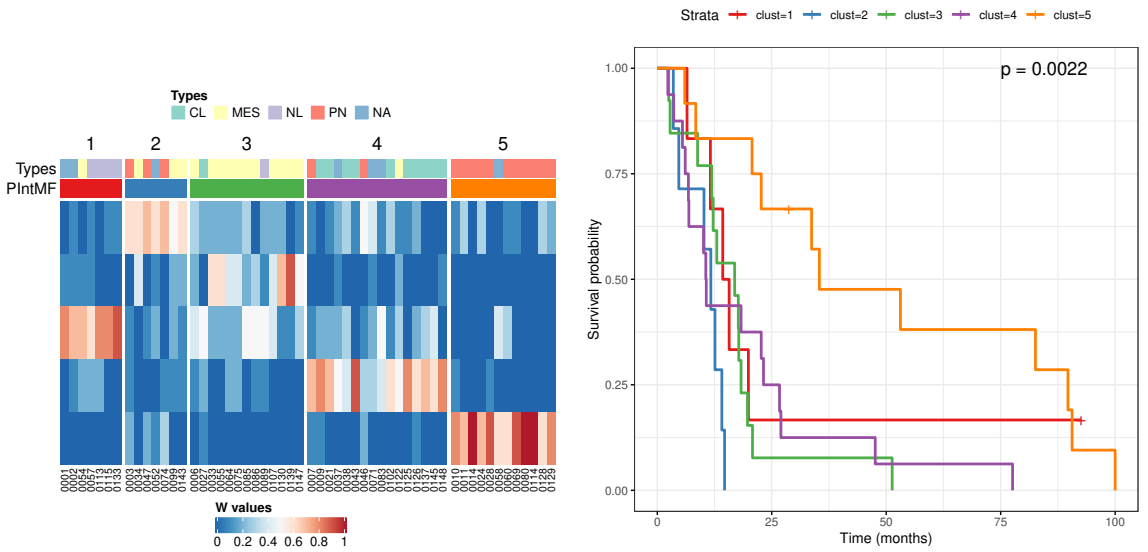
Daniele Ramazzotti, Avantika Lal, Bo Wang, Serafim Batzoglou, and Arend Sidow. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nature communications*, 9(1) :4453, 2018.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1) :193–218, 1985.



(a) ARI (Adjusted Rand Index)

(b) AUC (Area Under ROC Curve)



(c) Matrice $\mathbf{W}$ de PIntMF

(d) Courbes de survie

FIGURE 1 – (a) Rand Index Ajusté (ARI). (b) Aire sous la courbe ROC pour SNF, intNMF, SGCCA, MoCluster, iClusterPlus, CIMLR et PIntMF sur 4 Benchmarks de simulations. (c) Matrice $\mathbf{W}$ inferrée par PIntMF sur des données de glioblastome. (d) Courbes de survie pour chacun des groupes inferrés par PIntMF.