Chemometrics and Intelligent Laboratory Systems 137 (2014) 21–32

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Joint analysis of flow cytometry data and fluorescence spectra as a non-negative array factorization problem

David Brie [a,*], Rémi Klotz [a], Sebastian Miron [a], Saïd Moussaoui [b], Christian Mustin [c], Philippe Bécuwe [a], Stéphanie Grandemange [a]

[a] Centre de Recherche en Automatique de Nancy (CRAN), Université de Lorraine, CNRS, Campus Sciences, B.P. 70239, F-54506 Vandœuvre-lès-Nancy, France
[b] Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), l'UNAM, Ecole Centrale de Nantes, 1 rue de la Noë, 1, B.P. 92101, F-44321 Nantes, France
[c] Laboratoire Interdisciplinaire des Environnements Continentaux (LIEC), Université de Lorraine, CNRS, Campus Sciences, B.P. 70239, F-54506 Vandœuvre-lès-Nancy, France

ARTICLE INFO

Article history:
Received 21 February 2014
Received in revised form 28 May 2014
Accepted 5 June 2014
Available online 16 June 2014

Keywords:
Flow cytometry
Fluorescence spectroscopy
Multivariate probability density functions
Non-negative block CANDECOMP/PARAFAC decomposition
Non-negative matrix factorization
Mitochondrial membrane potential
JC-1 probe

ABSTRACT

The paper presents a novel approach to the processing of flow cytometry data sequences. It consists in decomposing a sequence of multidimensional probability density functions by using the multilinear block tensor decomposition approach [1,2]. Also a formal link between flow cytometry data and fluorescence spectra is provided allowing the joint processing of both data. To illustrate the effectiveness of the approach, a study of the T47D cell line mitochondrial membrane potential as a function of the CCCP decoupling agent concentration is performed. The main advantages of the method are: (i) the flow cytometry data compensation is no longer necessary, and (ii) the cell sorting capacity of the method is significantly improved as compared to classical clustering methods. As a byproduct, it was possible to observe directly on the result of the processing, the dependence of the cell mitochondrial membrane potential with respect to the cell cycle phase. The proposed method is quite general provided that it is possible to design an experiment allowing the observation of the response of cell populations to an environmental/chemical/biological parameter.

1. Introduction

Flow cytometry is an investigation technique widely used in biology and medicine for the characterization and quantification of the morphological, density and fluorescence properties of cells. A highly insightful approach, often used in biology and biomedical studies, consists in studying the evolution (response) of a cell population with respect to environmental/chemical/biological parameters (*e.g.* temperature, chemical drugs, and gene expression). An overview of recent techniques of flow cytometry data analysis is given in [3] and several challenges, that have to be dealt with, are also pointed out. In particular, it appears that the recent technological progresses of cytometers allow the design of complex multiparameter experiments yielding a large amount of data. Classical flow cytometry data analysis methods are no longer adapted to these data and hence there is a need for new algorithms that can efficiently retrieve the relevant information from this large amount of data. Classical flow cytometry data processing consists in a sequence of procedures mainly relying on the user expertise, and is, therefore, somewhat subjective. In general, the different operations are as follows:

Gating consists in manually selecting a cell population of interest within the dot plot.
Compensation aims at minimizing the influence of the spectral overlapping for different cell sub-populations. It consists in applying a linear transformation to the data, whose parameters are user-defined.
Clustering is a technique to perform pattern classification in unsupervised environments that may be used in cell sorting. In most manufacturer provided software, this operation consists in manually defining quadrants on the dot plots. Most of these methods require a user decision step which may strongly affect the relevance of the results. This is especially the case when the sub-population distributions strongly overlap, as often in practical applications.

In the last decades, a number of effective clustering algorithms have been proposed, including k-means related methods (*e.g.* [4–7]) or Gaussian mixture models (*e.g.* [6,8–10]). In [7], Lo et al. propose the use of *t*-distribution mixtures instead of Gaussian mixtures, as they allow better handling of outliers, due to their heavier tail. In [11], the performances of the available automated cytometry data analysis

* Corresponding author.
E-mail address: david.brie@univ-lorraine.fr (D. Brie).

http://dx.doi.org/10.1016/j.chemolab.2014.06.003
0169-7439/

techniques are assessed on three datasets having different complexities. It appears that automated methods (eventually coupled) yield effective results as compared to the manual clustering approach. However, the population overlapping still remains a challenging problem requiring further development. Another critical issue mentioned in [3], corresponds to the case of infrequent and/or heterogeneous populations.

In [12,13], an information geometry approach allows the definition of similarity measures between cytometry datasets facilitating data interpretation by clinicians and resulting in a low dimension representation of the data. This approach is somewhat related to the one proposed here since it jointly considers multiple cytometry datasets.

The contributions of this work are twofold: *i*) we introduce a cell population sorting method based on a non-negative block tensor decomposition of the data histograms. The key point is that it is a fully multidimensional approach in which cell sorting is done according to the response of the different sub-populations to a parameter (the CCCP concentration in this paper). The model identifiability is also studied. The main advantages of our method are: it is almost fully unsupervised (the only input parameter is the number of sub-populations sought in the data), it is non-parametric (there is no underlying parametric probability density function), it is not much affected by the overlapping of the sub-population distributions and it does not require any decision step since the problem is addressed as a source separation problem which provides both the amount of each sub-population and its probability density function; *ii*) we prove that there is a relationship between flow cytometry and bulk spectroscopy data, and propose a joint processing approach of the two data modalities. This improves the separation accuracy and provides a more complete description of the analyzed cell populations. This joint modality data analysis can be regarded as a data fusion approach, as proposed in [14] in the context of polarized Raman spectroscopy. This type of data fusion approach has gained a lot of interest recently in various domains of application (see *e.g.* [15,16]).

The remainder of this paper is organized as follows: in the next section, we introduce the notations and some general assumptions that are used throughout this paper; in Section 3, the data model for the proposed approach is derived and the link between spectroscopy and cytometry data is highlighted; in Section 4 we analyze the identifiability of the proposed model and propose a three-step algorithm for sorting the cell sub-populations; Section 5 gives an illustration of the applicability of the proposed approach to the study of the T47D cell line mitochondrial membrane potential as a function of the CCCP decoupling agent concentration. It includes the experiment description and the results of the proposed approach applied to different datasets; some conclusions are given in Section 6.

2. Preliminaries

Lowercase letters ($x, y, \ldots$) denote scalars, boldface lowercase ($\mathbf{x}, \mathbf{y}, \ldots$) are used for vectors, boldface capitals ($\mathbf{X}, \mathbf{Y}, \ldots$) symbolize matrices and tensors are written in boldface calligraphic capital letters ($\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}, \ldots$). A *tensor* or *N*-way array ($N \geq 3$) can be seen as the generalization of matrices to the multidimensional case. The number of dimensions *N* is called the *order* of the tensor. Thus, a vector is a first order tensor, a matrix is a second order tensor, *etc.* Consider a 3-way data array (third order tensor) $\boldsymbol{\mathcal{X}}(I \times J \times K)$ admitting the following decomposition in a sum of *K* terms:

$$\boldsymbol{\mathcal{X}} = \sum_{k=1}^{K} \mathbf{a}_k \circ \mathbf{b}_k \circ \mathbf{c}_k \tag{1}$$

where $\mathbf{a}_k(I \times 1)$, $\mathbf{b}_k(J \times 1)$ and $\mathbf{c}_k(L \times 1)$ are vectors and "$\circ$" denotes the outer product. The three dimensions of $\boldsymbol{\mathcal{X}}$ are referred to as *modes*. The quantity $\mathbf{a}_k \circ \mathbf{b}_k \circ \mathbf{c}_k$ represents a rank-1 tensor and the decomposition in Eq. (1) is commonly known as CANDECOMP/PARAFAC (CP) [17,18]. If *K* is the minimum number of rank-1 tensors that yield exactly $\boldsymbol{\mathcal{X}}$, then *K* is called the *rank* of the CP decomposition. An alternative notation for Eq. (1) is

$$\boldsymbol{\mathcal{X}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!], \tag{2}$$

where $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_K]$, $\mathbf{B} = [\mathbf{b}_1 \ldots \mathbf{b}_K]$ and $\mathbf{C} = [\mathbf{c}_1 \ldots \mathbf{c}_K]$ denote the component/loading matrices. All these notions, defined here for the 3-way case, generalize straightforwardly to *N*-way arrays ($N > 3$).

For simplicity, throughout this paper, the noise/error term in data model expressions will be ignored, which nothing detracts from the generality of the presented method. To introduce the theoretical data model in Section 3, continuous probability density functions (*pdf*s) should be employed. However, in practice, the recorded data are represented by histograms, implying discretized versions of these *pdf*s. For the clarity of the presentation, we will use lowercase letters to denote the continuous *pdf*s and boldface lowercase for their discretized versions. The length of the vectors representing the discretized *pdf*s (*i.e.* the number of bins) will not be explicitly mentioned, unless it is crucial for the comprehension of the presentation. Also, for simplification, the distinction continuous/discretized *pdf* will not always be made in the text, but can be easily deduced from the context.

3. Data model

3.1. The probability density function of N-dimensional flow cytometry data

Consider *N*-dimensional flow cytometry data. Each of the analyzed cells yields a length *N* vector measuring the amplitudes at *N* different wavelength values of the emitted fluorescence light. The set of measurements collected on a population of *M* different cells can be gathered in an $N \times M$ matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_M]$, where $\mathbf{x}_m = [x_m(1), \cdots, x_m(N)]^T$ and $m = 1, \ldots, M$. As illustrated in Fig. 1, this data matrix $\mathbf{X}$ can be characterized by its *N*-variate *pdf* denoted by $p(\mathbf{x}) = P(\mathbf{x}_m = \mathbf{x})$. At this point, it is necessary to pay some attention to the *pdf* estimation. We follow the main lines of [19] (chapter 4, pp 164–168). Density estimation can be done using a kernel approach among which the most well known is the Parzen window method. As a special case, the Parzen window can be chosen as a hypercube characterized by its widths along each dimension. However, other window functions can be used (*e.g.* Gaussian window). The choice of these widths is crucial because it will determine the accuracy of the estimated *pdf*. Basically, choosing a too large window will result in an over-smoothed *pdf* suffering from too little resolution while a too small width results in a high variance *pdf* estimate. Also, it is possible to let the width slowly go to zero as the number of samples

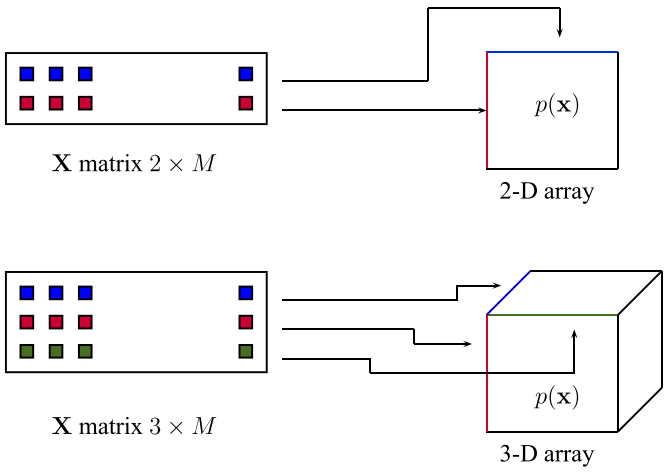


Fig. 1. Transforming the $(N \times M)$ matrix $\mathbf{X}$ into *N*-D histograms; illustrations for $N = 2$ and 3.

increases. The reader is referred to [19] for the precise statement of the convergence properties of the Parzen window estimator. In this work, we have only considered bi-dimensional *pdf* estimation. While the proposed approach can be extended to *N*-dimensional density estimation, theoretical arguments suggest that *pdf* estimation in high dimension ($N > 4$ or 5) is fruitless. This is referred to as the curse of dimensionality in [20]. This is a point requiring to be carefully studied in forthcoming studies. Future directions will be given in the Conclusions.

3.2. The bilinear model of a sequence of N-D pdfs

Let us consider a cell population composed of K different sub-populations. A sub-population is defined as a set of cells exhibiting identical/similar behaviors with respect to the variation of a physical parameter. We represent the N-D data points in the analyzed sample by the *pdf* $p(\mathbf{x})$ of the measurement vector $\mathbf{x}$. This *pdf* is expressed as a mixture of K density functions f_k, corresponding to the K sub-populations:

$$P(\mathbf{x}_m = \mathbf{x}) = p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k f_k(\mathbf{x}) \tag{3}$$

with $\sum_{k=1}^{K} \alpha_k = 1$. Assume that we study the response of this cell population to the variation of some physical parameter, denoted as s hereafter. For each value of s, a flow cytometry dataset can be recorded, resulting in a sequence of flow cytometry data matrices. Hence, each data matrix obtained for a given physical condition yields a *pdf* denoted by $p(\mathbf{x},s)$, that can be modeled as:

$$p(\mathbf{x}, s) = \sum_{k=1}^{K} \alpha_k(s) f_k(\mathbf{x}). \tag{4}$$

The sequence of N-D histograms, obtained for different values of s, can thus be gathered into a $(N+1)$-D array (tensor) denoted by $\mathcal{P}$. By unfolding this tensor along the dimensions corresponding to the N different wavelengths, we obtain a matrix $\mathbf{P}$ that admits the following bilinear factorization:

$$\mathbf{P} = \mathbf{A}\mathbf{F}^T. \tag{5}$$

In Eq. (5), $\mathbf{P}$ has a number of rows equal to the number of values of the parameter s; the columns of $\mathbf{A}$, symbolized by $\mathbf{a}_k$ ($k = 1, \ldots, K$), contain the mixing coefficients of the K sub-populations for the different values of s; the columns of $\mathbf{F}$ are the "unfolded" N-D density functions $\mathbf{f}_k$ of the K sub-populations.

3.3. The block-CANDECOMP/PARAFAC model of a sequence of N-D pdfs

Assuming the independence of each coordinate of $\mathbf{x} = [x_1, \cdots, x_N]^T$, the multivariate *pdf* $f_k(\mathbf{x})$ can be factorized as a product of N univariate *pdf*s: $f_k(\mathbf{x}) = f_k^1(x_1) \cdot f_k^2(x_2) \cdots f_k^N(x_N)$. Thus the data array can be written as a CP model of order $N+1$:

$$\mathcal{P} = \sum_{k=1}^{K} \mathbf{a}_k \circ \mathbf{f}_{k,1} \circ \cdots \circ \mathbf{f}_{k,N}. \tag{6}$$

Eq. (6) clearly expresses an $N+1$ CP model of rank K which can be alternatively written as:

$$\mathcal{P} = [\![\mathbf{A}, \mathbf{F}_1, \cdots, \mathbf{F}_N]\!], \tag{7}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$ and $\mathbf{F}_n = [\mathbf{f}_{1,n}, \ldots, \mathbf{f}_{K,n}]$, with $n = 1, \ldots, N$. The link between models (7) and (5) is obtained by unfolding $\mathcal{P}$ into a matrix, according to: $\mathbf{P} = \mathbf{A}(\mathbf{F}_1 \odot \cdots \odot \mathbf{F}_N)^T$, where "$\odot$" stands for the Khatri–Rao product. Thus, we have $\mathbf{F} = \mathbf{F}_1 \odot \cdots \odot \mathbf{F}_N$. Admittedly, assuming the independence of each coordinate of $\mathbf{x}$ does not allow the representation of the general probability density function. Restricting our attention to the case of bi-dimensional density functions ($N = 2$), we propose to adopt for the data the rank-$(L_k, L_k, 1)$ Block Component Model, equally known as the block-CANDECOMP/PARAFAC (BCP) model, introduced by De Lathauwer in [1,2]. By doing so, it is possible to consider the more general case of non-separable *pdf*s. In fact, this is nothing but performing a low rank approximation of the (discretized) *pdf*s. Hence, the $(N+1)$-D data array can be written as:

$$\mathcal{P} = \sum_{k=1}^{K} \mathbf{a}_k \circ \mathbf{E}_k, \tag{8}$$

where the rank L_k matrices $\mathbf{E}_k$ can be decomposed as:

$$\mathbf{E}_k = \sum_{l=1}^{L_k} \mathbf{f}_{k,1}^l \circ \mathbf{f}_{k,2}^l = \sum_{l=1}^{L_k} \mathbf{f}_{k,1}^l \mathbf{f}_{k,2}^{l\,T}. \tag{9}$$

A graphical illustration of this model is given in Fig. 2.

The BCP model can be seen as a CP model in which some of the loading vectors (columns) of the matrix $\mathbf{A}$ are collinear. The model in Eq. (8) can be easily generalized to higher dimensional ($N > 2$) data. The higher-dimensional case does not yield more complicated data processing situations, since it is known that higher-order CP models require less restrictive identifiability conditions. For example, some 4-order CP models, with collinear loading in at most three modes, are provably identifiable [21].

3.4. Joint analysis of fluorescence cytometry and bulk spectroscopy data

The question investigated in this subsection is the following: is there a formal link between the fluorescence spectra and the cytometry data for a given cell population? We are going to show that the answer is actually yes, allowing us to propose a joint analysis of the two types (modalities) of data. A bulk fluorescence spectrum corresponds to the fluorescence measured on a large number of cells. In that respect, it can be seen as an average of all the individual cell fluorescence spectra. Thus, considering the variation of the same physical parameter s, the measured spectra $\mathbf{m}(s)$ can be written as:

$$\mathbf{m}(s) = \int \tilde{\mathbf{x}}\, p(\tilde{\mathbf{x}}, s)\, d\tilde{\mathbf{x}}. \tag{10}$$

The vector $\tilde{\mathbf{x}}$ contains $\mathbf{x}$ as a "sub-vector", and therefore we can write[1]:

$$p(\tilde{\mathbf{x}}, s) = \sum_{k=1}^{K} \alpha_k(s) f_k(\tilde{\mathbf{x}}), \tag{11}$$

where the mixing coefficients $\alpha_k(s)$ are the same as in Eq. (10). By replacing Eq. (11) in Eq. (10), we obtain:

$$\mathbf{m}(s) = \sum_{k=1}^{K} \alpha_k(s) \underbrace{\int \tilde{\mathbf{x}} f_k(\tilde{\mathbf{x}}) d\tilde{\mathbf{x}}}_{\tilde{\mathbf{f}}_k} = \sum_{k=1}^{K} \alpha_k(s) \tilde{\mathbf{f}}_k. \tag{12}$$

If we regroup the set of spectra (after normalization to unit energy) recorded for the different values of s on the rows of a matrix $\mathbf{M}$, the following mixture model can be written:

$$\mathbf{M} = \mathbf{A}\tilde{\mathbf{F}}^T. \tag{13}$$

[1] In reality, the data vector $\mathbf{x}$ is obtained by integrating the emitted light over a wavelength interval $\Delta\lambda$ around the different "colors" used by the cytometer. However, for small values of $\Delta\lambda$, Eq. (11) is a good approximation of $p(\tilde{\mathbf{x}}, s)$.

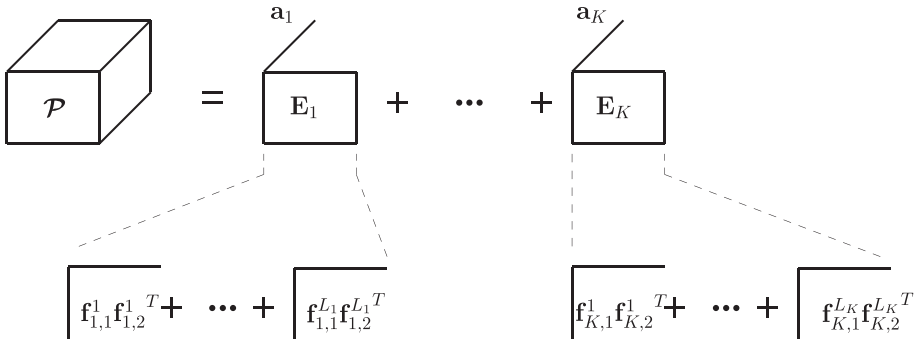


Fig. 2. Graphical illustration of the rank-(L_k,$L_{k,1}$) BCP model of $\mathcal{P}$.

In Eq. (13), the columns of $\widetilde{\mathbf{F}}$ are the vectors $\widetilde{\mathbf{f}}_k$ and correspond to spectra characterizing the (averaged) spectral response of the K cell sub-populations and the mixing matrix $\mathbf{A}$ is exactly the same as in the cytometry data model (5). We are now able to propose the joint model by gathering the two models (Eqs. (5) and (13)) into a single one. This is possible because of the common mixing matrix $\mathbf{A}$. We follow an approach quite similar to the one proposed in [14] which consists in concatenating the data matrices $\mathbf{P}$ and $\mathbf{M}$ according to:

$$[\mathbf{P} \quad \mathbf{M}] = \mathbf{A}\begin{bmatrix}\mathbf{F} \\ \widetilde{\mathbf{F}}\end{bmatrix}^T. \tag{14}$$

The joint use of a sequence of fluorescence flow cytometry and spectroscopy data provides a very complete description of cell population. It yields a decomposition of the whole population into homogeneous sub-populations characterized by their common:

- probability density function
- (averaged) spectral response
- response to a physical parameter.

4. Model identifiability and data processing

The problem at hand can be embedded in the general framework of non-negative approximation of non-negative tensors using multilinear decompositions. This is still an open problem but a milestone was reached with the work of Lim and Comon [22] where it is proved that non-negativity ensures the well posedness of the non-negative tensor approximation. As mentioned in [22], this can actually be associated with the sparse naive Bayes probabilistic model for *pdf* [23], since the underlying probabilistic model is a mixture of densities having independent variables. Here we go one step further since, by using the BCP decomposition, we can relax the independence assumption. However, the question of the validity of the non-negative BCP decomposition as an approximation tool is an open problem which would deserve to be studied. We did not pay further attention to this point, but as a first attempt to illustrate the regularization property of the non-negativity, we show that rank (L_k, L_k, 1) exact non-negative BCP decomposition can be unique without any additional assumption. This is not the case for general BCP unless some other constraints (such as orthogonality) are enforced. In Section 3, it was shown that NMF was involved in all the different models considered. In that respect, uniqueness of the NMF plays a central role and in the sequel we recall some results on the NMF uniqueness and we give a sufficient condition which allows checking directly on the data if a NMF is likely to be unique. We then use these results to study the uniqueness of the non-negative BCP model and we give some practical consequences of theses results. To conclude this section, we present the three different steps of the data processing algorithm.

4.1. Identifiability of the non-negative bilinear model

In this section, we address the identifiability of the bilinear model which arises in different contexts in this work, in particular for models (5) and (13). Such non-negative bilinear models are also involved in BCP decomposition, as we will show in Subsection 4.2. Assume that a non-negative matrix $\mathbf{W}$ admits an exact bilinear model representation:

$$\mathbf{W} = \mathbf{H}\mathbf{G}^T. \tag{15}$$

Depending on the considered case, the matrix $\mathbf{W}$ may represent different quantities: matrix $\mathbf{P}$ for model (5), Matrix $\mathbf{M}$ for Eq. (13), matrix $[\mathbf{P} \quad \mathbf{M}]$ for model (14), and matrix $\mathbf{E}_k$ for the BCP model (8).

It is well known, that in general the bilinear decomposition (Eq. (15)) does not admit a unique solution since for any non-singular matrix $\mathbf{T}$ it is possible to write:

$$\mathbf{W} = \mathbf{H}\mathbf{T}\,\mathbf{T}^{-1}\mathbf{G}^T = \widetilde{\mathbf{H}}\,\widetilde{\mathbf{G}}^T \tag{16}$$

which is another admissible solution. Regularization is thus needed to obtain a unique decomposition. Among the possible additional constraints which can be considered, we focus on the non-negativity assumption, leading to the non-negative matrix factorization (NMF) problem [24]. In order to discuss the identifiability of the NMF model (15), the notion of *simplicial cone* needs to be introduced.

Definition 1. Simplicial cone

The simplicial cone generated by a family of vectors $\{\mathbf{g}_n\}_{n=1}^N$ is

$$\mathcal{C}(\{\mathbf{g}_n\}) = \left\{\mathbf{w} : \mathbf{w} = \sum_n \alpha_n \mathbf{g}_n, \alpha_n > 0\right\}.$$

The *order* of a simplicial cone is the dimension of the subspace span $(\{\mathbf{g}_n\}_{n=1}^N)$. Based on the definition above, a necessary and sufficient condition for NMF identifiability has been provided by Chen in [25]:

Theorem 1. Necessary and sufficient uniqueness condition

Denoting $\mathcal{K}$ the convex hull of the data matrix $\mathbf{W}$, the decomposition of $\mathbf{W}$ according to $\mathbf{W} = \mathbf{H}\mathbf{G}^T, \mathbf{H} \geq 0, \mathbf{G} \geq 0$ is unique if and only if the simplicial cone $\mathcal{C}(\mathbf{G})$, such as $\mathcal{K} \subset \mathcal{C}(\mathbf{G})$, is unique.[2]

Clearly, Theorem 1 does not provide any numerical conditions to check if a NMF is unique or not. This motivated the work of [26] from which it appears that uniqueness relies on the number of zero entries in both matrices $\mathbf{H}$ and $\mathbf{G}$. New results can also be found in the very recent paper [27]. Unfortunately, even if these approaches do give numerical rules to check if a NMF model is identifiable, they do not provide any

[2] The notation H ≥ 0 means that each entry of the matrix $\mathbf{H}$ is non-negative.

effective means to check directly from the data if the NMF is unique. This is the goal of the following proposition and corollary. However, before going any further, it is necessary to introduce the notion of *monomial matrix* and some related properties.

Definition 2. Monomial matrix

A positive matrix $\mathbf{T}$ of dimension (p, p) is called a monomial matrix if every row and every column of this matrix contains exactly one non-null element [28], that is

$$\forall i = 1, \ldots, p, \exists k; t_{ik} > 0 \text{ and } t_{jk} = 0 \forall j \neq i. \tag{17}$$

Property 1. *See* [29]

An arbitrary positive square matrix $\mathbf{T}$ *has a positive inverse matrix if and only if* $\mathbf{T}$ *is a monomial matrix. Then* $\mathbf{T}^{-1}$ *is also monomial.*

Property 2. *See* [29]

Each monomial matrix $\mathbf{T}$ *may be decomposed as* $\mathbf{T} = \Delta\mathbf{U}$*, where* Δ *is a positive diagonal matrix and* $\mathbf{U}$ *is a permutation matrix, that is, a monomial matrix whose non-zero elements are equal to 1. Such a transformation, when applied to* $\mathbf{G}$ *yields the scaling and ordering indeterminacies.*

We are now ready to formulate a sufficient condition for the uniqueness of NMF, from which we derive another sufficient condition, which can be applied directly to the data.

Proposition 1. Sufficient uniqueness condition

The decomposition of $\boldsymbol{W}$ *into* $\boldsymbol{H}$ *and* $\boldsymbol{G}$ *according to*

$$\mathbf{W} = \mathbf{H}\,\mathbf{G}^T, \quad \text{with} \quad \mathbf{G} \geq 0, \mathbf{H} \geq 0, \tag{18}$$

is unique if the following conditions are satisfied:

(B1) *There exists a submatrix of* $\mathbf{H}$ *of dimension* (K, K) *which is monomial.*
(B2) *There exists a submatrix of* $\mathbf{G}$ *of dimension* (K, K) *which is monomial.*

Proof. See Appendix A. □

From this result, we may immediately deduce the following corollary which gives a sufficient condition on $\mathbf{W}$ to admit a unique non-negative factorization.

Corollary 1. *The decomposition of* $\mathbf{W}$ *into* $\mathbf{H}$ *and* $\mathbf{G}$ *according to*

$$\mathbf{W} = \mathbf{H}\,\mathbf{G}^T \quad \text{with} \quad \mathbf{H} \geq 0, \mathbf{G} \geq 0, \tag{19}$$

is unique if the following condition is satisfied:

(C1) *After line and column permutations, the matrix of* $\mathbf{W}$ *can be written as:*

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \vdots & \mathbf{W}_{12} \\ \cdots & \cdots & \cdots \\ \mathbf{W}_{21} & \vdots & \mathbf{W}_{22} \end{bmatrix},$$

where $\mathbf{W}_{11}$ *is a non-singular diagonal matrix of dimension* (K, K).

The proof of this corollary is trivial using the fact that a non-negative monomial matrix can be factorized only as a product of two other non-negative monomial matrices of the same size.

4.2. Identifiability of the non-negative BCP model

Before addressing the BCP model identifiability, some uniqueness results of the CP decomposition must be presented. A key notion to the uniqueness of the CP decomposition is due to Kruskal [30], and relies on the concept of "Kruskal-rank" or simply k-rank. The k-rank of an $I \times K$ matrix $\mathbf{A}$, denoted by $k_{\mathbf{A}}$, is the maximum value $l \in \mathbb{N}$ such that every l columns of $\mathbf{A}$ are linearly independent. By definition, the k-rank of a matrix is less than or equal to its classical rank. Kruskal proved that [30]

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2K + 2 \tag{20}$$

is a sufficient condition for ensuring the uniqueness of the CP decomposition in Eq. (1). Furthermore, it becomes a necessary and sufficient condition in the case $K = 2$ or 3.

This condition no longer holds when one matrix (say $\mathbf{A}$) has a k-rank equal to 1, that is, when the matrix $\mathbf{A}$ has collinear columns. Unfortunately, this is what happens for the BCP model. In this case, we have to resort to the notion of partial uniqueness which means that only "part of the model" can be unique (see [31,32] for details). Restricting our attention to the BCP decomposition at hand, and based on the results of [32], the identifiability of $\mathbf{A}$ is ensured if:

$$r_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2K + 2, \tag{21}$$

where r_{A} is the classical rank of matrix $\mathbf{A}$. In particular, in the case considered, if $\mathbf{B}$ and $\mathbf{C}$ are full-column rank matrices, the identifiability of $\mathbf{A}$ requires only $r_{\mathrm{A}} \geq 2$. Some other identifiability results for $\mathbf{A}$ can be found in [32].

The key point is that, provided that $\mathbf{A}$ can be uniquely estimated from the data, theorem 3.1 of [32] ensures that the identifiability of the entire CP model can be assessed by checking the identifiability of several independent lower-rank CP models. Coming back to the BCP problem (Eq. (8)) at hand, this means that the uniqueness of the decomposition can be assessed by investigating the uniqueness of each bilinear sub-problem (Eq. (9)). In general, uniqueness cannot be guaranteed and the bilinear problem is unique up to rotational ambiguities. This is the *essential uniqueness* of BCP, as introduced by De Lathauwer in [2]. However, in the particular case of non-negative *pdf*s, the results of Section 3 can be used. This means that the BCP decomposition (Eq. (8)) is unique if the uniqueness of $\mathbf{A}$ is ensured and the non-negative bilinear factorizations (Eq. (9)) are unique. This result clearly shows the interest of non-negativity for the uniqueness of CP-like decompositions.

4.3. From theory to practice

The theoretical identifiability results presented in the previous subsections, all involve the identifiability of the NMF model. However, they may be a bit difficult to interpret for users who are not familiar with matrix factorization. In this subsection, we first give some graphical illustrations corresponding to practical situations for which the NMF identifiability is ensured or not. Fig. 3 gives 4 examples of data matrices consisting in the superposition of 3 rank-1 non-negative matrices. The left data matrix of the top row satisfies the condition of Corollary 1, meaning that its non-negative rank-3 decomposition is unique. That on the right of the top row gives an example where Donoho's condition [26] is fulfilled; thus, its non-negative rank-3 decomposition is also unique. On the contrary, the two bottom row matrices do not admit a unique solution since the necessary condition of [27] is not fulfilled (see also [33]).

This discussion on the identifiability of the BCP model naturally brings up the following question: is the uniqueness of the BCP decomposition really important in practical applications? This is actually a question which deserves to be discussed in detail since, for our problem of approximating the 2-D *pdf*s represented by the matrices $\mathbf{E}_k$, the rotational ambiguities do not matter. Indeed, regardless of the rotational ambiguities, the reconstructed $\mathbf{E}_k$ remains the same. Nevertheless, if the objective is the biological interpretation of the results, then uniqueness does matter because it turns out that *pdf*s having complex shapes result from mixtures of sub-populations behaving in a similar way. To

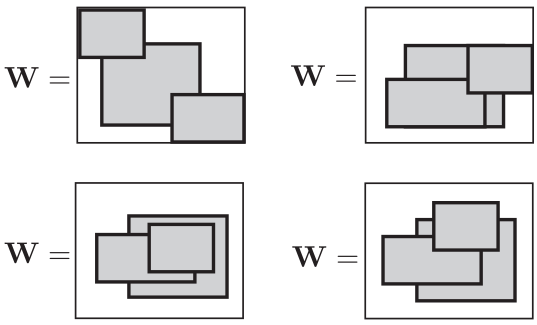


Fig. 3. Four data matrices admitting a NMF decomposition. The two matrices at the top of the figure admit a unique decomposition. Those of the bottom line do not.

illustrate the idea, let us consider a simple analogy: consider a population composed of children and adults having to run away from danger and observed at different locations with devices measuring the height and the weight (two parameters) of each individual. Roughly, 3 types of behavior are expected to be observed: a fast running sub-population (the adults), a slow running sub-population (children) and a last sub-population composed of adults holding children's hands and therefore running faster than children but slower than adults. From a decomposition point of view this is a single sub-population since they are all moving at the same speed. But if the uniqueness of the corresponding subblock is ensured, then the population of adults and children running together can be decomposed into adults only and children only, without any ambiguity. In such situations, the density $\mathbf{E}_k$ is multimodal. If its NMF decomposition is provably unique, thus each unimodal distribution can be uniquely associated with a sub-population: this is the property which is used to give a valid interpretation of the dependence of the mitochondrial membrane potential with respect to the cell cycle (see Subsection 5.3). In this context, the graphical illustration of the uniqueness of the NMF problem is mainly intended at verifying *a posteriori* if the decomposition of each density $\mathbf{E}_k$ (Eq. (9)) of the non-negative BCP decomposition can be considered as unique or not. Concerning the coupling of flow cytometry with fluorescence data, the uniqueness of the non-negative BCP implies the uniqueness of model (14). Otherwise, no general rule can be provided. However, it appears that gathering data is expected to improve the NMF model identifiability (see [14] for details).

4.4. Algorithms for data processing

The data processing method proposed in this paper consists of three steps. The first two steps deal with the processing of the flow cytometry data while the third addresses the coupling (fusion) of the flow cytometry data with the cell fluorescence spectra.

4.4.1. Estimation of the probability density functions

The flow cytometry data *pdf* is performed by computing the *N*-D histograms of the data which correspond to the hypercube Parzen window function. The developed algorithm requires to define the number of bins (one for each dimension) on which the histograms are calculated. Here, only 2-D histograms are considered. For all experiments, the number of bins was fixed to 50 along each dimension which results in *pdf* estimations showing a good tradeoff between resolution and accuracy. In fact, in practice the setting of this parameter value remains supervised and depends on the data at hand.

4.4.2. Non-negative BCP decomposition of the data

For the needs of this step, the non-negative BCP algorithm developed in the tensorlab toolbox [34] can be used. We also developed a procedure to estimate the ranks L_k and K of the decomposition. It first consists in performing a non-negative high-rank CP decomposition of the data. Thanks to the partial uniqueness properties of the CP model [32], the matrix containing the collinear loadings is ensured to be unique. In practice, because of the noise/error terms, the estimated loading may not be strictly collinear. The "most" collinear loadings, *i.e.* those presenting a correlation greater than a specified threshold (typically 0.9), are collapsed into a single loading by an averaging procedure. The K resulting averaged loadings are gathered into a full column rank matrix $\mathbf{A}$. The corresponding loadings of the other two modes are gathered to form matrices $\mathbf{E}_k$, $k = 1, \ldots, K$. This results in a non-negative BCP decomposition which is used as an initial solution for the non-negative BCP algorithm of [34]. Once the non-negative BCP decomposition is achieved, the corresponding *N*-D *pdf*s are normalized to have unit sum and the normalization factor is then transferred on the corresponding loadings, representing the responses of the K sub-populations to the physical parameter.

4.4.3. Coupling of the flow cytometry data with fluorescence spectra

There are at least two ways to couple the two datasets. A first approach is to estimate first the mixing matrix $\mathbf{A}$ using only the cytometry fluorescence data. Once matrix $\mathbf{A}$ is estimated, the source spectra are obtained from the bulk spectroscopy data by a least-squares procedure under non-negativity constraints. A second approach is to actually combine the data into a single data matrix following model (14) and then to decompose the large data matrix using a non-negative factorization algorithm (*e.g.* the Bayesian Positive Source Separation algorithm developed in [35]). In the next section, only the results corresponding to the first approach of step 3 are presented. In fact, no significant difference was observed between the two different approaches. The reason is that, in the considered example, the uniqueness of the corresponding rank-2 non-negative matrix factorization is provably unique. The interested reader is referred to [33], where a necessary and sufficient condition for having the uniqueness of the rank-2 NMF is provided.

The next section illustrates the effectiveness of the proposed approach on real flow cytometry and spectroscopy data. These data result from an experiment aiming at studying the response of the mitochondrial membrane potential of a particular cell line to a decoupling agent.

5. Analyzing mitochondrial membrane potential with JC-1

5.1. Mitochondrial membrane potential

Mitochondrial membrane potential ($\Delta\Psi m$) is an important indicator of the mitochondrial membrane integrity and mitochondrial efficacy through the coupling between oxidative phosphorylation and ATP synthesis. Indeed, $\Delta\Psi m$ is an indicator of cell viability since a drastic decrease of this potential is associated with cytochrome *c* release during apoptosis [36,37]. The membrane permeant dye JC-1[3] is largely used to monitor this mitochondrial parameter [38–40]. This lipophilic cationic dye enters cells and accumulates in mitochondria as monomers or oligomers (J-aggregates) that exhibit two different emission spectra often referred to as green and red respectively.[4] JC-1 monomers are associated with depolarized mitochondria whereas J-aggregates are formed when $\Delta\Psi m$ is high. Thus the mitochondrial membrane potential can be estimated by following the red/green ratio of the JC-1 dye to distinguish between mitochondria with high and low $\Delta\Psi m$. Qualitative and quantitative analysis of $\Delta\Psi m$ is usually performed by flow cytometry after an excitation of the probe at 488 nm with an argon laser. After excitation, JC-1 monomer fluorescence is measured in the FL1 channel 515–545 nm and the JC-1 aggregates are measured in the FL2 channel 564–606 nm. Because of the overlap of the two emission spectra, compensation is needed, around 20–30% of the green signal (FL1) has to be subtracted from the red signal (FL2). This compensation value needs to

[3] 5′,6,6′-tetrachloro-1,1′,3,3′-tetraethylbenzimidazolylcarbocyanine iodide.

[4] We will see, in the sequel, that this has to be understood as "mostly green" or "mostly red" since the emission spectra are not purely green or red.

be well calibrated in each experiment. Recently, Perelman et al. [41] have demonstrated that a new generation cytometer, equipped with another excitation laser, in particular at 405 nm, can be used for JC-1 measurements. These new lasers considerably reduce the overlap of the monomer fluorescence (green) with the J-aggregate fluorescence (red). These findings simplify the procedure since the fluorescence compensation can be avoided. However, it is worth mentioning that, regardless of the laser excitation, there will be situations where the peak overlapping cannot be completely avoided. It is also clear that changing the laser excitation is not always possible.

The present experimental study aims at validating the proposed data processing approach on real multicolor flow cytometry data corresponding to the response of a cell line to a widely used and well understood uncoupling agent (carbonyl cyanide p-chlorophenylhydrazone — CCCP, Sigma-Aldrich). In particular, we address the following questions:

What is the gain of coupling the flow cytometry together with the fluorescence spectroscopy? This is an original point since, so far, no previous work proposed to couple the two techniques.

Is it really necessary to perform the data compensation? We believe that this is a very important practical issue since, from our own experience, depending on the way the compensation is performed, the analysis results may be strongly affected. Avoiding this preprocessing step may certainly represent a major step in the development of quantitative analysis in flow cytometry.

Does the proposed approach bring new insights into the analysis and the understanding of cytometry data? Here, the stake is to evaluate, from a biological perspective the benefit of an accurate separation of the contributions of the different cell sub-populations.

5.2. Cell culture and data acquisition

Human ductal breast epithelial tumor cell line, T47D (from ATCC) was grown in RPMI 1640 medium supplemented with 10% fetal calf serum, 2 mM L-glutamine and 5 μg/ml Gentamicin at 37 °C in a humidified atmosphere containing 5% CO_2. The mitochondrial membrane potential-sensitive dye JC-1 was prepared as a stock solution in dimethyl sulfoxide (DMSO, Sigma-Aldrich) and stored at −20 °C. Before use, JC-1 stock solution was diluted 100× in assay buffer (delivered by manufacturer). Cells were stained following the manufacturer specifications. Briefly, 1 ml of each cell suspension was centrifuged at 400 *g* for 5 min at RT. The pellets were resuspended in 0.5 ml of JC-1 freshly diluted and containing various concentrations of carbonyl cyanide p-chlorophenylhydrazone (CCCP, Sigma-Aldrich). CCCP is an ionophore used to uncouple oxidative phosphorylation in mitochondria. It causes a mitochondrial proton leak, leading to a depolarization of the mitochondrial membrane. Thus, it is frequently used as a negative control in mitochondrial membrane potential measurements by flow cytometry. The concentration of CCCP for which it is well accepted that the cells are fully depolarized, ranges between 50 and 100 μM. Thus, we chose 6 CCCP concentrations varying between 0 and 100 μM ($[CCCP] = 0, 5, 10, 25, 50, 100$) to ensure that the whole CCCP response range is observed.

The samples were incubated for 15 min at 37 °C in a CO_2 incubator. At the end of the incubation period, each tube was washed twice with assay buffer and cells were resuspended in 0.3 ml of culture medium. Half of the cells were analyzed by flow cytometry (BD FACSCalibur) and the rest were analyzed by a fluorescence plate reader (Safas). Fig. 4 gives an example of experimental data. This is a sequence of six cytometry datasets, each one corresponding to a given CCCP concentration.

The first step of the processing consists in estimating the 2-D histograms. For each dimension, the number of bins is fixed to 50 resulting in a 50×50 data matrix. Then the six matrices are gathered into a 3-way array of dimension $(50 \times 50 \times 6)$. For this dataset, the compensation was fixed to have a maximum separation along the horizontal axis. Fig. 5 shows the corresponding sequence of fluorescence spectra. All the spectra are normalized to have a unit energy.

5.3. Results and discussion

Fig. 6 shows the results of the non-negative BCP decomposition corresponding to the dataset of Fig. 4 (see Subsection 3.3) and Fig. 7 shows the corresponding spectral source estimated by the joint analysis (see Subsection 3.4).

The number of block-component is expected to be equal to 2: a highly polarized cell sub-population whose response to CCCP is expected to decrease and a depolarized cell sub-population whose response is expected to increase with the CCCP concentration. The experimental parameters L_k of BCP were determined after successive trials. The first block rank is (3, 3, 1) while the second block rank is (2, 1, 1). The responses of the two sub-populations are in very good agreement with what was expected.

From a biological point of view, the top left-hand side plot on Fig. 6 represents the distribution of cells with a low mitochondrial membrane potential (referred to as "green" fluorescence) while the top right-hand side figure corresponds to the cells with high mitochondrial membrane potential (referred to as "red" fluorescence). The associated responses

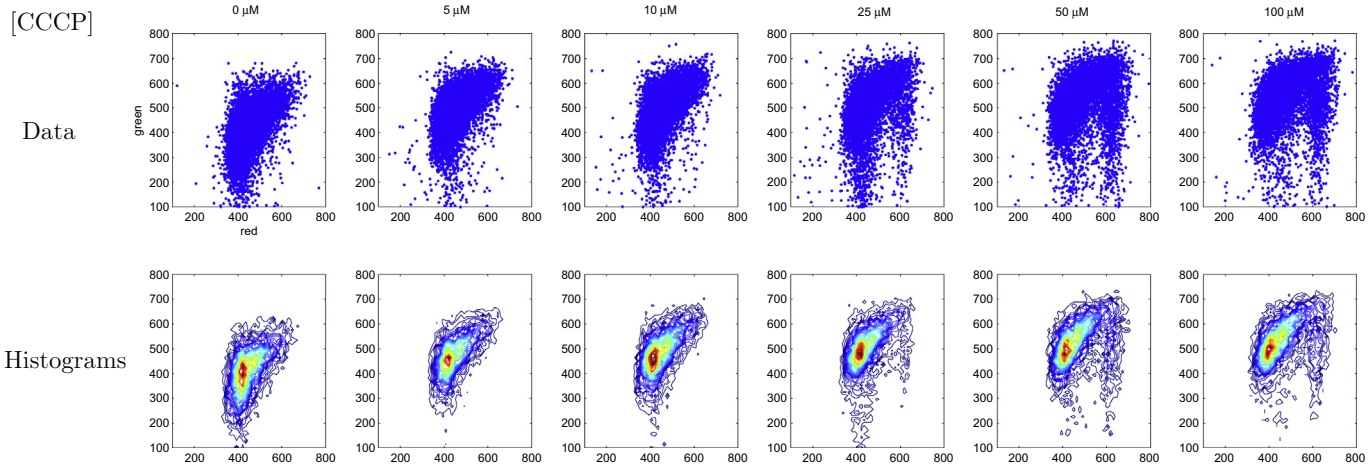


Fig. 4. A sequence of flow cytometry data showing the response of T47D cells to CCCP. The upper line figures correspond to the data and the lower line figures are the corresponding 2-D histograms.

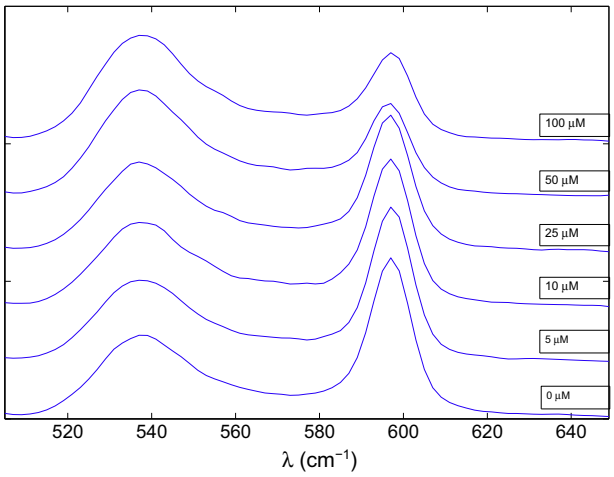


Fig. 5. A sequence of fluorescence spectra showing the response of T47D cells to CCCP. All the spectra are normalized to have unit area.

show that the low mitochondrial membrane potential sub-population increases with the concentration of CCCP while the high mitochondrial membrane potential sub-population decreases. In particular, it can be observed that the full cell population depolarization is reached after a concentration of 50 µM which is corresponding to the value generally accepted by practitioners.

As one can see on Fig. 7(a), the estimated spectra for the two sub-populations are highly correlated. From a signal processing point of view, separating these two spectral signatures using only the bulk spectroscopy data on Fig. 5, is a very difficult and challenging problem. The coupling of the two data modalities makes this separation possible without imposing additional constraints on the source parameters. From the biological point of view, the estimated spectra provide interesting insights into the understanding of the average behavior of the two sub-populations. Each sub-population, treated or not with CCCP, exhibits green and red fluorescence corresponding to JC-1 monomers and JC-1 aggregates, respectively. Cell sub-population with high (respectively low) mitochondrial membrane potential is more red than green (respectively more green than red). The estimated spectra also show that, for the two sub-populations, there are no other discriminant wavelengths in their emitted fluorescence light. This could represent an interesting tool for the efficient choice of the adequate wavelengths in flow cytometry.

The joint analysis of cytometry and spectroscopy data yields a complete characterization of the two types of cell sub-populations: distribution, fluorescence spectra, and response to CCCP concentration. The cytometry characterizes the cell sub-population distribution with respect to the fluorescence intensity while the spectroscopy provides information on the sub-population distributions with respect to the wavelength.

To evaluate the reproducibility of the experiments, we repeated it three times. The rank of the BCP decomposition was fixed to the same values as in the first experiment. The results of the non-negative BCP of the other two datasets are given in Fig. 8 and it appears that the reproducibility of the results is quite good: not only the CCCP responses are quite similar but also the cell sub-population distributions are quite similar.

The next experiment objective was twofold.

When using standard cytometry data analysis tools (as provided by the manufacturer), the red and green fluorescence analyses of JC-1 always require a user defined compensation procedure which may strongly affect the quantitative interpretation of the data. As mentioned in [42], compensation is in fact a linear transformation of the data and therefore, in our experiments, it is not supposed to affect the response to CCCP. In the following experiments, the data have been acquired without and with compensation.

Having a closer look at the *pdf*s of the cell sub-populations revealed that they were not unimodal, which was rather surprising from a biological point of view, since a single cell line is considered. Our conjecture was that the cell asynchronicity was responsible of this multimodal distribution, each mode corresponding to a particular mitochondrial

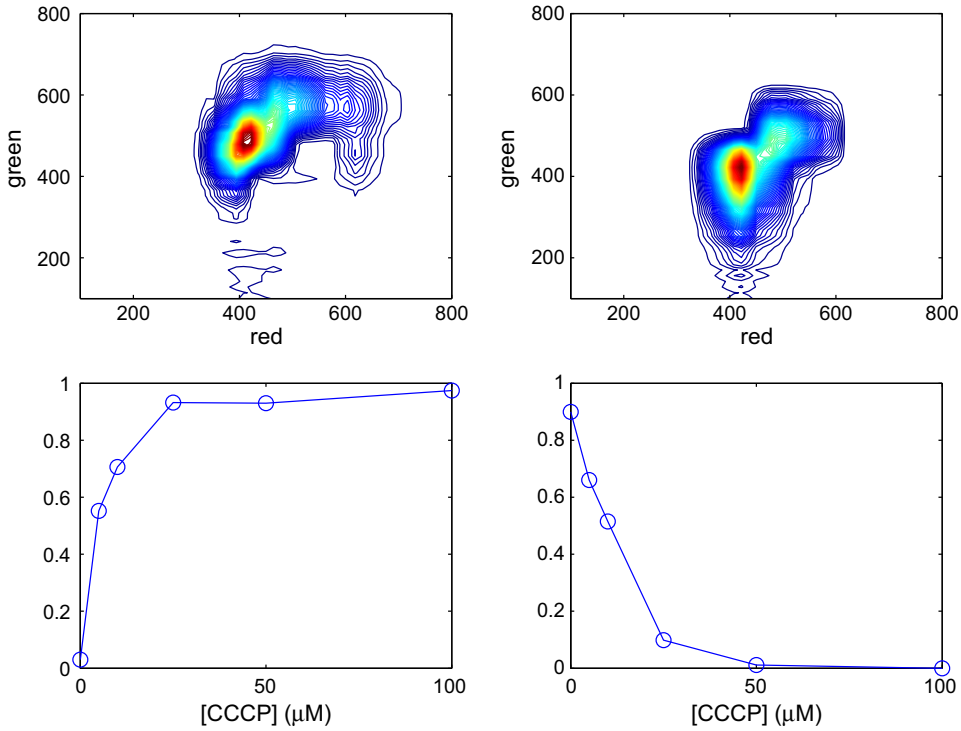


Fig. 6. Non-negative BCP decomposition results of flow cytometry dataset corresponding to the first experiment.

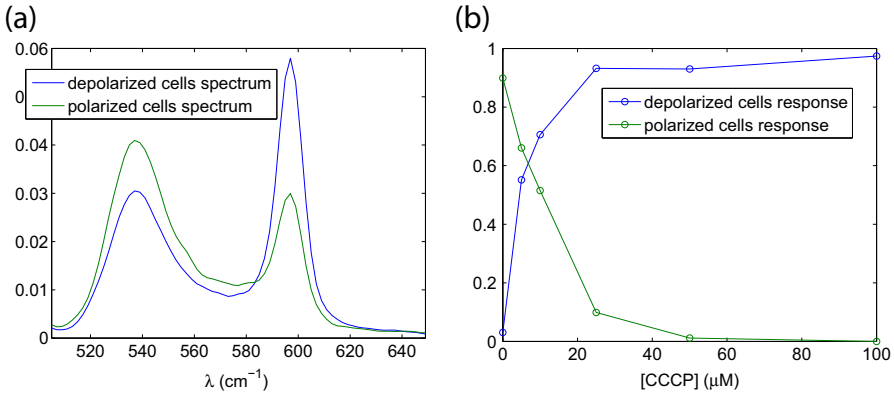


Fig. 7. Estimation of the pure fluorescence spectra (a) corresponding to the first experiment. The mixing coefficients (b) are those obtained from the non-negative BCP decomposition.

membrane potential associated with a specific cell cycle phase. Indeed, the dependance of the mitochondrial membrane potential of the cell cycle phase was already mentioned in [43–45]. They showed that there is a global increase of the mitochondrial membrane potential of cells in the G2 phase of the cell cycle as compared to cells in the G1 phase. Thus the next experiment consisted in studying the distributions of cell sub-populations before and after synchronization in the G1 phase.

This results in 4 different datasets referred to respectively as: (a) no synchronization and no compensation, (b) synchronization and no compensation, (c) no synchronization and compensation and (d) synchronization and compensation.

To synchronize cells at the G1 phase of the cell cycle, cells were exposed to 2 mM thymidine (Sigma-Aldrich) for 48 h. Then, synchronized cells were collected, counted (TC10 Automated Cell Counter, Bio-Rad) and adjusted to the density of 300,000 cells/ml for analysis of mitochondrial membrane potential. The cell cycle synchronization was monitored by measurements of the DNA content per cell. The rate of DNA was estimated by propidium iodide staining. Cells were fixed and permeabilized by 70% icecold ethanol and stored at −20 °C for at least 24 h. They were washed with PBS (Phosphate Buffered Saline) and resuspended in 1 ml of DNA staining solution (2.5 μg/ml propidium iodide, and 0.5 mg/ml RNase A in PBS). The labeling of the fluorescent probe was measured by flow cytometry (Becton, Dickinson, FACSCalibur). Cell cycle synchronization was verified by flow cytometric analysis of DNA content. Representative histograms are shown in Fig. 9. Treatment with thymidine results in a G1/S-phase arrest in contrast to untreated cells.

The results of the data processing are shown in Fig. 10(a–d). On the one hand, comparing Fig. 10(a) and (c) as well as (b) and (d), it can be observed that the compensation does not affect much the shape of the response to CCCP. In other words, by using the proposed approach, compensation is no longer necessary. On the other hand, it appears that the synchronization modifies the shape of the cell population distributions. Indeed, the not-synchronized low mitochondrial membrane potential sub-population includes two main modes, one centered on (350, 300) and a second centered on (450, 400). This second mode significantly decreases after synchronization in the G1 phase, resulting in a shift toward the low value of the red and thus increasing the relative importance of the green. This is much more visible on the distribution of the high mitochondrial membrane potential subpopulation. This can be attributed to the fact that cells in the G1 phase have a lower mitochondrial membrane potential that those in the S and G2 phases which is in accordance with the literature. Also, looking at the response to CCCP, it seems to indicate that the dynamic of the responses of the low and high mitochondrial membrane potential subpopulations is stronger after synchronization.

6. Conclusions

In this paper, we proposed a novel flow cytometry data analysis methodology, based on a non-negative block-CANDECOMP/PARAFAC model of the data, and highlighted the link between bulk spectroscopy and flow cytometry data. A sufficient condition allowing the guarantee of the uniqueness of data decomposition was also derived. The joint processing of the two data modalities results in an effective tool that reveals the full analysis potential of flow cytometry; the approach was validated on real data produced using the human ductal breast epithelial tumor cell line T47D for which the mitochondrial membrane potential was estimated.

The main underlying idea is to exploit the different behaviors of cell sub-populations with respect to a physical parameter (the CCCP concentration in this paper). Unlike the classically employed clustering methods, strongly relying on the user expertise, our method requires only the knowledge of the number of sub-populations to be extracted from the data, and is very little sensitive to the overlapping of cell sub-population distributions. Thus, similarly to Perelman et al. [41], our method also yields an effective way to avoid compensation, but without requiring the use of a different laser excitation adapted to the mitochondrial membrane potential measurement with a JC-1 probe. The ability of the method to separate overlapping densities has revealed two sub-populations within a single cell line in both high and low mitochondrial membrane potential cell populations. This unexpected but insightful side result has been attributed to cells being in different cell cycle phases, having slightly different mitochondrial membrane potentials. The price to pay for these interesting features is an increased complexity of the experiments generating the data. However, the joint data processing and experiment design is a promising research direction in which biologists and data analysts may develop very fruitful collaborations.

This work raises a number of questions which deserve to be investigated in forthcoming work. In this work, the cell sorting is addressed as a non-negative source separation problem; thus no decision is required since only the response to CCCP is sought. The proposed approach performs significantly better that a classification based approach in particular when the probability density function of the different populations strongly overlaps: even if it is possible to design optimal decision rules minimizing the probability of error, highly overlapping densities necessarily result in a high classification error. However it is clear that a decision is sometimes required. For example, in the case of infrequent cell population (as it is the case of stem cells in tissue), it is necessary to perform a “physical” cell sorting (generally referred to as gating) to increase the proportion of interesting cells in the population. A quite relevant problem is concerning the possibility to use the proposed approach to design optimal classification rules yielding improved gating

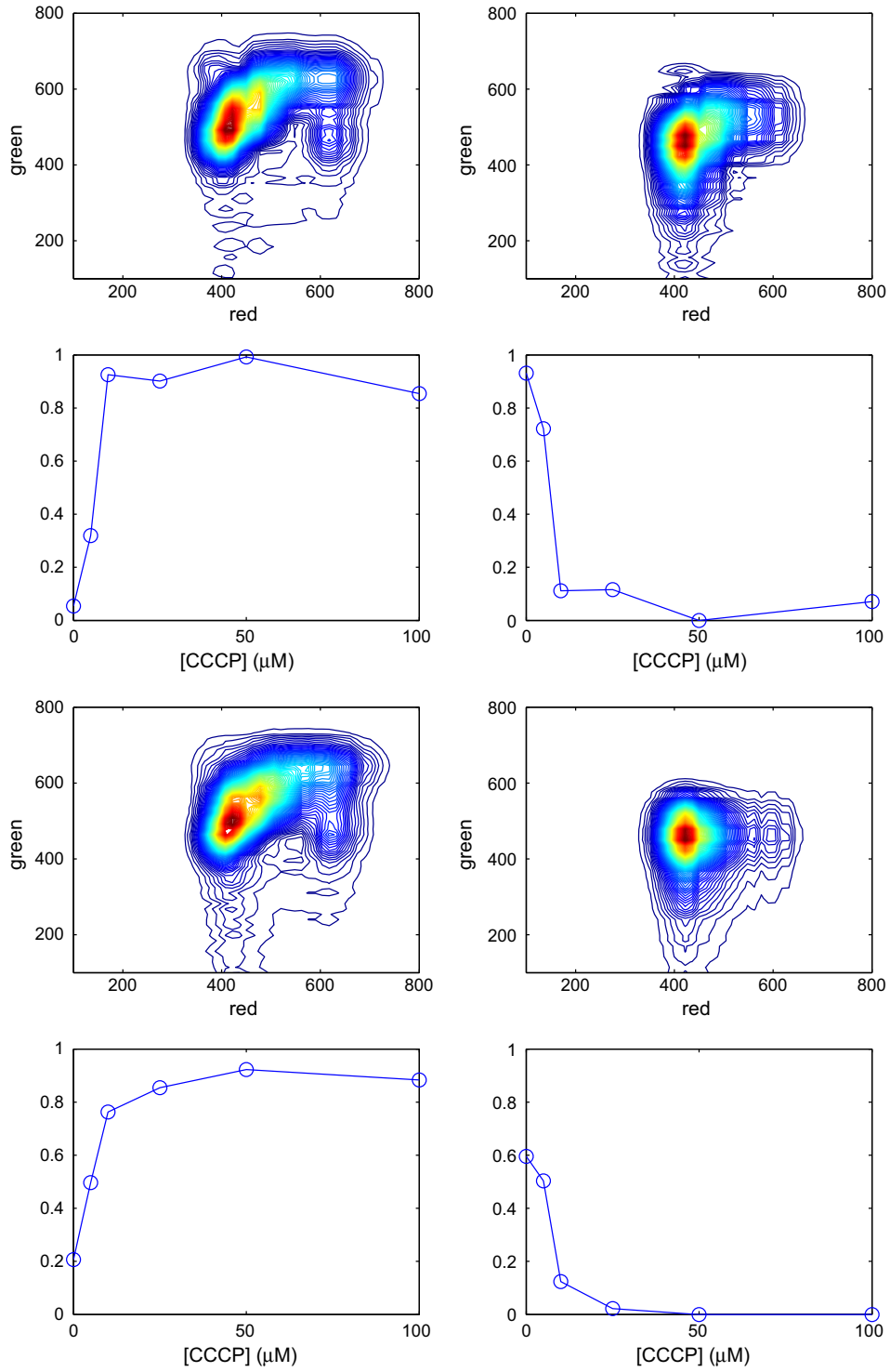


Fig. 8. Non-negative BCP decomposition of the two datasets obtained by repeating the same experiment as for the dataset on Fig. 6.

strategies. The problem complexity clearly increases with the number of classes (*i.e.* populations).

The proposed approach allows us to efficiently model the multivariate density function[5] and a very appealing feature of the proposed approach relates to model identifiability which becomes less restrictive as the dimension increases. However, this is only part of the problem and a challenging question is: how do we handle the curse of dimensionality? A possible way to tackle it is to perform dimensionality reduction (*e.g.* PCA, projection pursuit, Informative component analysis). In our future work, we will address this problem using a sparse approximation approach. Not only, this is expected to yield effective high dimensional density estimation but this is also expected to be helpful in designing efficient algorithms to perform high-order (*i.e.* dimension) non-negative Block CP decompositions.

[5] To give some figures, a N-dimensional pdf represented by a (smoothed) histogram having M bins in each dimension corresponds to an array having M^N entries. A rank K CP model of the pdf allows the representation of the array with $KMN \ll M^N$.

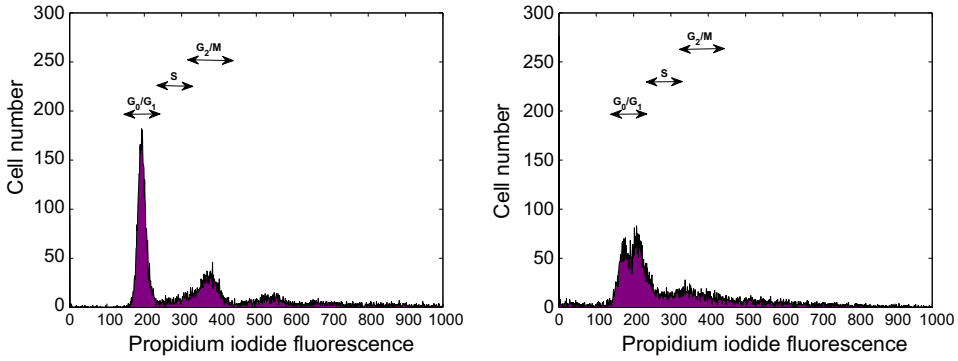
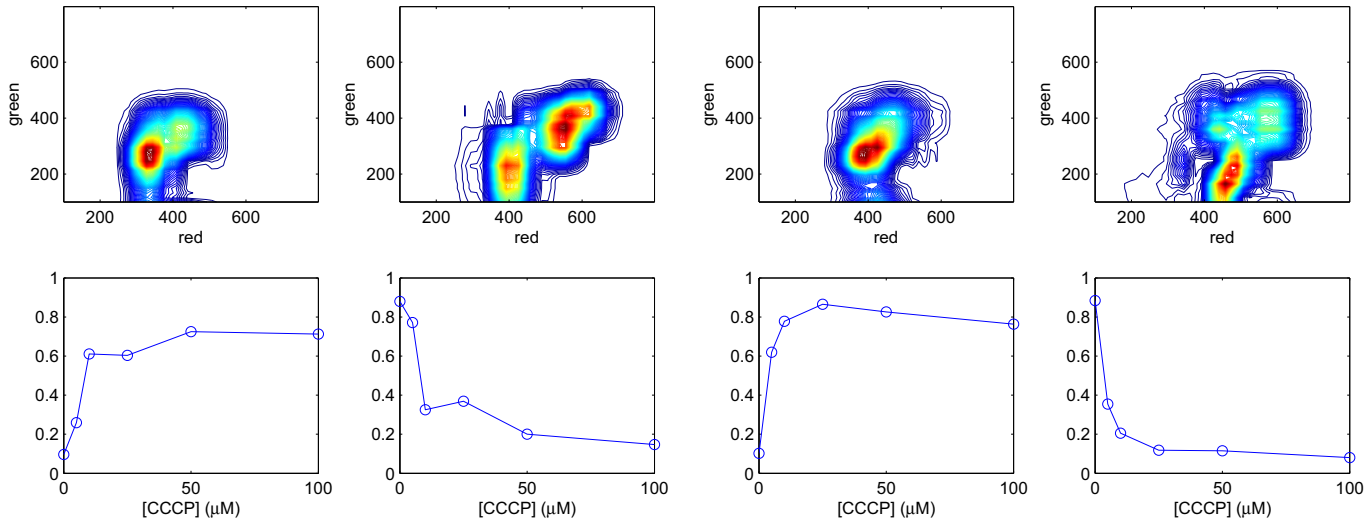
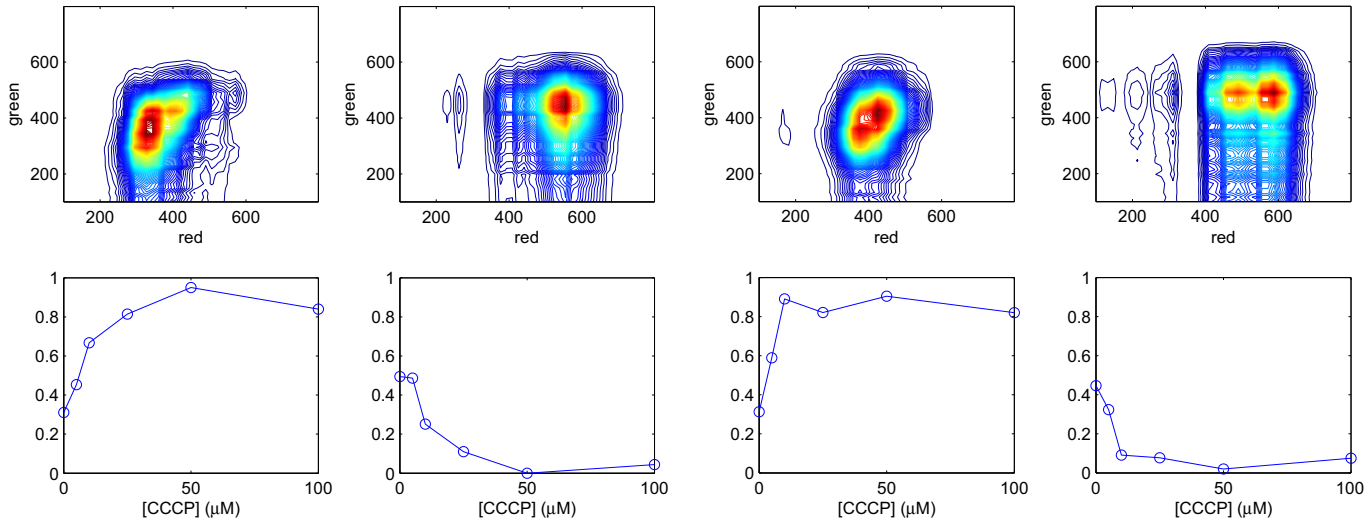


Fig. 9. Cell cycle distribution of T47D untreated (left) or synchronized (right) with thymidine (2 mM) for 2 days.



(a) no synchronization and no compensation

(b) synchronization and no compensation



(c) no synchronization and compensation

(d) synchronization and compensation

Fig. 10. Non-negative block decomposition results of datasets after cell synchronization in G1 phase or not and compensation or not.

Appendix A. Proof of Proposition 1

Suppose that conditions (B1) and (B2) are satisfied. After a possible permutation of its columns, the matrix $\mathbf{G}$ can be rewritten as:

$$\mathbf{G} = \begin{bmatrix} g_{11} & 0 & 0 & g_{1(K+1)} & \cdots \\ 0 & \ddots & 0 & \vdots & \\ 0 & 0 & g_{KK} & g_{K(K+1)} & \cdots \end{bmatrix}^T.$$

Similarly, after a possible permutation of its rows, the matrix $\mathbf{H}$ can be re-written as:

$$\mathbf{H} = \begin{bmatrix} h_{11} & & 0 \\ & \ddots & \\ 0 & & h_{KK} \\ h_{(K+1)1} & \cdots & h_{(K+1)K} \\ \vdots & & \vdots \end{bmatrix}.$$

Let us consider the regular $(K \times K)$ matrix $\mathbf{T} = [t_{ij}]$ whose inverse is noted as $\mathbf{T}^{-1} = [t_{ij}^{\#}]$. We have

$$\mathbf{TG}^T = \begin{bmatrix} t_{11}g_{11} & \cdots & t_{1K}g_{KK} & \cdots \\ \vdots & & \vdots & \\ t_{K1}g_{11} & \cdots & t_{KK}g_{KK} & \cdots \end{bmatrix}.$$

$$\mathbf{HT}^{-1} = \begin{bmatrix} h_{11}t_{11}^{\#} & \cdots & h_{11}t_{1K}^{\#} \\ \vdots & & \vdots \\ h_{KK}t_{1K}^{\#} & \cdots & h_{KK}t_{KK}^{\#} \\ \vdots & & \vdots \end{bmatrix}.$$

from which it turns out that $\mathbf{TG}^T \geq 0$ and $\mathbf{HT}^{-1} \geq 0$ if and only if $\mathbf{T} > 0$ and $\mathbf{T}^{-1} > 0$. From Property 1, this is equivalent to say that $\mathbf{T}$ is a monomial matrix and according to Property 2, the solution is unique up to scaling and ordering indeterminacies.

References

[1] L. De Lathauwer, Decompositions of a higher-order tensor in block terms — part I: lemmas for partitioned matrices, SIAM J. Matrix Anal. Appl. 30 (3) (2008) 1022–1032.
[2] L. De Lathauwer, Decompositions of a higher-order tensor in block terms — part II: definitions and uniqueness, SIAM J. Matrix Anal. Appl. 30 (3) (2008) 1033–1066.
[3] C.E. Pedreira, E.S. Costa, Q. Lecrevisse, J.J. van Dongen, A. Orfao, Overview of clinical flow cytometry data analysis: recent advances and future challenges, Trends Biotechnol. 31 (7) (2013) 415–425.
[4] M.F. Wilkins, S.A. Hardy, L. Boddy, C.W. Morris, Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data, Cytometry 44 (3) (2001) 210–217.
[5] Q.T. Zeng, J.P. Pratt, J. Pak, D. Ravnic, H. Huss, S.J. Mentzer, Feature-guided clustering of multi-dimensional flow cytometry datasets, J. Biomed. Inform. 40 (3) (2007) 325–331.
[6] J. Frelinger, T.B. Kepler, C. Chan, Flow: statistics, visualization and informatics for flow cytometry, Source Code Biol. Med. 3 (1) (2008) 10.
[7] K. Lo, R.R. Brinkman, R. Gottardo, Automated gating of flow cytometry data via robust model-based clustering, Cytom. Part A 73A (4) (2008) 321–332.
[8] C.E. Pedreira, E.S. Costa, M.E. Arroyo, J. Almeida, A. Orfao, A multidimensional classification approach for the automated analysis of flow cytometry data, IEEE Trans. Biomed. Eng. 55 (3) (2008) 1155–1162.
[9] C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, T.B. Kepler, Statistical mixture modeling for cell subtype identification in flow cytometry, Cytom. Part A 73A (8) (2008) 693–701.
[10] M.J. Boedigheimer, J. Ferbas, Mixture modeling approach to flow cytometry data, Cytom. Part A 73A (5) (2008) 421–429.
[11] N. Aghaeepour, G. F. N., H. Hoos, T.R. Mosmann, R. Brinkman, R. Gottardo, D. Consortium, Critical assessment of automated flow cytometry data analysis techniques, Nat. Methods 10 (2013) 228–238.
[12] K.M. Carter, R. Raich, W.G. Finn, A.O. Hero, Information preserving component analysis: data projections for flow cytometry analysis, IEEE J. Sel. Top. Signal Proc. 3 (1) (2009) 148–158.
[13] K. Carter, R. Raich, W. Finn, A. Hero, Information-geometric dimensionality reduction, IEEE Signal Process. Mag. 28 (2) (2011) 89–99.
[14] S. Miron, M. Dossot, C. Carteret, S. Margueron, D. Brie, Joint processing of the parallel and crossed polarized Raman spectra and uniqueness in blind nonnegative source separation, Chemometr. Intell. Lab. 105 (1) (2011) 7–18.
[15] E. Acar, M.A. Rasmussen, F. Savorani, T. Næs, R. Bro, Understanding data fusion within the framework of coupled matrix and tensor factorizations, Chemometr. Intell. Lab. 129 (1) (2013) 53–63.
[16] M. Sorensen, L. De Lathauwer, Coupled tensor decompositions for applications in array signal processing, Computational Advances in Multi-sensor Adaptive Processing (CAMSAP), 2013 IEEE 5th International Workshop on, IEEE, 2013, pp. 228–231.
[17] J.D. Carroll, J.-J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart–Young" decomposition, Psychometrika 35 (3) (1970) 283–319.
[18] R.A. Harshman, Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multimodal factor analysis, UCLA Working Papers in Phonetics, 16, 1970, pp. 1–84.
[19] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2001.
[20] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, 1992.
[21] D. Brie, S. Miron, F. Caland, C. Mustin, An uniqueness condition for the 4-way CANDECOMP/PARAFAC model with collinear loadings in three modes, International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011, 2011.
[22] L.-H. Lim, P. Comon, Nonnegative approximations of nonnegative tensors, J. Chemometr. 23 (7–8) (2009) 432–441.
[23] D. Lowd, P. Domingos, Naive Bayes models for probability estimation, Proceedings of the 22nd International Conference on Machine Learning, ACM, 2005, pp. 529–536.
[24] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[25] J.C. Chen, Nonnegative rank factorisation of nonnegative matrices, Linear Algebra Appl. 62 (1984) 207–217.
[26] D. Donoho, V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts? Advances in Neural Information Processing Systems, 16, MIT Press, Cambridge, United States, 2003.
[27] K. Huang, N. Sidiropoulos, A. Swami, Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition, IEEE Trans. Signal Proc. 62 (1) (2014) 211–224.
[28] J. Van den Hof, Realization of positive linear systems, Linear Algebra Appl. 256 (1997) 287–308.
[29] R. Berman, B. Plemmons, Nonnegative Matrices in the Mathematical Sciences, Siam, 1994.
[30] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear Algebra Appl. 18 (2) (1977) 95–138.
[31] J.M.F. ten Berge, Partial uniqueness in CANDECOMP/PARAFAC, J. Chemometr. 18 (1) (2004) 12–16.
[32] X. Guo, S. Miron, D. Brie, A. Stegeman, Uni-mode and partial uniqueness conditions for CANDECOMP/PARAFAC of three-way arrays with linearly dependent loadings, SIAM J. Matrix Anal. Appl. 33 (1) (2012) 111–129.
[33] S. Moussaoui, D. Brie, J. Idier, Non-negative source separation: range of admissible solutions and conditions for the uniqueness of the solution, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2005, 2005.
[34] L. Sorber, M. Van Barel, L. De Lathauwer, Tensorlab v1.0, http://esat.kuleuven.be/sista/tensorlab/ 2013 (Available, online, February 2013).
[35] S. Moussaoui, D. Brie, A. Mohammad-Djafari, C. Carteret, Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling, IEEE Trans. Signal Proc. 54 (11) (2006) 4133–4145.
[36] N. Zamzami, T. Hirsch, B. Dallaporta, P. Petit, G. Kroemer, Mitochondrial implication in accidental and programmed cell death: apoptosis and necrosis, J. Bioenerg. Biomembr. 29 (2) (1997) 185–193.
[37] A. Cossarizza, S. Salvioli, Analysis of mitochondria during cell death, Methods Cell Biol. 63 (2001) 467–486.
[38] M. Reers, T. Smith, L. Chen, J-aggregate formation of a carbocyanine as a quantitative fluorescent indicator of membrane potential, Biochemistry 30 (18) (1991) 4480–4486.
[39] S. Salvioli, A. Ardizzoni, C. Franceschi, A. Cossarizza, Jc-1, but not dioc6(3) or rhodamine 123, is a reliable fluorescent probe to assess delta psi changes in intact cells: implications for studies on mitochondrial functionality during apoptosis, FEBS Lett. 411 (1) (1997) 77–82.
[40] A. Mathur, Y. Hong, B. Kemp, A. B. AA, J.D. Erusalimsky, Evaluation of fluorescent dyes for the detection of mitochondrial membrane potential changes in cultured cardiomyocytes, Cardiovasc. Res. 46 (1) (2000) 126–138.
[41] A. Perelman, C. Wachtel, M. Cohen, S. Haupt, H. Shapiro, A. Tzur, Jc-1: alternative excitation wavelengths facilitate mitochondrial membrane potential cytometry, Cell Death Dis. (2012), http://dx.doi.org/10.1038/cddis.2012.171.
[42] B. Rajwa, Just compensation? Cytom. Part A 79 (12) (2011) 973–974.
[43] M. Martínez-Diez, G. Santamaría, Á.D. Ortega, J.M. Cuezva, Biogenesis and dynamics of mitochondria during the cell cycle: significance of 3′UTRs, PLoS ONE 1 (1) (2006) e107.
[44] S.M. Schieke, J.P. McCoy Jr., T. Finkel, Coordination of mitochondrial bioenergetics with G1 phase cell cycle progression, Cell Cycle 7 (12) (2008) 1782–1787.
[45] W. Xiong, Y. Jiao, W. Huang, M. Ma, M. Yu, Q. Cui, D. Tan, Regulation of the cell cycle via mitochondrial gene expression and energy metabolism in HeLa cells, Acta Biochim. Biophys. Sin. 44 (4) (2012) 347–358.