Factorisation couplée de tenseurs pour l'analyse de données de cytométrie en flux

Philippe FLORES, Guillaume HARLÉ, Konstantin USEVICH, Stéphanie GRANDEMANGE, David BRIE

Université de Lorraine, Centre de Recherche en Automatique de Nancy, CNRS Campus Sciences BP 70239, 54506 Vandoeuvre-lès-Nancy, France

prenom.nom@univ-lorraine.fr

Résumé – Nous proposons une méthode d'analyse automatique de données de cytométrie en flux. Le problème est abordé comme un problème d'estimation de densité en grandes dimensions. En modélisant la distribution comme un mélange de densités séparables, l'estimation de la densité jointe est reformulée comme une factorisation tensorielle couplée de marginales 3D. Afin de réduire la charge de calcul, nous proposons différentes stratégies de couplage partiel. Les termes de rang 1 sont alors regroupés par clustering hiérarchique. L'ensemble du traitement est intégré au sein d'un nouvel outil de visualisation des résultats. La pertinence de l'approche est illustrée sur des données simulées ainsi que sur des données réelles correspondant à un mélange contrôlé de populations de cellules.

Abstract – We propose a method for automated flow cytometry data analysis. The problem is addressed as a high dimension density estimation problem. Modeling the distribution as a mixture of separable distributions, the estimation of the joint density can be reformulated as a coupled tensor factorization of 3D-marginals. To reduce the computational load, partial coupling strategies are proposed. The rank-1 components are grouped by a hierarchical clustering. The whole data processing pipeline is integrated into a new visualization tool. The usefulness of the proposed methodology is illustrated on simulated and real data corresponding to a controlled mixture of cell populations.

1 Introduction

La cytométrie en flux (CMF) est une technique d'analyse de cellules biologiques largement utilisée dans de nombreux domaines comme l'agriculture, la médecine ou la biologie [1]. C'est la technique de référence en immunologie car elle permet d'identifier des populations cellulaires rares et contribue ainsi à l'amélioration de la connaissance du système immunitaire [2]. D'un point de vue analyse de données, un cytomètre produit un nuage de points dans un espace à M dimensions. Le but est d'identifier dans ce nuage de points les différentes populations de cellules. L'analyse conventionnelle réalisée manuellement par les biologistes/médecins, consiste essentiellement à enchaîner une suite d'analyse en 2 dimensions ; elle devient complexe, subjective et coûteuse en temps/homme lorsque M augmente. Cela a motivé le développement de méthodes automatiques [3, 4, 5] qui restent cependant coûteuses et peuvent difficilement être appliquées à de grands ensembles de données. En outre, ces méthodes présentent des performances limitées pour l'analyse de populations de cellules rares et les outils de visualisations associés sont difficilement interprétables par les utilisateurs finaux. Nous proposons une approche probabiliste qui repose sur l'estimation de la densité jointe des données, en s'inspirant de [6]. Pour faire face à la malédiction de la dimension, nous adoptons un modèle bayésien naïf de la densité jointe : ainsi, estimer l'histogramme en M dimensions revient à estimer les facteurs d'un modèle tensoriel CP [7] dont la complexité demeure linéaire avec le nombre de dimensions. En s'inspirant de [8], l'estimation du modèle CP d'ordre M est formulé comme un problème de factorisation couplée des marginales 3D. Afin de réduire la complexité de l'algorithme, différentes stratégies de couplage partiel sont proposées et évaluées. Les différentes populations cellulaires sont obtenues en appliquant un clustering hiérarchique aux termes de rang 1.

2 Modèle bayésien naïf

Soit $\boldsymbol{x} = \left(X^{(1)}, \ldots, X^{(M)}\right)$ un vecteur aléatoire prenant des valeurs dans $\mathcal{I}^{(1)} \times \ldots \times \mathcal{I}^{(M)}$, où $\mathcal{I}^{(m)} = \left[x_{\min}^{(m)}, x_{\max}^{(m)}\right]$. Notre but est d'estimer la densité de probabilité $\mathrm{p}(X^{(1)}, \ldots, X^{(M)})$ (PDF) du vecteur aléatoire $\boldsymbol{x}$ à partir de la matrice d'observation $\boldsymbol{X}$ qui regroupe les N réalisations de $\boldsymbol{x}$ notées $\boldsymbol{x}_{n,:}$, $n = 1, \cdots, N$. Une approche possible d'estimation de densités est de considérer un histogramme en M dimensions. Dans ce cas, chaque intervalle $\mathcal{I}^{(m)}$ est séparé en I intervalles égaux allant de $\Delta_1^{(m)} = \left[x_0^{(m)}, x_1^{(m)}\right]$ à $\Delta_I^{(m)} = \left[x_{I-1}^{(m)}, x_I^{(m)}\right]$, où $x_0^{(m)} = x_{\min}^{(m)}$ et $x_I^{(m)} = x_{\max}^{(m)}$. Cet histogramme, noté $\boldsymbol{\mathcal{H}}$, peut être interprété comme une PDF jointe discrétisée :

$$\begin{aligned}\mathcal{H}_{i_1 \ldots i_M} &= \Pr\left(X^{(1)} \in \Delta_{i_1}^{(1)}, \ldots, X^{(M)} \in \Delta_{i_M}^{(M)}\right) \\ &= \int_{X^{(1)} \in \Delta_{i_1}^{(1)}} \cdots \int_{X^{(M)} \in \Delta_{i_M}^{(M)}} \mathrm{p}(X^{(1)}, \ldots, X^{(M)}) dX^{(1)} \ldots dX^{(M)}.\end{aligned} \tag{1}$$

Pour estimer l'histogramme à partir de $\boldsymbol{X}$, les échantillons sont décomptés dans chaque intervalle en M dimensions :

$$\widetilde{\mathcal{H}}_{i_1 \ldots i_M} = \frac{1}{N} \operatorname{Card}\left\{ n \in [\![1, N]\!] \,\middle|\, \boldsymbol{x}_n \in \Delta_{i_1}^{(1)} \times \ldots \times \Delta_{i_M}^{(M)} \right\} \tag{2}$$

Cependant, cette approche nécessite un nombre d'échantillons qui croît de manière exponentielle avec le nombre de dimensions. Pour donner un ordre de grandeur avec $M = 8$ et $I = 20$, l'histogramme est décrit par $I^M \approx 10^{10}$ valeurs et requiert donc beaucoup plus d'échantillons pour obtenir une estimation précise : c'est la malédiction de la dimension. Pour pallier ce problème, nous suivons l'approche de [8] qui utilise un modèle bayésien Naïf (MBN) dont la complexité demeure linéaire avec le nombre de dimensions. Ce modèle permet de représenter de

manière efficace la loi de probabilité de $\boldsymbol{X}$ [9, 8]. Le MBN introduit une variable latente discrète L telle que les éléments de $\boldsymbol{x}$ sont conditionnellement indépendants par rapport à L :

$$\mathrm{p}(X^{(1)},\ldots,X^{(M)}) = \sum_{r=1}^{R} \Pr(L{=}r) \prod_{m=1}^{M} \mathrm{p}\Big(X^{(m)}|L{=}r\Big) \quad (3)$$

Ainsi, le MBN conduit à une Décomposition Canonique Polyadique (CPD) d'ordre M [10] de l'histogramme $\boldsymbol{\mathcal{H}}$ [8] :

$$\mathcal{H}_{i_1\ldots i_M} = \sum_{r=1}^{R} \Pr(L{=}r) \prod_{m=1}^{M} \Pr\left(X^{(m)} \in \Delta_{i_m}^{(m)} \Big| L{=}r\right)$$

$$\boldsymbol{\mathcal{H}} = \sum_{r=1}^{R} \boldsymbol{\lambda}_r \boldsymbol{a}_r^{(1)} \circ \ldots \circ \boldsymbol{a}_r^{(M)} = \Big[\!\Big[\boldsymbol{\lambda};\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(M)}\Big]\!\Big] \quad (4)$$

où R représente le nombre de composantes (rang de la décomposition de $\boldsymbol{\mathcal{H}}$). Afin que les matrices $\boldsymbol{A}^{(m)} = \left(\boldsymbol{a}_1^{(m)} \cdots \boldsymbol{a}_R^{(m)}\right) \in \mathbb{R}^{I_m \times R}$ et le vecteur $\boldsymbol{\lambda} \in \mathbb{R}^R$ soient interprétables comme des probabilités, ces quantités doivent satisfaire les conditions de non-négativité : $\boldsymbol{\lambda} \geq 0$, $\boldsymbol{A}^{(m)} \geq 0$, et les contraintes de simplexes (*sum-to-one*) : $\mathbb{1}^\mathsf{T}\boldsymbol{\lambda} = 1$, $\mathbb{1}^\mathsf{T}\boldsymbol{A}^{(m)} = \mathbb{1}^\mathsf{T}$.

3 Factorisation tensorielle couplée

3.1 Couplage total de tenseurs

Comme précisé en section 2, la malédiction de la dimension rend l'estimation de l'histogramme en dimension M impossible. Pour pallier ce problème, nous procédons à une factorisation tensorielle couplée des histogrammes 3D, estimables avec un nombre d'échantillons raisonnable, et permettant de conserver des propriétés d'unicités intéressantes (que n'ont pas les matrices). Soit $\left(X^{(j)}, X^{(k)}, X^{(\ell)}\right)$ un triplet de variables aléatoires de $\boldsymbol{x}$, le MBN (4) peut être marginalisé pour obtenir un modèle d'ordre 3 qui approche l'histogramme 3D $\boldsymbol{\mathcal{H}}^{(jk\ell)}$.

$$\boldsymbol{\mathcal{H}}^{(jk\ell)} = \Big[\!\Big[\boldsymbol{\lambda};\boldsymbol{A}^{(j)},\boldsymbol{A}^{(k)},\boldsymbol{A}^{(\ell)}\Big]\!\Big] \quad (5)$$

Pour estimer les facteurs du modèle (4) , nous considérons l'ensemble de tous les triplets $\mathcal{T} = \Big\{(j,k,\ell) \in [\![1,M]\!]^3 \,\Big|\, j<k<\ell\Big\}$ et nous résolvons le problème d'optimisation :

$$\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{A}^{(1)}}, \ldots, \widehat{\boldsymbol{A}^{(M)}} \quad (6)$$

$$= \underset{\boldsymbol{\lambda},\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(M)}}{\operatorname{argmin}} \sum_{(j,k,\ell)\in\mathcal{T}} \left\| \widetilde{\boldsymbol{\mathcal{H}}}^{(jk\ell)} - \Big[\!\Big[\boldsymbol{\lambda};\boldsymbol{A}^{(j)},\boldsymbol{A}^{(k)},\boldsymbol{A}^{(\ell)}\Big]\!\Big] \right\|_F^2$$

$$\text{s.c.} \quad \boldsymbol{\lambda} \geq 0, \boldsymbol{A}^{(m)} \geq 0, \mathbb{1}^\mathsf{T}\boldsymbol{\lambda} = 1, \mathbb{1}^\mathsf{T}\boldsymbol{A}^{(m)} = \mathbb{1}^\mathsf{T},$$

appelé factorisation tensorielle couplée totalement. (6) est résolu avec une procédure d'AO-ADMM couplée [8].

3.2 Conditions d'identifiabilité

Les décompositions tensorielles possèdent des conditions d'identifiabilité intéressantes [10]. En particulier, si tous les $\boldsymbol{\mathcal{H}}^{(jk\ell)}$ sont individuellement génériquement identifiables, c'est-à-dire si $R < \frac{3I-2}{2}$, alors le tenseur de probabilité $\boldsymbol{\mathcal{H}}$ est aussi identifiable. Cependant, comme beaucoup de $\boldsymbol{\mathcal{H}}^{(jk\ell)}$ partagent des facteurs communs, les conditions d'identifiabilité peuvent être significativement relaxées. Par exemple, si $M \leq I$ alors $\boldsymbol{\mathcal{H}}$ est génériquement identifiable si $R \leq I(M-2)$ [8]. Notons que ces résultats d'identifiabilité correspondent au cas de décompositions exactes et sont formulées pour des matrices facteurs à valeurs réelles (éventuellement non-négatives). En pratique, comme le nombre d'échantillons est limité, seules des versions bruitées de $\boldsymbol{\mathcal{H}}^{(jk\ell)}$ sont disponibles ce qui mène à un problème d'approximation de rang faible de tenseurs. De ce point de vue, les contraintes de non-négativité sur les facteurs sont intéressantes car elles garantissent l'existence et l'unicité de l'approximation de rang faible, voir [11]. Enfin, une analyse attentive des résultats d'identifiabilité de [8] révèle que seul l'identifiabilité d'un tenseur étendu correspondant à une partition spécifique des variables est requise. En d'autres termes, seul un nombre limité de triplets (définis par la partition des M variables) est nécessaire pour assurer l'identifiabilité. Cette idée est développée dans la sous-section suivante pour réduire les coûts de calculs de la factorisation couplée de tenseurs.

TABLE 1 – Stratégies de couplage partiel pour M=10.

Strategie	Card $(\mathcal{T})$	Triplets
+2	5	$(1,2,3),(3,4,5),(5,6,7)$ $(7,8,9),(9,10,1)$
+1	10	$(1,2,3),(2,3,4),\ldots$ $(9,10,1),(10,1,2)$
1/8	15=120/8	triplets aléatoires
1/4	30=120/4	triplets aléatoires
1/2	60=120/2	triplets aléatoires
1	120=$\binom{10}{3}$	tous les triplets

3.3 Couplage partiel de tenseurs

Dans (6), tous les triplets possibles sont utilisés, soit $\binom{M}{3}$ triplets. Le principe de couplage partiel est de ne considérer qu'un sous-ensemble des histogrammes 3D, ce qui permet de réduire le coûts de calculs de la procédure d'optimisation. Plusieurs stratégies sont possibles mais toutes doivent contenir au moins une occurrence de chaque variable, et toutes les variables doivent pouvoir être reliées entre elles par au moins une suite de triplets. Six stratégies de complexité croissante sont présentées dans la Table 1.

3.4 Évaluation des performances

Nous avons appliqué la factorisation couplée de tenseurs avec les stratégies de la Table 1 sur des données synthétiques de dimension $M = 10$. $R = 20$ distributions gaussiennes de paramètres aléatoires ont été générées et ajoutées ensemble avec des poids $\boldsymbol{\lambda}$ pour créer un histogramme $\boldsymbol{\mathcal{H}}$ théorique. Nous avons ensuite généré un nombre d'échantillons $N = \{10^4,$

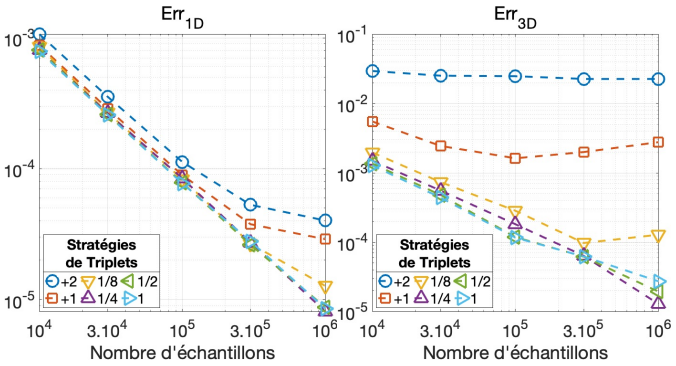

FIGURE 1 – Évolution des erreurs pour les différentes stratégies

$3.10^4, 10^5, 3.10^5, 10^6\}$ suivant la même distribution et calculé les histogrammes 3D pour $I = 30$ bins. Concernant les paramètres de l'AO-ADMM couplée, nous avons choisi $N_1 = 10^3$ itérations externes et $N_2 = 20$ itérations internes. Le rang de la décomposition a été choisi égal au rang théorique $R = 20$. Pour évaluer l'influence des différentes stratégies de couplage sur les performances d'estimation de notre méthode, nous avons calculé l'erreur sur les marginales 1D et 3D, moyennée sur 10 réalisations différentes.

$$\mathrm{Err}_{1D} = \sum_{m=1}^{M} \left\| \sum_{r=1}^{R} \boldsymbol{\lambda}_r \boldsymbol{a}_r^{(m)} - \sum_{r=1}^{R} \widehat{\boldsymbol{\lambda}_r} \widehat{\boldsymbol{a}_r^{(m)}} \right\|^2 \quad (7)$$

$$\mathrm{Err}_{3D} = \sum_{(j,k,\ell)\in\mathcal{T}} \left\| \left[\!\left[\boldsymbol{\lambda}; \boldsymbol{A}^{(j)}, \boldsymbol{A}^{(k)}, \boldsymbol{A}^{(\ell)} \right]\!\right] - \left[\!\left[\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{A}^{(j)}}, \widehat{\boldsymbol{A}^{(k)}}, \widehat{\boldsymbol{A}^{(\ell)}} \right]\!\right] \right\|_F^2 \quad (8)$$

La Figure 1 montre les erreurs sur les marginales 1D et 3D : les couplages '1/8', '1/4' et '1/2' permettent d'atteindre des performances similaires à la stratégie de couplage total. En revanche, les stratégies '+1' et '+2' conduisent à des performances dégradées. Les stratégies de couplage partiel offrent donc une alternative avantageuse lorsque le nombre de variables considérées est important.

4 Application à la cytométrie en flux

Dans cette partie, la méthode de factorisation couplée est appliquée à des données réelles de CMF. Elles ont été obtenues en mélangeant 3 populations cellulaires : Lymphocytes B (LB), Lymphocyte T (LT), Macrophages (MP). Les cellules ont ensuite été marquées avec 4 marqueurs (CFSE, CD4, CTV, MHCII) ayant des réponses différentes selon la population de cellules (voir Table 2). Pour chaque expérience, $N = 10^5$ cellules ont été analysées. Les résultats de la Figure 2 ont été obtenus par un traitement manuel (*gating*) réalisé par des biologistes à partir du nuage de points des marqueurs CFSE et CTV, et seront considérés comme vérité terrain dans notre étude.

TABLE 2 – Propriétés des 3 populations de l'expérience contrôlée. + correspond à des expressions fortes et - des expressions faibles.

	Expression des marqueurs			
Population	CFSE	CD4	CTV	MHCII
Macrophages	-	-	+	+
Lymphocytes B	+	-	-	++
Lymphocytes T	-	+-	-	-

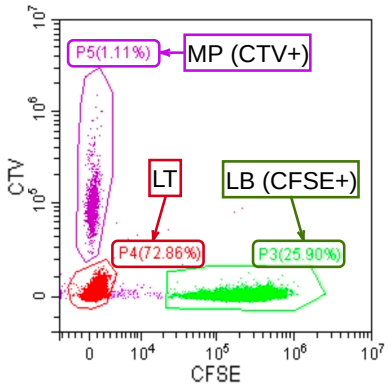

FIGURE 2 – *Gating* manuel. P3 sépare les cellules CFSE+ (LB), P4 les cellules CTV− et CFSE− (LT) et P5 les cellules CTV+ (MP).

4.1 Visualisation et clustering

Après application de la factorisation couplée, notre méthode fournit M matrices facteurs et le vecteur $\boldsymbol{\lambda}$ qui permettent d'approcher la densité M-dimensionnelle. Le choix du rang de la décomposition influe sur la qualité de l'approximation. Plus R est grand, meilleure est l'approximation mais plus le temps de calcul est important. Il apparaît également que pour bien représenter une population cellulaire, il est nécessaire de regrouper les termes de rang 1 ayant des propriétés similaires. Nous procédons alors à un clustering hiérarchique (*single linkage*) pour obtenir un dendrogramme. La métrique utilisée mesure la corrélation entre les termes de rang 1 selon :

$$D(r,s) = 1 - \prod_{m=1}^{M} \left\langle \widehat{\boldsymbol{a}_r^{(m)}}, \widehat{\boldsymbol{a}_s^{(m)}} \right\rangle \quad (9)$$

où $\langle ., . \rangle$ représente le produit scalaire. Les composantes sont alors groupées pour former une population si leur distance (9) est inférieure à une valeur seuil. En diminuant ce seuil, on peut explorer plus en profondeur les données alors qu'à l'inverse, augmenter le seuil donne une vue plus globale des données.

La Figure 3 montre les résultats obtenus sur la même expérience que celle de la Figure 2. Le rang de la décomposition est R=85 et le clustering permet d'obtenir 3 populations dont les propriétés sont en parfait accord avec les réponses attendues de la Table 2. La méthode permet d'estimer la taille

TABLE 3 – Dépendance entre le choix du rang sur l'estimation de la proportion des macrophages (MP).

		Proportion de Macrophage		
Gating		20.7%	8%	1.1%
Notre méthode	$R = 20$	20.2%	6.7%	0.83%
	$R = 40$	20.1%	7.1%	0.83%
	$R = 85$	20.6%	7.8%	0.9%

TABLE 4 – Estimation de la proportion de Macrophage pour les stratégies de couplage partiel et total (4 vs 2 triplets).

	Proportion de Macrophage	
Gating	Couplage total ('1')	Couplage partiel ('+1')
20.7%	20%	17.2%
8%	7.7%	8.3%
1.1%	0.91%	0.83%

de chacune des 3 populations. La Table 3 montre les résultats obtenus en fonction du rang de la décomposition pour 3 expériences différentes où la proportion de macrophages varie. Il apparaît qu'une augmentation du rang permet d'améliorer la qualité d'estimation , surtout dans le cas de populations rares. On note également que R=85 est supérieur à la condition d'identifiabilité de [8] qui ne fournit qu'une condition suffisante que l'on peut qualifier de pessimiste. La Table 4 permet de comparer les résultats obtenus avec couplage total et couplage partiel (stratégie '+1'). La stratégie de couplage partiel permet de retrouver des proportions du même ordre de grandeur avec cependant une perte de précision par rapport au couplage total. En outre, avec la PDF décomposée en R termes de rang 1, il est possible de calculer une distance entre chacune de ces composantes et d'alimenter des méthodes de clustering utilisées en CMF comme SPADE [3] ou viSNE [4] pour un coût extrêmement réduit.

5 Conclusion

Nous proposons dans cette communication une approche probabiliste pour l'analyse et l'interprétation des données de cytométrie. Notre méthode est capable de reconstruire des histogrammes en grandes dimensions par décomposition tensorielle couplée. Le résultat de la décomposition est utilisé pour le clustering de populations cellulaires. Elle permet également de visualiser les résultats sans réduction de dimension, a contrario des méthodes existantes. L'approche fournit des résultats des résultats très prometteurs sur des données à plus de 20 marqueurs différents.

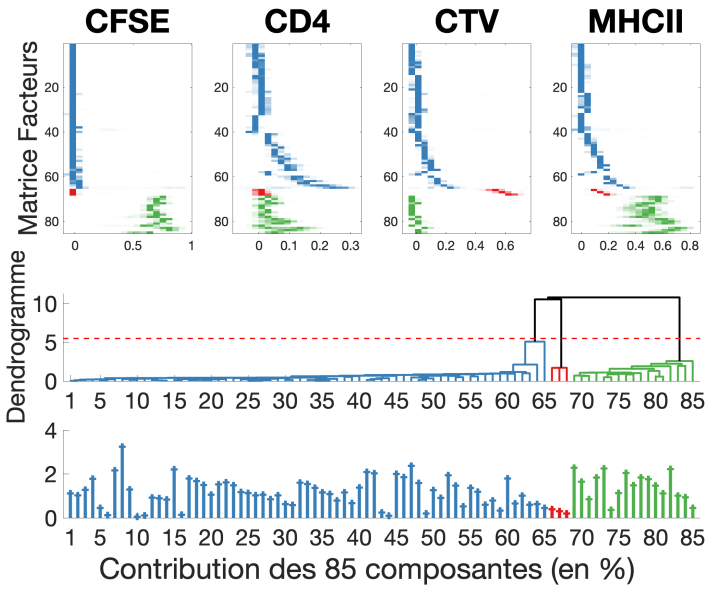

FIGURE 3 – Visualisation des résultats : **En haut** : Matrices facteurs représentant les $M = 4$ marqueurs ($R = 85$). **Au milieu** : clustering hiérarchique. **En bas** : taille de chaque composante en %.

Références

[1] Y. Saeys et al., "Computational flow cytometry : helping to make sense of high-dimensional immunology data," 2016.

[2] S. P. Perfetto et al., "Seventeen-colour flow cytometry : unravelling the immune system," *Nature Reviews Immunology*, 2004.

[3] P. Qiu et al., "Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE," *Nature biotechnology*, 2011.

[4] E. A. Amir et al., "viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia," *Nature Biotechnology*, 2013.

[5] S. Van Gassen et al., "FlowSOM : Using self-organizing maps for visualization and interpretation of cytometry data," *Cytometry Part A*, 2015.

[6] D. Brie et al., "Joint analysis of flow cytometry data and fluorescence spectra as a non-negative array factorization problem," *Chemometrics and Intelligent Laboratory Systems*, 2014.

[7] R. A. Harshman, *Foundations of the PARAFAC Procedure : Models and Conditions for an "explanatory" Multi-modal Factor Analysis*, UCLA working papers in phonetics, 1970.

[8] N. Kargas et al., "Tensors, learning, and "kolmogorov extension" for finite-alphabet random vectors," *IEEE Transactions on Signal Processing*, 2018.

[9] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes, "Identifiability of parameters in latent structure models with many observed variables," vol. 37, no. 6, pp. 3099–3132, Publisher : Institute of Mathematical Statistics.

[10] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Review*, 2009.

[11] Y. Qi et al., "Uniqueness of nonnegative tensor approximations," *IEEE Transactions on Information Theory*, 2016.