Coupled tensor factorization for flow cytometry data analysis

Philippe Flores, Guillaume Harlé, Anne-Béatrice Notarantonio, Konstantin Usevich, Maud d'Aveni, Stéphanie Grandemange, Marie-Thérèse Rubio, David Brie

2022 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, AGU. 22–25, 2022, XI'AN, CHINA

COUPLED TENSOR FACTORIZATION FOR FLOW CYTOMETRY DATA ANALYSIS

Philippe Flores[1], *Guillaume Harlé*[1], *Anne-Béatrice Notarantonio*[2,3], *Konstantin Usevich*[1], *Maud D'Aveni*[2,3], *Stéphanie Grandemange*[1], *Marie-Thérèse Rubio*[2,3], *David Brie*[1]

[1]CRAN, Université de Lorraine, CNRS, Nancy ; [2]IMoPA, Université de Lorraine, Nancy ;
[3]Hematology Unit, University Hospital of Nancy, Nancy.
firstname.lastname@univ-lorraine.fr

ABSTRACT

In this paper, we propose a new method for automated flow cytometry data analysis. By modeling a multidimensional probability distribution as a mixture of simpler distributions, we can reformulate the problem as a coupled tensor approximation of 3D marginals. In order to reduce the computational load, we use partially coupled strategies. We also propose a grouping of rank-one components together with a new visualization of the results. We demonstrate the usefulness of the proposed methodology on simulated and real data.

Index Terms— Flow cytometry, Naive Bayes model, Coupled tensor factorization

1. INTRODUCTION

Flow cytometry (FCM) is one of the most popular techniques for biological cell analysis. It is the reference technique in immunology because it allows for identification of rare cell population and thus improves knowledge of the human immune system [1]. From a data analysis point of view, a cytometer produces a point cloud in an M-dimensional space, where each point measured represents M characteristics called markers. The aim is to identify the different cell populations in this set of data points. Conventional analysis carried out manually by practitioners essentially consists of a series of 2-dimensional analyses; it becomes complex, subjective and costly in terms of manpower and time when the number of markers M increases. This has motivated the development of automatic methods [2, 3], which are still costly and difficult to apply to large data sets. Furthermore, these methods have limited performance for the analysis of rare cell populations, and their associated visualization tools are often difficult to interpret by end-users. In this paper, we propose a probabilistic approach based on the estimation of the joint density of the data. To cope with the curse of dimensionality, we adopt a naive Bayes model of the joint density: under this model, estimating the M-dimensional histogram can be reduced to estimating the factors of a tensor CP model [4] whose complexity remains linear with the number of dimensions. Inspired by [5], the estimation problem is reformulated as a coupled factorization problem of 3D marginals. In order to reduce the complexity of the algorithm, different partial coupling strategies are proposed and evaluated. The cell populations are obtained by applying a hierarchical clustering to the rank-one terms.

2. NAIVE BAYES MODEL

Let $\boldsymbol{x} = \left(X^{(1)}, \ldots, X^{(M)}\right)$ be a random vector taking values in $\mathcal{I}^{(1)} \times \cdots \times \mathcal{I}^{(M)}$ where $\mathcal{I}^{(m)} = [x_{\min}^{(m)}, x_{\max}^{(m)}]$.

978-1-6654-8547-0/22//$31.00

We assume that the N rows $\boldsymbol{x}_n$ of $\boldsymbol{X}$ are realizations of the random vector $\boldsymbol{x}$. Our goal is to estimate the multivariate probability density function (PDF) $\mathrm{p}(\boldsymbol{x})=\mathrm{p}(X^{(1)},\ldots,X^{(M)})$ from the observation matrix $\boldsymbol{X}$. One of the approaches for density estimation is to consider an M-dimensional histogram. In this case, each interval $\mathcal{I}^{(m)}$ is separated in I equal bins from $\Delta_1^{(m)} = [x_0^{(m)}, x_1^{(m)}]$ to $\Delta_I^{(m)} = [x_{I-1}^{(m)}, x_I^{(m)}]$, where $x_0^{(m)} = x_{\min}^{(m)}$ and $x_I^{(m)} = x_{\max}^{(m)}$. The histogram $\boldsymbol{\mathcal{H}} \in (\mathbb{R}^I)^M$ can be interpreted as the discretized joint PDF:

$$\begin{aligned}\mathcal{H}_{i_1\ldots i_M} &= \Pr(\boldsymbol{x} \in \Delta_{i_1}^{(1)} \times \cdots \times \Delta_{i_M}^{(M)}) \\ &= \int_{\Delta_{i_1}^{(1)}} \cdots \int_{\Delta_{i_M}^{(M)}} \mathrm{p}(\boldsymbol{x}) dX^{(1)} \cdots dX^{(M)}\end{aligned} \quad (1)$$

A naive approach to estimate the histogram from the samples is to count the number of samples $\boldsymbol{x}_n$ in each M-dimensional bin: $\widetilde{\mathcal{H}}_{i_1\cdots i_M} =$

$$\frac{1}{N}\,\mathrm{Card}\Big\{n \in [\![1,N]\!] \Big| \boldsymbol{x}_n \in \Delta_{i_1}^{(1)} \times \cdots \times \Delta_{i_M}^{(M)}\Big\}. \quad (2)$$

However, it requires a number of samples growing exponentially with the number of dimensions. To give some figures, with M=8 and I=20, the histogram is described with $I^M \approx 10^{10}$ values and much more samples are required to produce an accurate estimation. This drawback is referred to as the curse of dimensionality. To cope with it, we follow the approach of [5] which uses a Naive Bayes Model (NBM) whose complexity remains linear with the number of dimensions [6].

The NBM [5] introduces a discrete latent variable L taking values in $\{1,\ldots,R\}$, such that the element of $\boldsymbol{x}$ are conditionally independent on L:

$$\mathrm{p}(\boldsymbol{x}) = \sum_{r=1}^{R} \Pr(L=r) \prod_{m=1}^{M} \mathrm{p}(X^{(m)}|L=r). \quad (3)$$

By plugging (3) into (1), we get that the NBM corresponds to an order-M Canonical Polyadic Decomposition (CPD) [7] of $\boldsymbol{\mathcal{H}}$:

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

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

where $\boldsymbol{\lambda} \in \mathbb{R}^R$ is containing the probabilities $\Pr(L=r)$, and $\boldsymbol{a}_r^{(m)}$ is the 1D-conditional marginal in the m-th dimension (i.e., the values of $\Pr(X^{(m)} \in \Delta_{i_m}^{(m)} | L=r)$). Thus R corresponds to the tensor rank of $\boldsymbol{\mathcal{H}}$. In addition, as the factor matrices $\boldsymbol{A}^{(m)} = (\boldsymbol{a}_1^{(m)} \ldots \boldsymbol{a}_R^{(m)}) \in \mathbb{R}^{I\times R}$ and the vector $\boldsymbol{\lambda} \in \mathbb{R}^R$ represent probabilities, they should satisfy the non-negativity constraints ($\boldsymbol{\lambda} \geq 0$, $\boldsymbol{A}^{(m)} \geq 0$), and simplex (sum-to-one) constraints ($\mathbb{1}^\mathsf{T}\boldsymbol{\lambda} = 1$, $\mathbb{1}^\mathsf{T}\boldsymbol{A}^{(m)} = \mathbb{1}^\mathsf{T}$).

3. COUPLED TENSOR FACTORIZATION

3.1. Fully coupled tensor factorization

The coupled tensor factorization approach is based on the fact that the marginalized NBM [5] is still an NBM. Indeed, the 3D histogram $\boldsymbol{\mathcal{H}}^{(jk\ell)}$ of a subset of variables $(X^{(j)}, X^{(k)}, X^{(\ell)})$ has the CPD

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

due to sum-to-one properties. The 3D histograms for the subset are estimated with $\widetilde{\mathcal{H}}_{i_j i_k i_\ell}^{(jk\ell)} =$

$$\frac{1}{N}\,\mathrm{Card}\Big\{n \in [\![1,N]\!] \Big| \boldsymbol{x}_n \in \Delta_{i_j}^{(j)} \times \Delta_{i_k}^{(k)} \times \Delta_{i_\ell}^{(\ell)}\Big\},$$

which are easily computable compared to the full M-D histogram (2). The idea then is to estimate the factors in the full NBM via a coupled tensor approximation of 3D histograms. For a set of triples $\mathcal{T}$ we consider the following optimization problem:

$$\begin{aligned}&\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{A}^{(1)}}, \ldots, \widehat{\boldsymbol{A}^{(M)}} = \\ &\underset{\boldsymbol{\lambda}, \boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(M)}}{\operatorname{argmin}} \sum_{(j,k,\ell)\in\mathcal{T}} \Big\| \widetilde{\boldsymbol{\mathcal{H}}}^{(jk\ell)} - \Big[\!\Big[\boldsymbol{\lambda}; \boldsymbol{A}^{(j)}, \boldsymbol{A}^{(k)}, \boldsymbol{A}^{(\ell)}\Big]\!\Big] \Big\|_F^2 \\ &\text{s.t. } \boldsymbol{\lambda} \geq 0, \boldsymbol{A}^{(m)} \geq 0, \mathbb{1}^\mathsf{T}\boldsymbol{\lambda} = 1, \mathbb{1}^\mathsf{T}\boldsymbol{A}^{(m)} = \mathbb{1}^\mathsf{T}. \quad (6)\end{aligned}$$

If $\mathcal{T}=\{(j,k,\ell) \in [\![1, M]\!]^3 \,|j<k<\ell\}$, the problem is referred to as the fully coupled tensor factorization, which was initially proposed in [5]. However, as we will show later, it not necessary to consider all possible triples. The problem (6) is solved with a coupled AO-ADMM [5].

3.2. Identifiability conditions

Tensor decompositions possess strong uniqueness (identifiability) properties [7]. In particular, if each $\boldsymbol{\mathcal{H}}^{(jk\ell)}$ is individually generically identifiable, then the probability tensor $\boldsymbol{\mathcal{H}}$ is itself identifiable. For example, the Kruskal generic uniqueness condition for $\widetilde{\boldsymbol{\mathcal{H}}}^{(jk\ell)}$ is $R \leq (3M-2)/2$. However, since many $\boldsymbol{\mathcal{H}}^{(jk\ell)}$'s share common factors, the identifiability conditions can be significantly improved. Assuming $M \leq I$, $\boldsymbol{\mathcal{H}}$ can be shown to be generically identifiable if $R \leq I(M-2)$ [5]. Note that these identifiability results were derived in a noiseless setup (exact decomposition) and are stated for real (possibly non-negative) factor matrices. In practice, due to the limited sample size, only noisy $\widetilde{\boldsymbol{\mathcal{H}}}^{(jk\ell)}$ are available, which leads to a low rank tensor approximation problem. Adding nonnegativity constraints on the latent factors is beneficial since this ensures the existence and uniqueness of the low-rank tensor approximation, see [8]. Finally, a closer look at the proof of the identifiability results in [5] reveals that only the identifiability of an extended tensor defined by a specific partition of the M variables is required. In other words, only a limited number of triples are necessary to ensure identifiability. This idea is developed in the next subsection to reduce the computational burden of the coupled tensor factorization.

3.3. Partially Coupled Tensor Factorization

The partially coupled strategies are motivated by the high number of 3D histograms in the case of a fully coupled approach [5]. Indeed, $\binom{M}{3}$ histograms must be estimated which leads to difficulties in practice (lack of storage and prohibitive computational complexity). To overcome this problem, partial coupling considers a subset of triples in $\mathcal{T}$, thus less 3D marginals are considered in the coupling. Strategies of different complexity can be considered but each subset must always contain at least one occurrence of each variable. In our study, we considered 6 strategies explained in **Table 1**. The '+2' strategy, in which two consecutive triples have only one marker in common, is one of the smallest subset of triples possible, whereas '+1' consecutive triples have two markers in common. For the other strategies, a subset of triples is randomly selected from one eighth of all triples ('1/8') to all triples possible ('1').

Table 1: Coupling strategies for M=10.

Strategy	Card $(\mathcal{T})$	Triples
+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), … (9,10,1), (10,1,2)
1/8	15=120/8	random triples
1/4	30=120/4	random triples
1/2	60=120/2	random triples
1	120=$\binom{10}{3}$	random triples

3.4. Performance evaluation

To study the performance of coupling strategies, we applied our method with all strategies of **Table 1** with M=10 dimensions synthetic data. R=20 multivariate Gaussian distributions were generated randomly and added with weights $\boldsymbol{\lambda}$ together to create a theoretical histogram $\boldsymbol{\mathcal{H}}$. We generated different number of samples N=$\{10^4, 3\cdot 10^4, 10^5, 3\cdot 10^5, 10^6\}$ and computed the histograms for I=30 bins. We set N_1=10^3 outer iterations and N_2=20 inner iterations for the coupled AO-ADMM and the target rank equal to R. The reconstruction error

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

is evaluated for each strategy and averaged over 10 experiments. **Fig. 1** shows that the coupling strategy '1/8' yields similar performance with the fully

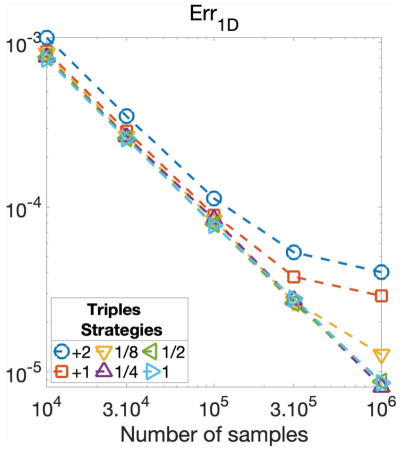

Fig. 1: Evolution of 1D marginals error for different strategies presented in **Table 1**

coupled strategy. This shows that partial coupling strategies are beneficial in terms of computational cost as the computational complexity is linear with the number of triples in consideration.

4. APPLICATION TO FLOW CYTOMETRY

4.1. Visualization and clustering

The output of the coupled tensor factorization are the M factor matrices $\widehat{\boldsymbol{A}^{(m)}}$ and $\widehat{\boldsymbol{\lambda}}$ which approximate the M-dimensional density function using R rank-one components. The choice of the CP decomposition rank R is crucial. The quality of the approximation increases with the rank R at the price of a longer computation time. In order to allow end users to properly analyze and interpret the data, adequate visualization tools should be developed. This is not a simple task when the number of dimensions is large. From this point of view, the CP decomposition is often appealing since it allows for joint visualization of the M variables representing the M-dimensional density. However, when R is larger than the actual number of populations, a single population is obtained by gathering the rank-one terms with similar properties. In that respect, a hierarchical clustering procedure [9] is applied to the rank-one components with the following distance:

$$D(r,s) = \left\| \mathbb{E}\left[\widehat{\boldsymbol{a}_r^{(m)}}\right] - \mathbb{E}\left[\widehat{\boldsymbol{a}_s^{(m)}}\right] \right\|_2^2 \qquad (8)$$

where $\mathbb{E}[\widehat{\boldsymbol{a}_r^{(m)}}] = \sum_{i=1}^{I} \overline{\Delta_i^{(m)}}\, \widehat{\boldsymbol{a}_{ri}^{(m)}}$ is the (scaled) expectation of the estimated marginal distributions ($\overline{\Delta_i^{(m)}}$ represents the centroid of the i-th bin). The clusters are then obtained by grouping rank-one terms in the dendrogram whose distance (8) is lower than a specified threshold. Each cluster is represented using the same color. The whole processing pipeline is termed *CTFlowHD*.

Table 2: Properties of the 3 populations used in the controlled experiment. + represents high marker expression and - low expression.

	Marker expression			
Population	CFSE	CD4	CTV	MHCII
Macrophage	-	-	+	+
Lymphocyte B	+	-	-	++
Lymphocyte T	-	+-	-	-

4.2. Applications on FCM controlled data

To validate our method, *CTFlowHD* is applied to FCM controlled data sets after compensation [10] and a non-linear transformation [11]. Three cell lines were considered: Lymphocytes B (LB), Lymphocyte T (LT) and Macrophage (MP), having different responses according to M=4 markers (see **Table 2**). Then, cells were mixed in different proportions resulting in 3 different data sets with N=10^5 cells, where the MP proportion varies (approximately 20%, 8% and 1%). 3D histograms with I=20 bins were estimated from the data sets. Ground-truth population sizes were obtained by a manual gating **Fig. 2** as the 3 cell lines are easily recognizable. With *CTFlowHD*: 3 groups were separated by the hierarchical clustering. As LB and LT did not represent rare cell populations in the 3 experiments (always above 25%), estimated sizes were accurately estimated and their estimation

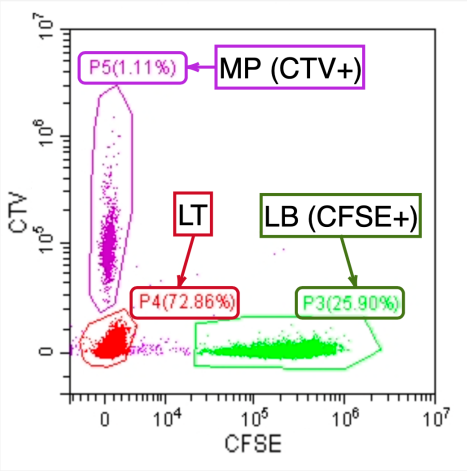

Fig. 2: Manual gating of the 3 populations. The plot shows three gates : P3 groups CFSE+ cells (LB), P4 groups LT cells and P5 CTV+ cells (MP). *Logicle* scale is used on this plot [11]

will not be addressed here. Concerning the MP cell population, **Fig. 3** shows the estimation of MP proportion for the 3 experiments and clearly shows comparable results between manual gating and *CTFlowHD*. Increasing the rank of the decomposition improves the accuracy of the estimated percentage. It is also noted that R=85 is higher than the rank bound given by the identifiability condition of [5] which appears to be pessimistic. **Table 3** gives a comparison of the results obtained with the full and partial coupling ('+1' strategy). The partial coupling strategy yields to similar sizes but with an accuracy loss as compared to the full coupling.

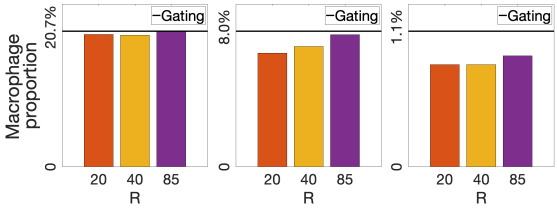

Fig. 3: Rank influence on the macrophage population size estimation. Gating is considered ground-truth.

Table 3: Estimation of MP population sizes for fully and partially coupled strategies (4 vs. 2 triples).

	Macrophage proportion	
Gating	Fully coupled	Partially coupled
20.7%	20%	17.2%
8%	7.7%	8.3%
1.1%	0.91%	0.83%

4.3. Application to 8-dimension FCM data

To validate partial coupling strategies, we ran *CTFlowHD* on a M=8 FCM real dataset of N=10^6 cells currently analyzed by immunologists. Our method was applied with I=30 bins and a R=100 decomposition rank for strategies '1' and '1/4' (see **Table 1**). **Fig. 4** shows *CTFlowHD* visualizations are similar, while **Table 4** shows similar population size estimation results for both strategies. This shows that partial coupling reduces the computational burden while keeping interpretability.

Table 4: *CTFlowHD* estimation of cell population sizes for fully and partially coupled strategies in $M = 8$ dimensions. Colors match with **Fig. 4**.

	Population size	
Cell population	Full ('1')	Partial ('1/4')
Blue	36.3%	34.4%
Red	11.8%	12.4%
Green	28.2%	29.4%
Purple	1.3%	1.1%
Orange	22.1%	22.4%
Brown	0.27%	0.23%

5. CONCLUSION

We provide a new probabilistic method that allows biologists to interpret the flow cytometry data jointly with all dimensions, compared to existing flow cytometry methods. This method is able to recover high-dimensional histograms and separate them into populations of cells. Even if coupled ten-

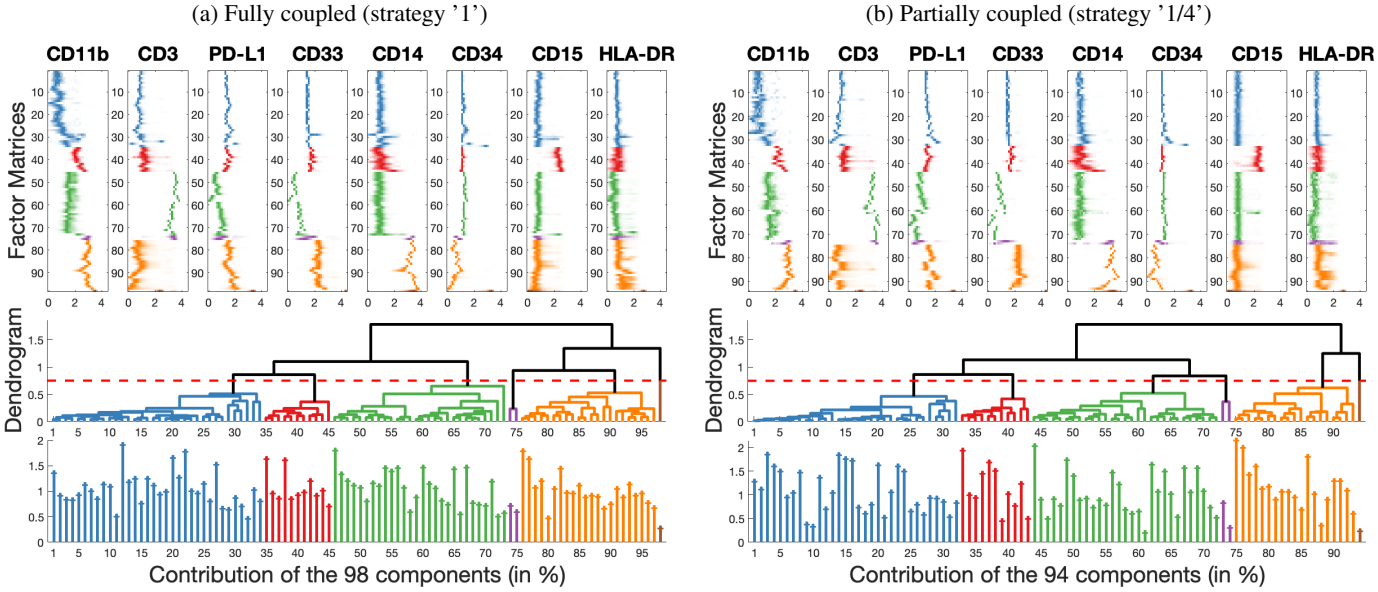

Fig. 4: *CTFlowHD* results in 8 dimensions for coupling strategies '1' and '1/4' (see **Table 1**). **Upper plots**: Factor matrices where 1D marginals are plotted in rows for each component. **Middle plot** : dendrogram obtained with a complete linkage clustering. **Lower plot** : size of each rank-one component in %.

sor factorization is able to cope with the curse of dimensionality, partial coupling permits to reduce even more the computational burden while keeping similar performance of the estimation.

6. REFERENCES

[1] S. P. Perfetto et al., "Seventeen-colour flow cytometry: unravelling the immune system," *Nature Reviews Immunology*, vol. 4, no. 8, pp. 648–655, 2004.

[2] P. Qiu et al., "Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE," *Nature biotechnology*, vol. 29, no. 10, pp. 886–891, 2011.

[3] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

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

[5] N. Kargas et al., "Tensors, learning, and "Kolmogorov extension" for finite-alphabet random vectors," *IEEE Transactions on Signal Processing*, vol. 66, no. 18, pp. 4854–4868, 2018.

[6] N. Kargas and N. D. Sidiropoulos, "Learning mixtures of smooth product distributions: Identifiability and algorithm," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. 2019, vol. 89, pp. 388–396, PMLR.

[7] T. Kolda and B. Bader, "Tensor Decompositions and Applications," *SIAM Review*, vol. 51, pp. 455–500, 2009.

[8] Y. Qi et al., "Uniqueness of nonnegative tensor approximations," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 2170–2183, 2016.

[9] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[10] M. Roederer, "Compensation in flow cytometry," *Current protocols in cytometry*, vol. 22, no. 1, pp. 1.14. 1–1.14. 20, 2002, ISBN: 1934-9297.

[11] D. R. Parks et al., "A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data," *Cytometry Part A*, vol. 69A, no. 6, pp. 541–551, 2006.