Algorithms for nonnegative matrix factorization with the beta-divergence

Cédric Févotte, Jérôme Idier

Algorithms for nonnegative matrix factorization with the β-divergence

Cédric Févotte[1] and Jérôme Idier[2]

[1] CNRS LTCI; Télécom ParisTech, France
Email: fevotte@telecom-paristech.fr

[2] CNRS IRCCyN; École Centrale de Nantes, France
Email: jerome.idier@irccyn.ec-nantes.fr

March 7, 2011

to appear in *Neural Computation*

Abstract

This paper describes algorithms for nonnegative matrix factorization (NMF) with the β-divergence (β-NMF). The β-divergence is a family of cost functions parametrized by a single shape parameter β that takes the Euclidean distance, the Kullback-Leibler divergence and the Itakura-Saito divergence as special cases ($\beta = 2, 1, 0$ respectively). The proposed algorithms are based on a surrogate *auxiliary function* (a local majorization of the criterion function). We first describe a *majorization-minimization* (MM) algorithm that leads to multiplicative updates, which differ from standard heuristic multiplicative updates by a β-dependent power exponent. The monotonicity of the heuristic algorithm can however be proven for $\beta \in (0, 1)$ using the proposed auxiliary function. Then we introduce the concept of *majorization-equalization* (ME) algorithm which produces updates that move along constant level sets of the auxiliary function and lead to larger steps than MM. Simulations on synthetic and real data illustrate the faster convergence of the ME approach. The paper also describes how the proposed algorithms can be adapted to two common variants of NMF: penalized NMF (i.e., when a penalty function of the factors is added to the criterion function) and convex-NMF (when the dictionary is assumed to belong to a known subspace).

Keywords: Nonnegative matrix factorization (NMF), β-divergence, multiplicative algorithms, majorization-minimization (MM), majorization-equalization (ME).

1 Introduction

Given a data matrix $\mathbf{V}$ of dimensions $F \times N$ with nonnegative entries, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $F\,K + K\,N \ll F\,N$, hence reducing the data dimension. The factorization is in general only approximate, so that the terms "approximate nonnegative matrix factorization" or "nonnegative matrix approximation" also appear in the literature. NMF has been used for various problems in diverse fields. To cite a few, let us mention the problems of learning parts of faces and semantic features of text (Lee and Seung, 1999), polyphonic music transcription (Smaragdis and Brown, 2003), object characterization by reflectance spectra analysis (Berry et al., 2007), portfolio diversification (Drakakis et al., 2007), DNA gene expression analysis (Brunet et al., 2004; Gao and Church, 2005), clustering of protein interactions (Greene et al., 2008), image denoising and inpainting (Mairal et al., 2010), etc. The factorization (1) is usually sought after through the minimization problem

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{W}\mathbf{H}) \text{ subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0 \tag{2}$$

where the notation $\mathbf{A} \geq 0$ expresses nonnegativity of the entries of matrix $\mathbf{A}$ (and not semidefinite positiveness), and where $D(\mathbf{V}|\mathbf{W}\mathbf{H})$ is a separable measure of fit such that

$$D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{W}\mathbf{H}]_{fn}) \tag{3}$$

where $d(x|y)$ is a scalar cost function. What we intend by "cost function" is a positive function of $y \in \mathbb{R}_+$ given $x \in \mathbb{R}_+$, with a single minimum for $x = y$.

A popular cost function in NMF is the β-divergence $d_\beta(x|y)$ of Basu et al. (1998); Eguchi and Kano (2001); Cichocki and Amari (2010), defined rigorously in Section 2.1. In essence, it is a parameterized cost function with a single parameter β, which takes the Euclidean distance, the generalized Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) divergence as special cases ($\beta = 2, 1$ and 0, respectively). NMF with the β-divergence has been widely used in music signal processing in particular, for transcription and source separation (O'Grady, 2007; O'Grady and Pearlmutter, 2008; FitzGerald et al., 2009; Bertin et al., 2009; Févotte et al., 2009; Vincent et al., 2010; Dessein et al., 2010; Hennequin et al., 2010). In these work the nonnegative data matrix $\mathbf{V}$ is a spectrogram which is decomposed into elementary spectra with NMF. The parameter β can be tuned so as to optimize transcription or separation accuracy on training data. While popular in music signal processing, NMF with the β-divergence (shortened as "β-NMF" in the rest of the paper) can be of interest to any field: the parameter β essentially controls the assumed statistics of the observation noise and can either be fixed or learnt from training data or by cross-validation. As noted by Févotte and Cemgil (2009), the values $\beta = 2, 1, 0$ respectively underly Gaussian additive, Poisson and multiplicative Gamma observation noise. The β-divergence offers a continuum of noise statistics that interpolate between these three specific cases, see (Basu et al., 1998; Eguchi and Kano, 2001; Minami and Eguchi, 2002; Cichocki and Amari, 2010).

The standard β-NMF algorithm used in the above-mentioned papers is presented as a gradient-descent algorithm where the step size is set adaptively and chosen such that the updates are multiplicative, as originally described by Cichocki et al. (2006). The same algorithm can be derived from the following heuristic, proposed by Févotte et al. (2009). Let θ be a coefficient of $\mathbf{W}$ or $\mathbf{H}$. As will be seen later, when using the β-divergence the derivative $\nabla_\theta D(\theta)$ of the criterion $D(\mathbf{V}|\mathbf{WH})$ with respect to (w.r.t) θ can be expressed as the difference of two nonnegative functions such that $\nabla_\theta D(\theta) = \nabla_\theta^+ D(\theta) - \nabla_\theta^- D(\theta)$. Then, a heuristic multiplicative algorithm simply writes

$$\theta \leftarrow \theta . \frac{\nabla_\theta^- D(\theta)}{\nabla_\theta^+ D(\theta)} \tag{4}$$

which ensures nonnegativity of the parameter updates, provided initialization with a nonnegative value. It produces a descent algorithm in the sense that θ is updated towards left (resp. right) when the gradient is positive (resp. negative). A fixed point $\theta^\star$ of the algorithm implies either $\nabla_\theta D(\theta^\star) = 0$ or $\theta^\star = 0$. Monotonicity of this algorithm has been proven by Kompass (2007) for the specific range of values of β for which the divergence $d_\beta(x|y)$ is convex w.r.t y (i.e., $\beta \in [1, 2]$, see Section 2.1). The proof is based on a *majorization-minimization* (MM) procedure: an *auxiliary function* is built and iteratively minimized for each column of $\mathbf{H}$ (given $\mathbf{W}$) and each row of $\mathbf{W}$ (given $\mathbf{H}$). The auxiliary function is built using Jensen's inequality, thanks to convexity of the cost for $\beta \in [1, 2]$. However, it was observed in practice that the multiplicative algorithm (4) is still monotone (i.e., decreases the criterion function at each iteration) for values of β out of the "convexity" interval $[1, 2]$, though no proof is to avail.

This paper studies three descent algorithms for β-NMF, based on an auxiliary function that unifies existing auxiliary functions for the Euclidean distance and KL divergence (De Pierro, 1993; Lee and Seung, 2001), the "generalized divergence" of Kompass (2007) and the IS divergence (Cao et al., 1999). This auxiliary function was also recently proposed independently by Nakano et al. (2010). The construction of the auxiliary function relies on the decomposition of the criterion function into its convex and concave parts, following the approach of Cao et al. (1999) for the IS divergence. An auxiliary function to the convex part is constructed using Jensen's inequality while the concave part is locally majorized by its tangent. It is shown that MM algorithms based on the latter auxiliary function yield multiplicative updates that coincide with the heuristic described by Eq. (4) for $\beta \in [1, 2]$, but differ from a β-dependent power exponent when $\beta \notin [1, 2]$, a result also obtained by Nakano et al. (2010). Additionally, we show that the monotonicity of the heuristic algorithm can however be proven for $\beta \in (0, 1)$, using the proposed auxiliary function (it is shown to produce a descent algorithm though it does not fully minimize the auxiliary function). Then we introduce the concept of *maximization-equalization* (ME) algorithm which produces updates that move along constant level sets of the auxiliary function and leads to larger steps than MM. This is akin to *overrelaxation* and is shown experimentally to produce faster convergence. Finally we show how the described MM, ME and heuristic algorithms can be adapted to two common variants of NMF: penalized NMF (i.e., when a penalty function of $\mathbf{W}$ or $\mathbf{H}$ is added to the criterion function) and "convex"-NMF (when the dictionary is assumed to belong to a known subspace, as proposed by Ding et al. (2010)).

	$\overset{\smile}{d}(x\|y)$	$\overset{\smile}{d}'(x\|y)$	$\overset{\frown}{d}(x\|y)$	$\overset{\frown}{d}'(x\|y)$	$\bar{d}(x)$
$\beta < 1$ and $\beta \neq 0$	$-\frac{1}{\beta-1} x\, y^{\beta-1}$	$-x\, y^{\beta-2}$	$\frac{1}{\beta} y^{\beta}$	$y^{\beta-1}$	$\frac{1}{\beta(\beta-1)} x^{\beta}$
$\beta = 0$	$x\, y^{-1}$	$-x\, y^{-2}$	$\log y$	y^{-1}	$x(\log x - 1)$
$1 \leq \beta \leq 2$	$d_\beta(x\|y)$	$d'_\beta(x\|y)$	0	0	0
$\beta > 2$	$\frac{1}{\beta} y^{\beta}$	$y^{\beta-1}$	$-\frac{1}{\beta-1} x\, y^{\beta-1}$	$-x\, y^{\beta-2}$	$\frac{1}{\beta(\beta-1)} x^{\beta}$

Table 1: Example of differentiable convex-concave-constant decomposition of the β-divergence under the form (8).

The paper is organized as follows. Section 2 defines and discusses the β-divergence, and then exposes in details the optimization task addressed in this paper. Section 3 recalls the concept of auxiliary function and then introduces a general auxiliary function for the β-NMF problem. Section 4 describes algorithms based on the proposed auxiliary function, namely MM and ME algorithms, and describe how they relate to the heuristic update (4). Section 5 reports simulations and convergence behaviors on synthetic and real data (with audio transcription and face interpolation examples). Section 6 describes extensions of the proposed algorithms to penalized and convex- NMF. Section 7 concludes and discusses open questions.

2 Preliminaries

In this section we present the β-divergence and more precisely specify the task that is addressed in this paper. A detailed exposition of the β-divergence can be found in (Cichocki and Amari, 2010).

2.1 Definition of the β-divergence

The β-divergence was introduced by Basu et al. (1998) and Eguchi and Kano (2001) and can be defined as

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta\,(\beta-1)} \left(x^\beta + (\beta-1)\, y^\beta - \beta\, x\, y^{\beta-1} \right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x \,\log \frac{x}{y} - x + y & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{5}$$

Basu et al. (1998) and Eguchi and Kano (2001) assume $\beta > 1$, but the definition domain can be extended to $\beta \in \mathbb{R}$, as suggested by Cichocki et al. (2006), which is the definition domain that is considered in this paper. The β-divergence can be shown continuous in β by using the identity $\lim_{\beta\to 0} (x^\beta - y^\beta)/\beta = \log(x/y)$. The limit cases $\beta = 0$ and $\beta = 1$ correspond to the IS and KL divergences, respectively. The β-divergence coincides up to a factor $1/\beta$ with the "generalized divergence" of Kompass (2007) which, in the context of NMF as well, was separately constructed so as to interpolate between the KL divergence ($\beta = 1$) and the Euclidean distance ($\beta = 2$). The β-divergence is plotted for various values of β on Figure 1. Note that in this paper we will abusively refer to $d_{\beta=2} = (x-y)^2/2$ as the Euclidean distance, though the latter is formally defined with a square root, and for vectors.

The first and second derivative of $d_\beta(x|y)$ w.r.t y are continuous in β, and write

$$d'_\beta(x|y) = y^{\beta-2}\,(y-x), \tag{6}$$

$$d''_\beta(x|y) = y^{\beta-3}\, \left[(\beta-1)y - (\beta-2)x\right]. \tag{7}$$

The derivative shows that $d_\beta(x|y)$, as a function of y, has a single minimum in $y = x$ and that it increases with $|y-x|$, justifying its relevance as a measure of fit. The second derivative shows that the β-divergence is convex w.r.t y for $\beta \in [1,2]$. Outside this interval the divergence can always be expressed as the sum of a convex, concave and constant part, such that

$$d_\beta(x|y) = \overset{\smile}{d}(x|y) + \overset{\frown}{d}(x|y) + \bar{d}(x) \tag{8}$$

where $\overset{\smile}{d}(x|y)$ is a convex function of y, $\overset{\frown}{d}(x|y)$ is a concave function of y and $\bar{d}(x)$ is a constant of y. The decomposition is not unique, since constant or linear terms (w.r.t y) are both convex and concave, or, less trivially, since any convex term can be added to $\overset{\smile}{d}(x|y)$ while subtracted from $\overset{\frown}{d}(x|y)$. In the following we will use the "natural conventions" given in Table 1.

As noted by Févotte et al. (2009), a noteworthy property of the β-divergence is its behavior w.r.t to scale, as the following equation holds for any value of β:

$$d_\beta(\lambda\, x|\lambda\, y) = \lambda^\beta\ d_\beta(x|y). \tag{9}$$

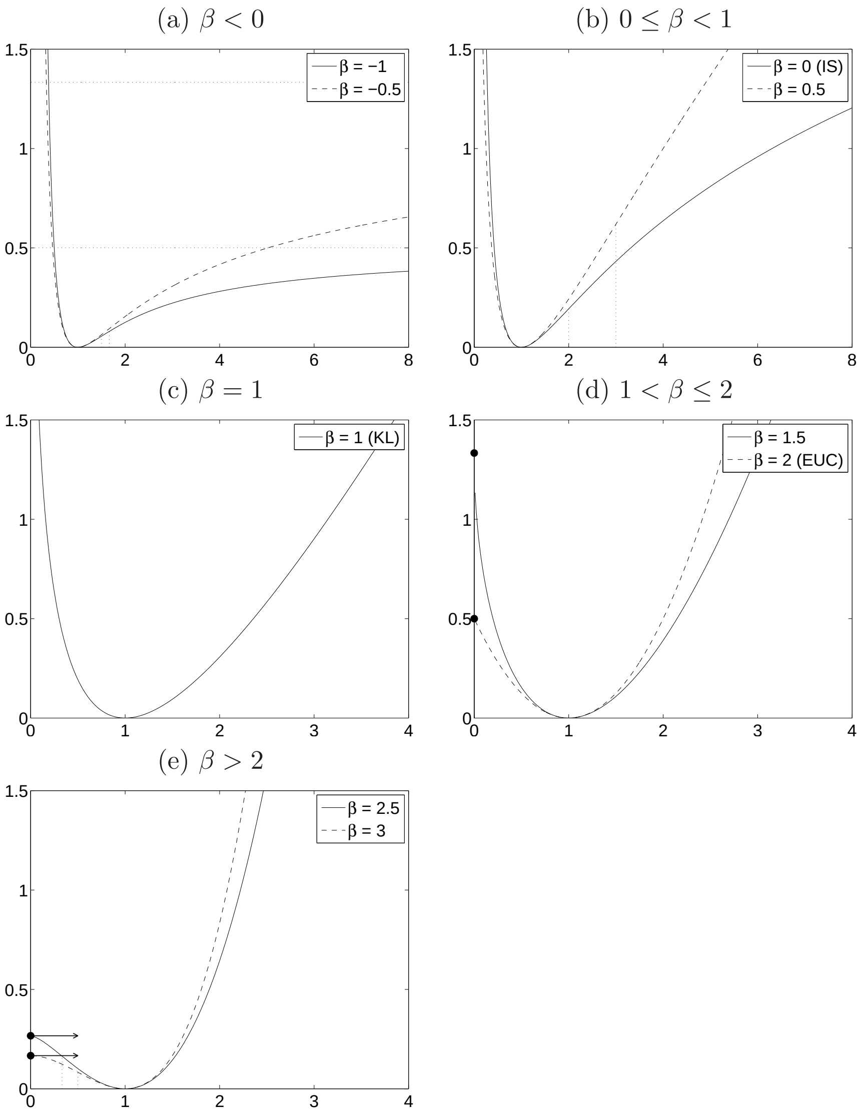


Figure 1: β-divergence $d_\beta(x|y)$ as a function of y (with $x = 1$). The subfigures illustrate the regimes of the β-divergence for its five characteristic ranges of values of β. The divergence is convex for $1 \leq \beta \leq 2$, as seen on subfigures (c) and (d). On the other subfigures, the inflection points are indicated with vertical dotted lines. For $\beta < 0$, the divergence possesses horizontal asymptotes of coordinate $x^\beta/(\beta(\beta-1))$ as $y \to \infty$. For $\beta > 1$, the divergence takes finite value $x^\beta/(\beta(\beta-1))$ at $y = 0$, where the derivative is zero for $\beta > 2$.

It implies that factorizations obtained with $\beta > 0$ (such as with the Euclidean distance or the KL divergence) will rely more heavily on the largest data values and less precision is to be expected in the estimation of the low-power components, and conversely factorizations obtained with $\beta < 0$ will rely more heavily on smallest data values. The IS divergence ($\beta = 0$) is scale-invariant, i.e., $d_{IS}(\lambda\, x|\lambda\, y) = d_{IS}(x|y)$, and is the only one in the family of β-divergences to possess this property. Factorizations with small positive values of β are relevant to decomposition of audio spectra, which typically exhibit exponential power decrease along frequency f and also usually comprise low-power transient components such as note attacks together with higher power components such as tonal parts of sustained notes. For example, Févotte et al. (2009) present the results of the decomposition of a piano power spectrogram with IS-NMF and show that components corresponding to very low residual noise and hammer hits on the strings are extracted with great accuracy, while these components are either ignored or severely degraded when using Euclidean or KL divergences. Similarly, the value $\beta = 0.5$ is advocated by FitzGerald et al. (2009); Hennequin et al. (2010) and has been shown to give optimal results in music transcription based on NMF of the magnitude spectrogram by Vincent et al. (2010).

The β-divergence belongs to the family of Bregman divergences. For $\beta \notin \{0, 1\}$, a suitable Bregman generating function is $\phi(y) = y^\beta/(\beta(\beta-1))$, as noted by Févotte and Cemgil (2009). This function, however, cannot generate the IS and KL divergences by continuity when β tends to 0 or 1. The latter divergences may nonetheless be generated "separately", using the functions $\phi(y) = -\log y$ and $\phi(y) = y \log y$, respectively. Cichocki and Amari (2010) give a general Bregman generating function of the β-divergence, defined for all $\beta \in \mathbb{R}$, in the form of $\phi_{\beta \neq 0,1}(y) = (y^\beta - \beta y + \beta - 1)/(\beta(\beta-1))$, $\phi_{\beta=0}(y) = y - \log y - 1$ and $\phi_{\beta=1}(y) = y \log y - y + 1$. NMF with Bregman divergences has been considered by Dhillon and Sra (2005), where the lack of results about the monotonicity of multiplicative algorithms in general has been noted.[1] This paper fills this gap for the specific case of β-divergence.

2.2 Task

Core optimization problem As to our best knowledge all algorithms in the literature to date, the NMF algorithms we describe in this paper sequentially update $\mathbf{H}$ given $\mathbf{W}$ and then $\mathbf{W}$ given $\mathbf{H}$. These two steps are essentially the same, by symmetry of the factorization ($\mathbf{V} \approx \mathbf{WH}$ is equivalent to $\mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$ and the roles of $\mathbf{W}$ and $\mathbf{H}$ are simply exchanged), and because we are not making any assumption on the relative values of F and N. Hence, we may concentrate on solving the following subproblem

$$\min_{\mathbf{H}} C(\mathbf{H}) \overset{\text{def}}{=} D(\mathbf{V}|\mathbf{WH}) \text{ subject to } \mathbf{H} \geq 0 \tag{10}$$

with fixed $\mathbf{W}$ and where in the rest of the paper $D(\mathbf{V}|\mathbf{WH})$ is as of Eq. (3) with $d(x|y) = d_\beta(x|y)$. The criterion function $C(\mathbf{H})$ separates into $\sum_n D(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n)$, where $\mathbf{v}_n$ and $\mathbf{h}_n$ are the n^{th} row of $\mathbf{V}$ and $\mathbf{H}$, respectively, so that we are essentially left with solving the problem

$$\min_{\mathbf{h}} C(\mathbf{h}) = D(\mathbf{v}|\mathbf{Wh}) \text{ subject to } \mathbf{h} \geq 0 \tag{11}$$

where $\mathbf{v} \in \mathbb{R}_+^F$, $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{h} \in \mathbb{R}_+^K$.

KKT necessary conditions An admissible solution $\mathbf{h}^\star$ to problem (11) must satisfy the Karush-Kuhn-Tucker (KKT) first order optimality conditions, which write

$$\nabla_{\mathbf{h}} C(\mathbf{h}^\star).\mathbf{h}^\star = 0 \tag{12}$$

$$\nabla_{\mathbf{h}} C(\mathbf{h}^\star) \geq 0 \tag{13}$$

$$\mathbf{h}^\star \geq 0 \tag{14}$$

where the dot notation '.' denotes entrywise operations (here term-to-term multiplication) and $\nabla_{\mathbf{h}} C(\mathbf{h})$ denotes the gradient of $C(\mathbf{h})$, given by

$$\nabla_{\mathbf{h}} C(\mathbf{h}) = \mathbf{W}^T \left[d'(v_f|[\mathbf{Wh}]_f)\right]_f \tag{15}$$

$$= \mathbf{W}^T [(\mathbf{Wh})^{\cdot(\beta-2)}(\mathbf{Wh} - \mathbf{v})] \tag{16}$$

where the notation $[x_f]_f$ refers to the column vector $[x_1, \ldots, x_F]^T$. The KKT conditions (12)-(14) can be summarized as

$$\min\{\mathbf{h}^\star, \nabla_{\mathbf{h}} C(\mathbf{h}^\star)\} = \mathbf{0}_K \tag{17}$$

where the "min" operator is entrywise and $\mathbf{0}_K$ is a null vector of dimension K.

Algorithms In the following, we will say that an algorithm is *monotone* if it produces a sequence of iterates $\{\mathbf{h}^{(i)}\}_{i \geq 0}$, such that $C(\mathbf{h}^{(i+1)}) \leq C(\mathbf{h}^{(i)})$ for all $i \geq 0$. An algorithm is said *convergent* if it produces a sequence of iterates $\{\mathbf{h}^{(i)}\}_{i \geq 0}$ which converges to a limit point $\mathbf{h}^\star$ satisfying the KKT conditions (12)-(14). Monotonicity does not imply convergence in general, nor is monotonicity necessary to convergence.

3 An auxiliary function for β-NMF

In this section we properly define the concept of auxiliary function and then exhibit a separable auxiliary function for the β-NMF problem.

[1] More precisely, Dhillon and Sra (2005) give proofs of monotonicity for the "reverse" problem of minimizing $D(\mathbf{WH}|\mathbf{V})$ instead of $D(\mathbf{V}|\mathbf{WH})$, while pointing that monotonicity of multiplicative algorithms based on the heuristic (4) for the latter problem is however observed in practice.

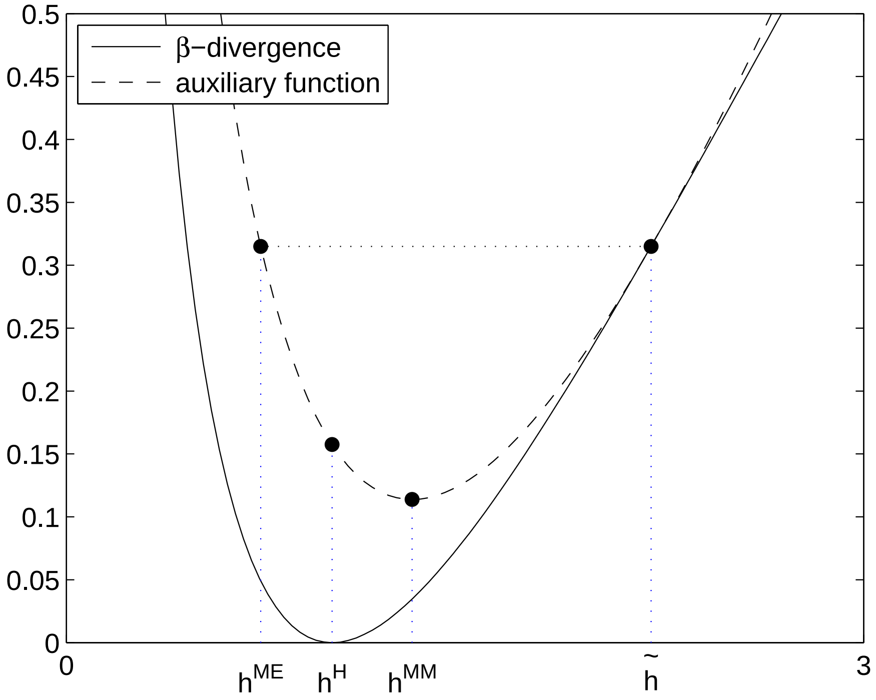


Figure 2: The β-divergence $d_\beta(x|y)$ for $\beta = 0.5$ (with $x = 1$) and its auxiliary function in dimension one (with $\tilde{h} = 2.2$). The MM update h^{MM} corresponds to the minimum of the auxiliary function, see Section 4.1. The heuristic update h^{H} given by Eq. (4) is discussed in Section 4.2 (the heuristic update minimizes the criterion function in the simple one-dimensional case but this is not true in larger dimensions). The ME update h^{ME} consists in selecting the next update "beyond the valley" defined by the auxiliary function, from the current solution $\tilde{h}$, see Section 4.3.

3.1 Definition of auxiliary function

Definition 1 (Auxiliary function). The $\mathbb{R}_+^K \times \mathbb{R}_+^K \to \mathbb{R}_+$ mapping $G(\mathbf{h}|\tilde{\mathbf{h}})$ is said to be an *auxiliary function* to $C(\mathbf{h})$ if and only if

- $\forall \mathbf{h} \in \mathbb{R}_+^K$, $C(\mathbf{h}) = G(\mathbf{h}|\mathbf{h})$
- $\forall (\mathbf{h}, \tilde{\mathbf{h}}) \in \mathbb{R}_+^K \times \mathbb{R}_+^K$, $C(\mathbf{h}) \leq G(\mathbf{h}|\tilde{\mathbf{h}})$

In other words an auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ is a *majorizing function* or *upper bound* of $C(\mathbf{h})$ which is tight for $\mathbf{h} = \tilde{\mathbf{h}}$. The optimization of $C(\mathbf{h})$ can be replaced by iterative optimization of $G(\mathbf{h}|\tilde{\mathbf{h}})$. Indeed, any iterate $\mathbf{h}^{(i+1)}$ satisfying

$$G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) \tag{18}$$

satisfies $C(\mathbf{h}^{(i+1)}) \leq C(\mathbf{h}^{(i)})$, because we have

$$C(\mathbf{h}^{(i+1)}) \leq G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) = C(\mathbf{h}^{(i)}). \tag{19}$$

The iterate $\mathbf{h}^{(i+1)}$ is typically chosen as

$$\mathbf{h}^{(i+1)} = \arg\min_{\mathbf{h} \geq 0} \; G(\mathbf{h}|\mathbf{h}^{(i)}) \tag{20}$$

which forms the basis of *maximization-minimization* (MM) algorithms, see, e.g., (Hunter and Lange, 2004) for a tutorial. However, any other iterate $\mathbf{h}^{(i+1)}$ satisfying (18) produces a monotone algorithm. As such, Figure 2 illustrates the three updates strategies that will be developed in this paper.

3.2 Separable auxiliary function for β-NMF

In this section we construct an auxiliary function to $C(\mathbf{h})$ for the specific case of the β-divergence. Our approach follows the one of Cao et al. (1999) for IS divergence, and consists of majorizing the convex part of the criterion using Jensen's inequality and majorizing the concave part by its tangent, as detailed in the proof of the following theorem. Here and henceforth, we denote $\mathbf{W}\tilde{\mathbf{h}}$ by $\tilde{\mathbf{v}}$, with entries $[\mathbf{W}\tilde{\mathbf{h}}]_f = \tilde{v}_f$.

Theorem 1 (Auxiliary function for β-NMF). Let $\tilde{\mathbf{h}}$ be such that

(i) $\forall f, \tilde{v}_f > 0$

(ii) $\forall k, \tilde{h}_k > 0$

Then, the function $G(\mathbf{h}|\tilde{\mathbf{h}})$ defined by

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \left[\sum_k \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} \overset{\smile}{d}\left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) \right] + \left[\overset{\frown}{d}'(v_f|\tilde{v}_f) \sum_k w_{fk}(h_k - \tilde{h}_k) + \overset{\frown}{d}(v_f|\tilde{v}_f) \right] + \bar{d}(v_f) \tag{21}$$

is an auxiliary function to $C(\mathbf{h}) = \sum_f d(v_f|[\mathbf{W}\mathbf{h}]_f)$, where $\overset{\smile}{d}(x|y) + \overset{\frown}{d}(x|y) + \bar{d}(x)$ is any differentiable convex-concave-constant decomposition of the β-divergence, such as the one defined in Table 1.

Proof. The condition $G(\mathbf{h}|\mathbf{h}) = C(\mathbf{h})$ is trivially met. The criterion $C(\mathbf{h})$ may be written as

$$C(\mathbf{h}) = \sum_f C_f(\mathbf{h}) \tag{22}$$

where $C_f(\mathbf{h}) \overset{\text{def}}{=} d(v_f|[\mathbf{W}\mathbf{h}]_f)$. We prove $C(\mathbf{h}) \leq G(\mathbf{h}|\tilde{\mathbf{h}})$ by constructing an auxiliary function to each part $C_f(\mathbf{h})$ of the criterion, and more precisely by treating the convex and concave part separately. Let us define $\overset{\smile}{C}_f(\mathbf{h}) \overset{\text{def}}{=} \overset{\smile}{d}(v_f|[\mathbf{W}\mathbf{h}]_f)$ and $\overset{\frown}{C}_f(\mathbf{h}) \overset{\text{def}}{=} \overset{\frown}{d}(v_f|[\mathbf{W}\mathbf{h}]_f)$, so that we can write

$$C_f(\mathbf{h}) = \overset{\smile}{C}_f(\mathbf{h}) + \overset{\frown}{C}_f(\mathbf{h}) + \bar{d}(v_f). \tag{23}$$

Convex part: We first prove that

$$\overset{\smile}{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_k \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} \overset{\smile}{d}\left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) \tag{24}$$

is an auxiliary function to $\overset{\smile}{C}_f(\mathbf{h})$. The condition $\overset{\smile}{G}_f(\mathbf{h}|\mathbf{h}) = \overset{\smile}{C}_f(\mathbf{h})$ is trivially met. The condition $\overset{\smile}{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) \geq \overset{\smile}{C}_f(\tilde{\mathbf{h}})$ is proven as follows. Let $\mathcal{K}$ be the set of indices k such that $w_{fk} \neq 0$. Define $\forall k \in \mathcal{K}$,

$$\tilde{\lambda}_{fk} = \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} = \frac{w_{fk}\tilde{h}_k}{\sum_{\ell \in \mathcal{K}} w_{f\ell}\tilde{h}_\ell}. \tag{25}$$

We have $\sum_{k \in \mathcal{K}} \tilde{\lambda}_{fk} = 1$ and

$$\begin{aligned}
\overset{\smile}{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) &= \sum_{k \in \mathcal{K}} \tilde{\lambda}_{fk} \overset{\smile}{d}\left(v_f | \frac{w_{fk}h_k}{\tilde{\lambda}_{fk}} \right) && (26)\\
&\geq \overset{\smile}{d}\left(v_f | \sum_{k \in \mathcal{K}} \tilde{\lambda}_{fk} \frac{w_{fk}h_k}{\tilde{\lambda}_{fk}} \right) && (27)\\
&= \overset{\smile}{d}\left(v_f | \sum_{k=1}^{K} w_{fk}h_k \right) && (28)\\
&= \overset{\smile}{C}_f(\mathbf{h}) && (29)
\end{aligned}$$

where we used Jensen's inequality, by convexity of $\overset{\smile}{d}(x|y)$.

Concave part: An auxiliary function $\overset{\frown}{G}_f(\mathbf{h}|\tilde{\mathbf{h}})$ to the concave part $\overset{\frown}{C}_f(\mathbf{h})$ can be taken as the first order Taylor approximation to $\overset{\frown}{C}_f(\mathbf{h})$ in the vicinity of $\tilde{\mathbf{h}}$, i.e.,

$$\overset{\frown}{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \overset{\frown}{C}_f(\tilde{\mathbf{h}}) + \nabla^T \overset{\frown}{C}_f(\tilde{\mathbf{h}})\ (\mathbf{h} - \tilde{\mathbf{h}}). \tag{30}$$

The function satisfies $\overset{\frown}{G}_f(\mathbf{h}|\mathbf{h}) = \overset{\frown}{C}_f(\mathbf{h})$ by construction and $\overset{\frown}{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) \geq \overset{\frown}{C}_f(\mathbf{h})$ by concavity of $\overset{\frown}{C}_f(\mathbf{h})$, using the property that the tangent to any point is an upper bound of a concave function.[2] Using

$$\nabla_{h_k} \overset{\frown}{C}_f(\mathbf{h}) = w_{fk} \overset{\frown}{d}'(v_f|[\mathbf{W}\mathbf{h}]_f) \tag{31}$$

the explicit form for $\overset{\frown}{G}_f(\mathbf{h}|\tilde{\mathbf{h}})$ is given by

$$\overset{\frown}{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \overset{\frown}{d}(v_f|\tilde{v}_f) + \overset{\frown}{d}'(v_f|\tilde{v}_f) \sum_k w_{fk}(h_k - \tilde{h}_k). \tag{32}$$

In the end a suitable auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ to $C(\mathbf{h})$ is obtained by summing up the auxiliary functions constructed for each individual part of the criterion, i.e.,

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \left(\overset{\smile}{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) + \overset{\frown}{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) + \bar{d}(v_f) \right) \tag{33}$$

which leads to Eq. (21). □

Properties of the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ is by construction separable in functions of the individual coefficients h_k of $\mathbf{h}$, which allows to decouple the optimization. It is convenient to rewrite the auxiliary function as such in order to derive some of the algorithms of Section (4). We may write

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_k G_k(h_k|\tilde{\mathbf{h}}) + cst \tag{34}$$

where cst is a constant w.r.t $\mathbf{h}$ and

$$G_k(h_k|\tilde{\mathbf{h}}) \overset{\text{def}}{=} \tilde{h}_k \left[\sum_f \frac{w_{fk}}{\tilde{v}_f} \overset{\smile}{d} \left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) \right] + h_k \left[\sum_f w_{fk} \overset{\frown}{d}'(v_f|\tilde{v}_f) \right]. \tag{35}$$

The gradient of the auxiliary function is given by

$$\nabla_{h_k} G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f w_{fk} \left[\overset{\smile}{d}' \left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) + \overset{\frown}{d}'(v_f|\tilde{v}_f) \right]. \tag{36}$$

Thanks to the separability of the auxiliary function into its variables the Hessian matrix is diagonal with

$$\nabla^2_{h_k} G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \tilde{v}_f \frac{w_{fk}}{\tilde{h}_k} \overset{\smile}{d}'' \left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right). \tag{37}$$

By convexity of $\overset{\smile}{d}(x|y)$ we have $\overset{\smile}{d}''(x|y) \geq 0$ which implies positive definiteness of the Hessian matrix and hence convexity of the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ (convexity more simply derives from the fact that the auxiliary function is built as a sum of convex functions).

Connections with other works The construction of $G(\mathbf{h}|\tilde{\mathbf{h}})$ employs standard mathematical tools (Jensen's inequality, Taylor approximation) that are well known from the MM literature, see, e.g., (Hunter and Lange, 2004). For $\beta \in [1, 2]$, $G(\mathbf{h}|\tilde{\mathbf{h}})$ coincides with the auxiliary function built by Kompass (2007). This latter paper proposed itself a generalization of the auxiliary functions proposed by Lee and Seung (2001) for the Euclidean distance ($\beta = 2$) and the generalized KL divergence ($\beta = 1$). For $\beta = 0$ (IS divergence), $G(\mathbf{h}|\tilde{\mathbf{h}})$ coincides with the auxiliary function proposed by Cao et al. (1999). It is worth recalling that in the algorithms proposed by Lee and Seung (2001) the update of $\mathbf{W}$ given $\mathbf{H}$ or $\mathbf{H}$ given $\mathbf{W}$ are instances of well known algorithms for image restoration (for which $\mathbf{W}$ acts as a fixed, known blurring matrix and $\mathbf{H}$ is a vectorized image to be reconstructed). These algorithms are the Iterative Space Reconstructing Algorithm (ISRA) of Daube-Witherspoon and Muehllehner (1986) and the Richardson-Lucy (RL) algorithm of Richardson (1972); Lucy (1974), which perform nonnegative linear regression with the Euclidean distance and KL divergence, respectively. The ISRA and RL algorithms are shown to be MM algorithms by De Pierro (1993). Similarly, the algorithms proposed by Cao et al. (1999) for nonnegative linear regression with the IS divergence were designed in the image restoration setting. Finally, let us mention that an auxiliary function based on Jensen's inequality for NMF with the α-divergence (which is always convex w.r.t to its second argument) is given by Cichocki et al. (2008).

[2] $\overset{\frown}{C}_f(\mathbf{h}) = \overset{\frown}{d}(v_f|[\mathbf{W}\mathbf{h}]_f)$ is concave as the composition of a concave function and a linear function.

	$\beta < 1$	$1 \leq \beta \leq 2$	$\beta > 2$
$\gamma(\beta)$	$\frac{1}{2-\beta}$	1	$\frac{1}{\beta-1}$

Table 2: Exponent in the multiplicative updates given by the MM algorithm.

4 Algorithms for β-NMF

In this section we describe algorithms for β-NMF based on the auxiliary function constructed in the latter section. In the following $\tilde{\mathbf{h}}$ should be understood as the current iterate $\mathbf{h}^{(i)}$ and we are seeking to obtain $\mathbf{h}^{(i+1)}$ such that Eq. (18) is satisfied.

4.1 Maximization-Minimization (MM) algorithm

An MM algorithm can be derived by minimizing the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ w.r.t to $\mathbf{h}$. Given the convexity and the separability of the auxiliary function the optimum is obtained by canceling the gradient given by Eq. (36). This is trivially done and leads to the following update:

$$h_k^{\mathrm{MM}} = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1}} \right)^{\gamma(\beta)} \tag{38}$$

where $\gamma(\beta)$ is given in Table 2. Note that $\gamma(\beta) \leq 1, \forall\beta$. As suggested in Section 1, the gradient of the criterion may be written as the difference of two nonnegative functions such that

$$\nabla_{h_k} C(\tilde{\mathbf{h}}) = \nabla^+_{h_k} C(\tilde{\mathbf{h}}) - \nabla^-_{h_k} C(\tilde{\mathbf{h}}) \tag{39}$$

$$\nabla^+_{h_k} C(\tilde{\mathbf{h}}) = \sum_f w_{fk}\, \tilde{v}_f^{\beta-1} \tag{40}$$

$$\nabla^-_{h_k} C(\tilde{\mathbf{h}}) = \sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2} \tag{41}$$

so that the update (38) can be rewritten in the more interpretable form

$$h_k^{\mathrm{MM}} = \tilde{h}_k \left(\frac{\nabla^-_{h_k} C(\tilde{\mathbf{h}})}{\nabla^+_{h_k} C(\tilde{\mathbf{h}})} \right)^{\gamma(\beta)} . \tag{42}$$

The conclusion is thus that the MM algorithm leads to multiplicative updates, but the latter differ from the "usual ones", obtained by setting $\gamma(\beta) = 1$ for all β and derived heuristically by Cichocki et al. (2006) through gradient descent with adaptative step or by Févotte et al. (2009) by splitting the gradient into two nonnegative functions as discussed above and in Section 1. The MM update differs from the heuristic update by the exponent $\gamma(\beta)$ which is not equal to one for $\beta \notin [1, 2]$.

4.2 Heuristic algorithm

This section discusses the properties of the heuristic update proposed by Cichocki et al. (2006); Févotte et al. (2009) and defined for all β by

$$h_k^{\mathrm{H}} = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1}} \right) . \tag{43}$$

Very few mathematical results exist for the heuristic update when β falls outside $[1, 2]$, i.e., when the β-divergence $d_\beta(x|y)$ is not convex. In such a case, the heuristic update can be erroneously interpreted as an MM algorithm by wrongly applying Jensen's inequality to $C(\mathbf{h})$. Yet, in the particular case $\beta = 0$, it holds that each heuristic update produces a decrease of $C(\mathbf{h})$ (Cao et al., 1999). One objective of this section is to extend this result to all values of β between 0 and 1.

Let us first introduce a scalar auxiliary function $g(y|\tilde{y}; x)$ as follows:

$$\forall y, \tilde{y}, x > 0, \quad g(y|\tilde{y}; x) = \breve{d}(x|y) + \overset{\frown}{d}(x|\tilde{y}) + (y - \tilde{y})\overset{\frown}{d}{}'(x|\tilde{y}) + \bar{d}(x) \tag{44}$$

where $\breve{d}(x|y)$, $\overset{\frown}{d}(x|y)$ and $\bar{d}(x|y)$ are defined in Table 1. By immediate application of Theorem 1 to the scalar case, $g(y|\tilde{y}; x)$ is an auxiliary function to $d(x|y)$. In particular, $g(\tilde{y}|\tilde{y}; x) = d(x|\tilde{y})$. Then, we have the following preliminary result.

Lemma 1. For all $\beta \in \mathbb{R}$,

$$G_k(h_k|\tilde{\mathbf{h}}) = \frac{1}{\tilde{h}_k^{\beta-1}} \left(\sum_f w_{fk} \tilde{v}_f^{\beta-1} \right) g(h_k|\tilde{h}_k; h_k^{\mathrm{H}}) + cst. \tag{45}$$

Proof. For each of the four possible expressions of $(\overset{\frown}{d}, \breve{d})$ given in Table 1, the validity of (45) can be checked straightforwardly by direct verification. □

As already mentioned in Section 3.1, the MM update (20) is not the only way of taking advantage of the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ to obtain a decrease of $C(\mathbf{h})$: any update satisfying (18) also ensures that $C(\mathbf{h})$ does not increase. This is a key remark to understand the behavior of the heuristic algorithm for $\beta \in (0,1)$, given the following property.

Theorem 2. For all $\beta \in (0,1)$, and all $\tilde{\mathbf{h}}$ such that conditions (i)-(ii) of Theorem 1 hold, the heuristic algorithm produces nonincreasing values of $C(\mathbf{h})$, according to the following inequality:

$$G(\mathbf{h}^{\mathrm{H}}|\tilde{\mathbf{h}}) \leq G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}). \tag{46}$$

Proof. For all $\beta \in (0,1)$, straightforward calculations yield

$$\begin{aligned} g(\tilde{y}|\tilde{y}; x) - g(x|\tilde{y}; x) &= \breve{d}(x|\tilde{y}) - \breve{d}(x|x) - (x - \tilde{y})\overset{\frown}{d}{}'(x|\tilde{y}) && (47) \\ &= \frac{1}{1-\beta}\, \tilde{y}^{\beta}(1 - \beta + \beta\theta - \theta^{\beta}) && (48) \end{aligned}$$

where $\theta = x/\tilde{y}$. Since $f(\theta) = \theta^{\beta}$ is a concave function of θ, we have $f(\theta) \leq f(1) + (\theta - 1)f'(1)$, which also reads $\theta^{\beta} \leq 1 + (\theta - 1)\beta$. Hence, $g(\tilde{y}|\tilde{y}; x) - g(x|\tilde{y}; x) \geq 0$ for all x, $\tilde{y}$. The latter inequality implies $\forall k$, $g(h_k^{\mathrm{H}}|\tilde{h}_k, h_k^{\mathrm{H}}) \leq g(\tilde{h}_k|\tilde{h}_k, h_k^{\mathrm{H}})$, so that we have $G_k(h_k^{\mathrm{H}}|\tilde{\mathbf{h}}) \leq G_k(\tilde{h}_k|\tilde{\mathbf{h}})$ according to (45), which leads to the result by summation over k. □

Cao et al. (1999) show that inequality (46) becomes an equality in the case $\beta = 0$, so that each heuristic update yields $G(\mathbf{h}^{\mathrm{H}}|\tilde{\mathbf{h}}) = G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}})$. In this particular case, the heuristic algorithm can be called a "majorization-equalization" algorithm, a class of algorithms described in next section. For values of β outside the range $[0,2]$, inequality (46) does not hold anymore.[3] Of course, this does not mean that the heuristic updates produce increasing values of $C(\mathbf{h})$. On the contrary, numerical simulations tend to indicate that they always produce nonincreasing values of $C(\mathbf{h})$, but proving this is still an open issue for $\beta \notin [0,2]$. Compared to MM updates, heuristic updates produce larger or equal steps for all β, since it can trivially be shown that

$$\forall k, \quad |h_k^{\mathrm{H}} - \tilde{h}_k| \geq |h_k^{\mathrm{MM}} - \tilde{h}_k|. \tag{49}$$

For $\beta \notin [1,2]$, numerical simulations indicate that the heuristic algorithm is faster than the MM algorithm (and we recall that the two algorithms coincide for $\beta \in [1,2]$). Given (49), skipping from the latter to the former has an effect comparable to that of overrelaxation: on the average, stretching the steps allows to reduce their number to reach convergence. This will be discussed in more details in Section 4.4.

In order to produce even larger steps for $\beta \in [0,2]$, and yet nonincreasing values of $C(\mathbf{h})$, the following section explores the concept of majorization-equalization.

4.3 Majorization-Equalization (ME) algorithm

Let us introduce the general notion of ME update by the fact that the new iterate $\mathbf{h}^{\mathrm{ME}}$ fulfills

$$G(\mathbf{h}^{\mathrm{ME}}|\tilde{\mathbf{h}}) = G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}). \tag{50}$$

Eq. (50) actually defines a level set rather than a single point. Let us concentrate on the following more constrained and manageable condition, given the separability of $G(\mathbf{h}|\tilde{\mathbf{h}})$:

$$\forall k, \quad G_k(h_k^{\mathrm{ME}}|\tilde{\mathbf{h}}) = G_k(\tilde{h}_k|\tilde{\mathbf{h}}).$$

Given (45), this amounts to solve the following equation for y, for any $\tilde{y}, x > 0$:

$$g(y|\tilde{y}; x) = g(\tilde{y}|\tilde{y}; x). \tag{51}$$

[3] Indeed, we can prove that the reversed inequality holds for all $\beta < 0$, while no systematic result is known for $\beta > 2$.

$\beta \leq 0$	$0 \leq \beta \leq 1$	$1 \leq \beta \leq 2$	$\beta \geq 2$	d
0	0	2	2	1
-1	1/2	3/2	3	2
-2	2/3	4/3	4	3
-3	3/4	5/4	5	4

Table 3: Values of β for which ME updates are closed-form, by root extraction of polynomials of degree d.

Since $g(y|\tilde{y};x)$ is strictly convex w.r.t y, (51) has not more than two solutions, one of them being $\tilde{y}$. By construction, the selection of the other solution (provided that it exists) will provide ME steps that are larger than MM updates, i.e.,

$$\forall k, \quad |h_k^{\mathrm{ME}} - \tilde{h}_k| \geq |h_k^{\mathrm{MM}} - \tilde{h}_k|, \tag{52}$$

as illustrated by Figure 2. To go further on the determination of this solution, a case-by-case analysis must be performed, depending on the range of β.

Case 1: $\beta \in [0,1)$ In that case we have

$$g(y|\tilde{y};x) = \frac{1}{1-\beta} x\, y^{\beta-1} + y\tilde{y}^{\beta-1} + cst. \tag{53}$$

Let us remark that

$$\forall\, \tilde{y}, x > 0, \quad \lim_{y \to 0} g(y|\tilde{y};x) = \lim_{y \to \infty} g(y|\tilde{y};x) = \infty, \tag{54}$$

so that (51) always admits two positive solutions (or one double positive solution if $\tilde{y} = x$), one of the two being $y = \tilde{y}$. The other one is the solution of interest. However, it is not closed-form, except for specific values of β (see Table 3). More precisely, when $\beta = 1 - 1/d$ and d is an integer, the solution can be found by solving the following polynomial equation of degree d, for $z = y^{1/d}$:

$$(1-\beta) \sum_{\ell=1}^{d} \tilde{z}^{d-\ell} z^{\ell} - x = 0 \tag{55}$$

where $\tilde{z} = \tilde{y}^{1/d}$. Not surprisingly, the simplest case $\beta = 0$ ($d = 1$) leads us to $y = x$, and thus to $h_k^{\mathrm{ME}} = h_k^{\mathrm{H}}$. The case $\beta = 0.5$ ($d = 2$) is more interesting. The extraction of the positive root of (55) then provides the following update formula:

$$h_k^{\mathrm{ME}} = \frac{\tilde{h}_k}{4} \left(\sqrt{1 + 8 \frac{h_k^{\mathrm{H}}}{\tilde{h}_k}} - 1 \right)^2. \tag{56}$$

Let us remark that this expression does not correspond to a multiplicative update, although it ensures that positivity is maintained.

Case 2: $\beta \in (1,2]$ In that case we have

$$g(y|\tilde{y};x) = \frac{1}{\beta} y^{\beta} - \frac{1}{\beta-1} x y^{\beta-1} + cst. \tag{57}$$

$g(y|\tilde{y};x)$ tends toward ∞ for $y \to \infty$, but it remains finite for $y \to 0$. As a consequence, Eq. (51) only admits the trivial solution $y = \tilde{y}$ if $g(\tilde{y}|\tilde{y};x) > g(0|\tilde{y};x)$, and also the unwanted solution 0 if $g(\tilde{y}|\tilde{y};x) = g(0|\tilde{y};x)$. It is only when $g(\tilde{y}|\tilde{y};x) < g(0|\tilde{y};x)$ that a positive, non trivial solution exists. This solution is closed-form for specific values of β given in Table 3. They correspond to $\beta = 1 + 1/d$, where d is an integer. Eq. (51) then amounts to solve the following polynomial equation of degree d, for $z = y^{1/d}$:

$$\sum_{\ell=0}^{d} \tilde{z}^{d-\ell} z^{\ell} - (d+1)\, x = 0, \tag{58}$$

with $\tilde{z} = \tilde{y}^{1/d}$. The simplest case is $\beta = 2$ ($d = 1$), and the solution is then given by $y = 2x - \tilde{y}$ if $\tilde{y} < 2x$, which yields the overrelaxed update

$$h_k^{\mathrm{ME}} = 2h_k^{\mathrm{MM}} - \tilde{h}_k, \tag{59}$$

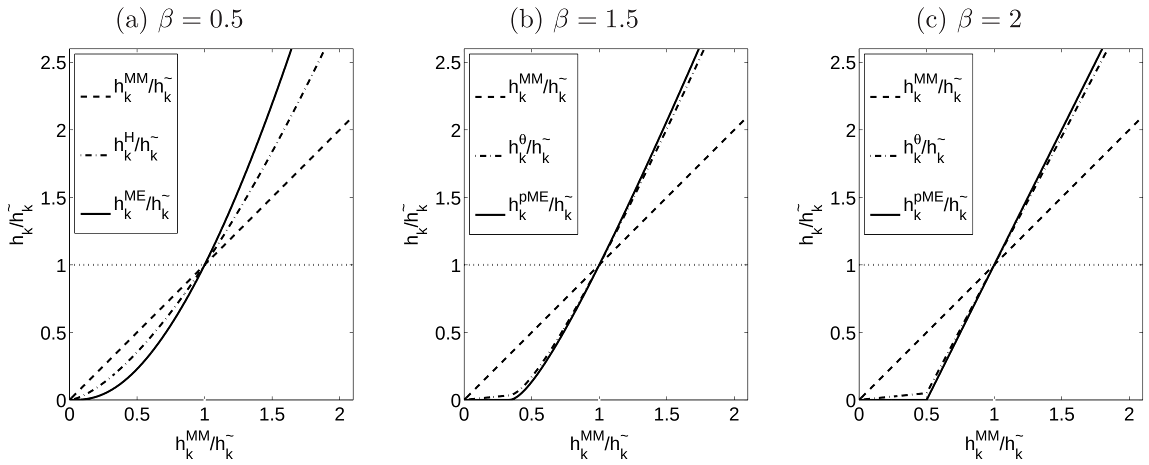


Figure 3: Normalized updates $h_k/\tilde{h}_k$ as functions of $h_k^{\mathrm{MM}}/\tilde{h}_k$ ($\theta = 0.9$). The region between the dotted, horizontal line and the solid line correspond to the steps that fulfill Eq. (18). The larger departure from the horizontal line, the larger the step.

provided that $\tilde{h}_k < 2h_k^{\mathrm{MM}}$. Note that this result more simply stems from the fact that when $\beta = 2$ the auxiliary function is parabolic and thus symmetric with respect to h_k^{MM}. In the case $\beta = 1.5$ ($d = 2$), a positive ME update exists if $\tilde{h}_k < 3h_k^{\mathrm{MM}}$, and it takes the following form:

$$h_k^{\mathrm{ME}} = \frac{\tilde{h}_k}{4}\left(\sqrt{12\frac{h_k^{\mathrm{MM}}}{\tilde{h}_k} - 3} - 1\right)^2. \tag{60}$$

As we need an update strategy that is defined everywhere, we propose to rely on a linear mixture between the MM update and a prolonged version of ME, defined as

$$h_k^{\theta} = \theta h_k^{\mathrm{pME}} + (1-\theta)h_k^{\mathrm{MM}} \tag{61}$$

where $\theta \in (0,1)$ and h_k^{pME} prolongs the ME update by zero when the latter does not exist:

$$h_k^{\mathrm{pME}} = \begin{cases} h_k^{\mathrm{ME}} & \text{if } h_k^{\mathrm{ME}} \text{ is defined} \\ 0 & \text{otherwise} \end{cases} \tag{62}$$

It is mathematically easy to check that h_k^{θ} fulfills Eq. (18) for all $\theta \in [0,1]$, and that positivity is maintained for all $\theta \in [0,1)$. In practice, values of θ near one may be favored to produce larger steps.

When $\beta < 0$ or $\beta > 2$, similar analyses can be conducted. In particular, there are specific values of β for which a closed-form expression of ME updates is available according to Table 3.

When $\beta < 0$, ME updates always exist since (53) and (54) still hold. Moreover, they provide nonincreasing values of $C(\mathbf{h})$, while the latter monotonicity property is not yet proved for the heuristic algorithm. However, simulations tend to indicate that the heuristic algorithm is faster than the ME algorithm (which is itself faster than the MM algorithm) in the case $\beta < 0$. This is in conformity with the fact that ME steps can then be proved to be smaller than heuristic steps (on the basis of the reversed inequality mentioned in Footnote 3).

When $\beta > 2$, ME updates do not necessarily exist, akin to the case $\beta \in (1,2]$. When they exist, they provide nonincreasing values of $C(\mathbf{h})$, while the latter is not yet proved for the heuristic formula. However, since this range of β values is not of utter practical interest, we will not go further into a detailed analysis here.

4.4 Overrelaxation properties of the heuristic and ME updates

As already stated, the heuristic and ME updates produce larger steps than the MM update, i.e., $|h_k^{\mathrm{H}} - \tilde{h}_k|$ and $|h_k^{\mathrm{ME}} - \tilde{h}_k|$ are larger than $|h_k^{\mathrm{MM}} - \tilde{h}_k|$, for all values of $\beta \in \mathbb{R}$. This is a form of overrelaxation, which will be shown in Section 5 to produce faster convergence in practice. The normalized ME (or pME) and heuristic updates studied in the above sections can be written as a function of $h_k^{\mathrm{MM}}/\tilde{h}_k$, such that

$$\frac{h_k}{\tilde{h}_k} = f\left(\frac{h_k^{\mathrm{MM}}}{\tilde{h}_k}\right) \tag{63}$$

where h_k is either h_k^{H}, h_k^{ME}, h_k^{pME} or h_k^{θ}. For the heuristic update, the function is simply given by $f(x) = x^{1/\gamma(\beta)}$. For $\beta \in \{0.5, 1.5, 2\}$, the function f corresponding to the ME/pME update is easily derived from equations (56), (60) and (59). Figure 3 displays the latter functions for the updates studied in this work when $\beta \in \{0.5, 1.5, 2\}$. Overrelaxation appears from the fact that $h_k < h_k^{\mathrm{MM}}$ whenever $h_k^{\mathrm{MM}} < \tilde{h}_k$ (steps towards left) and $h_k > h_k^{\mathrm{MM}}$ whenever $h_k^{\mathrm{MM}} > \tilde{h}_k$ (steps towards right).

General results about overrelaxation of MM algorithms are given by Salakhutdinov and Roweis (2003), and in particular in the case of NMF. The authors consider the specific case of KL divergence but their study holds for any divergence. They show that, *in a neighborhood of a stationary point*, for any $\eta \in (0, 2)$, relaxed updates h_k^{R} of the form

$$\frac{h_k^{\mathrm{R}}}{\tilde{h}_k} = \left(\frac{h_k^{\mathrm{MM}}}{\tilde{h}_k}\right)^{\eta} \tag{64}$$

will converge to the same point than h_k^{MM}, with a different, possibly better, rate of convergence. In particular, the optimal learning rate η, providing the largest rate of convergence, can be computed from the eigenvalues of the Jacobian, at convergence, of the mapping that relates h_k^{MM} at iteration (i) to h_k^{MM} at iteration $(i + 1)$. The optimal learning rate is shown to be always greater or equal to 1. A similar result was recently obtained by Badeau et al. (2010). However, these results do not translate into a practical algorithm, because the latter relaxation property only holds locally, and the computation of the optimal learning rate requires the stationary point to be known. As such, Salakhutdinov and Roweis (2003) propose an adaptive scheme which incrementally proposes values of η greater than one at each iteration, and backtrack to $\eta = 1$ when the criterion ceases decreasing.

Our results show that for $\beta \in (0, 1)$ the learning rate $\eta = 1/\gamma(\beta) = 2-\beta$, corresponding to the heuristic update, ensures descent of the criterion everywhere. The results of Salakhutdinov and Roweis (2003) indicate that the learning rate can be increased to $\eta = 2$ when the algorithm approaches the solution. Note that in the neighborhood of the solution the Taylor approximation $f(x) \approx f(1)+f'(1)(x-1)$ applied to $f(x) = x^{\eta}$ implies that

$$(h_k^{\mathrm{R}} - \tilde{h}_k) \approx \eta(h_k^{\mathrm{MM}} - \tilde{h}_k). \tag{65}$$

A similar approximation carried out with the ME/pME updates defined by equations (56), (60) for $\beta \in \{0.5, 1.5\}$ reveals that in these two cases $f'(1) = 2$ (and by construction $f(1) = 1$), so that, in a neighborhood of the solution we have

$$(h_k^{\mathrm{ME}} - \tilde{h}_k) \approx 2(h_k^{\mathrm{MM}} - \tilde{h}_k). \tag{66}$$

This means that the ME algorithms produce the largest admissible learning rate $\eta = 2$ in the neighborhood of the solution, while avoiding to adapt the learning rate so as to ensure monotonicity of the criterion. This results holds everywhere for $\beta = 2$, see (59), by symmetry of the auxiliary function w.r.t to h_k^{MM}. The interested reader may also refer to (Lantéri et al., 2001, 2010) for relaxation of multiplicative algorithms using adaptative learning rates computed through line search.

4.5 Implementation and complexity of the algorithms

As seen in previous section, the update rules of all the studied algorithms can be expressed as functions of the ratio $\nabla^{-}_{h_k} C(\tilde{\mathbf{h}})/\nabla^{+}_{h_k} C(\tilde{\mathbf{h}})$, which dominates the algorithmic complexities. Fortunately, the latter ratio takes a simple matrix form that leads to efficient implementations. As such, getting back to the original factorization problem, the heuristic update (43) for factors $\mathbf{H}$ and $\mathbf{W}$ can conveniently be expressed in the following matrix form

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T \left[(\mathbf{WH})^{\cdot(\beta-2)}.\mathbf{V}\right]}{\mathbf{W}^T \left[\mathbf{WH}\right]^{\cdot(\beta-1)}} \tag{67}$$

$$\mathbf{W} \leftarrow \mathbf{W}.\frac{\left[(\mathbf{WH})^{\cdot(\beta-2)}.\mathbf{V}\right]\mathbf{H}^T}{\left[\mathbf{WH}\right]^{\cdot(\beta-1)}\mathbf{H}^T} \tag{68}$$

where the division $\cdot/\cdot$ is here taken entrywise. The MM update simply involves bringing the corrective ratio to the power $\gamma(\beta)$, and the ME update involves applying a function specific to the value of β. Hence, the algorithms have similar complexity $\mathcal{O}(FKN)$ and their implementation take simple forms. MATLAB implementations of the algorithms discussed in this paper are available online at `http://perso.telecom-paristech.fr/~fevotte/Code/code_beta_nmf.zip`

5 Simulations

In this section we report performance results of β-NMF algorithms for the specific values $\beta \in \{0.5, 1.5, 2\}$. These values are chosen for their practical interest and because a simple ME algorithms exist in their case. As such this section will evidence the performance improvement brought by the ME approach over the MM or heuristic approaches, with similar computational burden. More precisely, the ME algorithm considered in this section is the mixture of prolonged ME and MM, defined by Eq. (61) and with $\theta = 0.95$, but we will still refer to it as ME for simplicity. The algorithms for all three considered values of β are compared on small-sized synthetic data in Section 5.1. The algorithms for $\beta = 0.5$ are analyzed in Section 5.2 on the basis of a small music transcription example as this specific value of β has proven efficient for this task (FitzGerald et al., 2009; Vincent et al., 2010; Hennequin et al., 2010).

In the following results we will display the cost values through iterations as well as, following Gonzalez and Zhang (2005), "KKT residuals". The residuals allow to monitor convergence to a stationary point and are here defined as

$$\mathrm{KKT}(\mathbf{W}) = \| \min \Big\{ \mathbf{W}, [(\mathbf{WH})^{\cdot(\beta-2)}.(\mathbf{WH} - \mathbf{V})]\mathbf{H}^T \Big\} \|_1 / FK \tag{69}$$

$$\mathrm{KKT}(\mathbf{H}) = \| \min \Big\{ \mathbf{H}, \mathbf{W}^T[(\mathbf{WH})^{\cdot(\beta-2)}.(\mathbf{WH} - \mathbf{V})] \Big\} \|_1 / KN. \tag{70}$$

They are meant to converge to zero, by Eq. (17). Again, the monotonicity of the heuristic, MM and ME algorithms does not imply convergence of the iterates to a stationary point. Hence, displaying the KKT residuals allows to experimentally check whether convergence is achieved in practice.

One iteration of each algorithm consists of updating $\mathbf{W}$ given $\mathbf{H}^{(i-1)}$ and $\mathbf{H}$ given $\mathbf{W}^{(i)}$, and then normalize $\mathbf{W}^{(i)}$ and $\mathbf{H}^{(i)}$ to eliminate trivial scale indeterminacies that leave the cost function unchanged. The normalization step consists of rescaling each column of $\mathbf{W}$ so that $\|\mathbf{w}_k\|_1 = 1$ and rescale the k^{th} row of $\mathbf{H}$ accordingly. The normalization step is not required *per se* but is useful to display and compare the KKT residuals, which are scale-sensitive.

5.1 Factorization of synthetic data

We consider a synthetic data matrix $\mathbf{V}$ constructed as $\mathbf{V} = \mathbf{W}^*\mathbf{H}^*$ where the ground truth factors are generated as the absolute values of Gaussian noise.[4] The matrix can be exactly factorized so that all algorithms should converge to a solution such that $D(\mathbf{V}|\mathbf{WH}) = 0$. The dimensions are $F = 10$, $N = 25$, $K = 5$. The algorithms (heuristic, MM, ME for $\beta = 0.5$, MM and ME for $\beta \in \{1.5, 2\}$) are run for 10^5 iterations and initialized with positive random values. Fig. 4, 5 and 6 display for each of the 3 values of β the normalized cost values $D(\mathbf{V}|\mathbf{WH})/FN$, the KKT residuals, as well as "fit residuals" computed as $\|\mathbf{W}^{(i)} - \hat{\mathbf{W}}\|_F/FK$ and $\|\mathbf{H}^{(i)} - \hat{\mathbf{H}}\|_F/KN$, where $\hat{\mathbf{W}}$ and $\hat{\mathbf{H}}$ are the factor estimates at the end of the 10^5 iterations and $\|.\|_F$ is the Frobenius norm. The fit residuals allow to measure the closeness of the current iterates to their end value.

The cost values in all three cases converge to zero as an exact factorization is reached (oscillations appear in the end iterations as machine precision is reached). Convergence is achieved in all three cases, as shown by both the cost values and KKT residuals. We visually inspected the factorizations returned by the algorithms. For each value of $\beta \in \{0.5, 1.5\}$, the different algorithms appeared to converge to the same solution (and solutions obtained for the two values of β appeared comparable). This was less clear for $\beta = 2$, where ME appeared to reach out a different solution than MM. Still, in this run ME provides fastest convergence for every considered value of β. Other runs, obtained from other starting points (obtained randomly), tend to show that when the compared algorithms converge to the same solution, then ME converges faster. Convergence to a common solution can be controlled in the specific case where $\mathbf{W}$ is fixed and $\beta \in \{1.5, 2\}$, because the objective function is then convex w.r.t $\mathbf{H}$. In this scenario ME was found to always converge faster than MM. These simulations are reported in the companion report available online at `http://perso.telecom-paristech.fr/~fevotte/Samples/neco11/beta_nmf_supp.pdf`

The fit residuals in Fig. 4, 5 and 6 show that full convergence will not need be attained to obtain satisfying solutions for most applications as the fit residual will be considered sufficiently small after a few hundred iterations. Note that the factor iterates do not necessarily converge to the ground truth values $\mathbf{W}^*$ and $\mathbf{H}^*$ (and this is what we observed indeed) because of the identifiability ambiguities inherent to NMF (Donoho and Stodden, 2004; Laurberg et al., 2008).

[4] E.g., in MATLAB notation `V = abs(randn(F,K))*abs(randn(K,N))`.

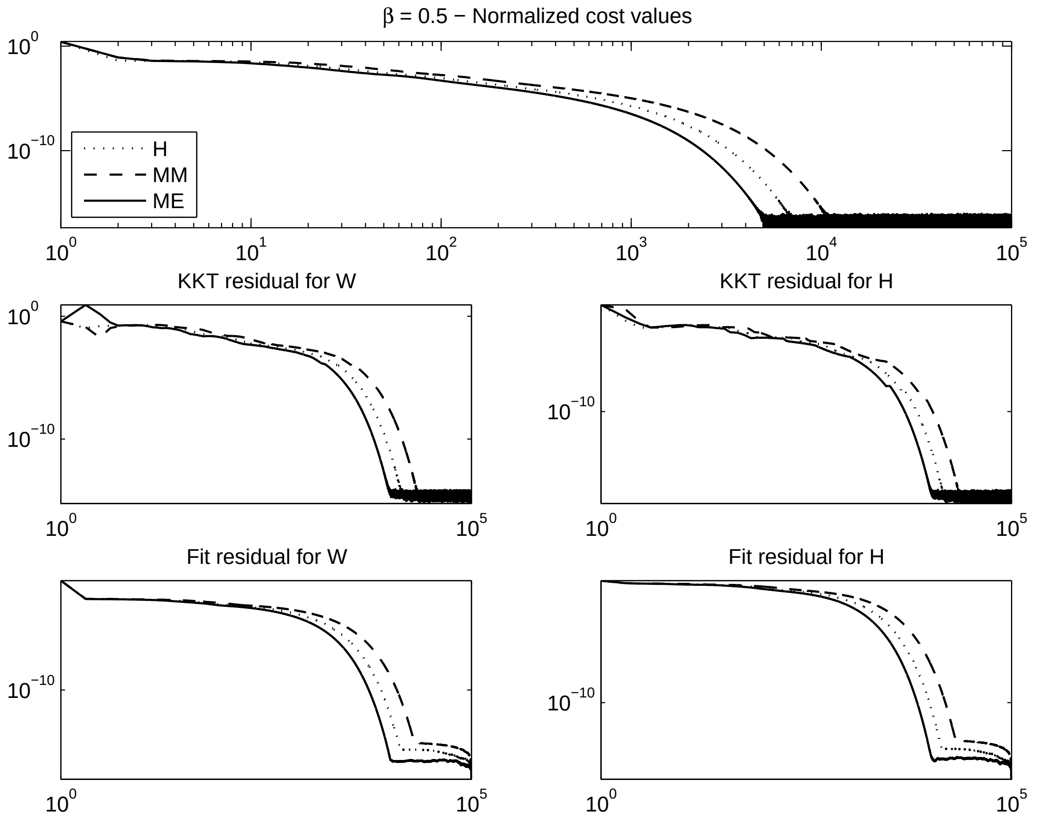


Figure 4: One run of the heuristic (H), ME and MM algorithms on synthetic data with $\beta = 0.5$. Logarithmic scales for both x- and y- axes.

Using a MATLAB implementation run on Mac 2.6 GHz with 2 GB RAM, the CPU time required by each algorithm for the 10^5 iterations is about 60 s for $\beta \in \{0.5, 1.5\}$ and 20 s for $\beta = 2$, including the computation of the cost values and KKT residuals. The ME algorithm is marginally more expensive than MM, itself only slightly more expensive than the heuristic algorithm, for $\beta = 0.5$. The CPU times incurred by the algorithms when $\beta = 2$ is considerably lower thanks to simplifications in Eq. (67) and (68). Indeed, in latter case the term $(\mathbf{WH})\mathbf{H}^T$ appearing at the denominator can more efficiently be computed as $\mathbf{W}(\mathbf{HH}^T)$, which involves a multiplication of matrices with smaller sizes.

5.2 Audio spectrogram decompostion

This section addresses the comparison of the heuristic, MM and ME algorithms for $\beta = 0.5$ applied to an audio spectrogram. We consider the short piano sequence of (Févotte et al., 2009), recorded in live conditions, composed of 4 musical notes, played all at once in the first measure and then played by pairs in all possible combinations in the subsequent measures. A magnitude spectrogram of the audio signal is computed, leading to nonnegative matrix data $\mathbf{V}$ of size $F = 513$ frequency bins by $N = 674$ time frames. The data is represented top-left of Fig. 7.

As discussed in (Févotte et al., 2009), K was set to 6 so as to retrieve in $\mathbf{W}$ the individual spectra of each of the 4 notes and supplementary spectra corresponding to transients and residual noise. The three algorithms were initialized with common positive random values and run for 10^5 iterations. Figure 7 displays the cost values and KKT residuals along the 10^5 iterations. It was manually checked that the algorithms converged to the desired “ground-truth” solution, i.e., the notes, transients and residual noise spectra are correctly unmingled. The three plots show that the ME provides fastest convergence overall though, judging from the KKT residuals, it appears that convergence is not achieved within the 10^5 iterations. However, the musical pitch values (computed from $\mathbf{W}$ at every iteration) converge to their ground truth values after only 30, 50 and 580 iterations for ME, heuristic and MM, respectively. Other initializations yielded two types of results. In a minority of cases, either the heuristic and MM update, on one side, or the ME update, on the other side, converged to a local solution. In the large majority of cases the three algorithms converge to the same solution and the results are similar to those of Figure 7: the heuristic algorithm produces largest decreases of the objective function in the early iterations and is then supplanted by ME. In some runs the pitch values converged faster with the heuristic algorithm

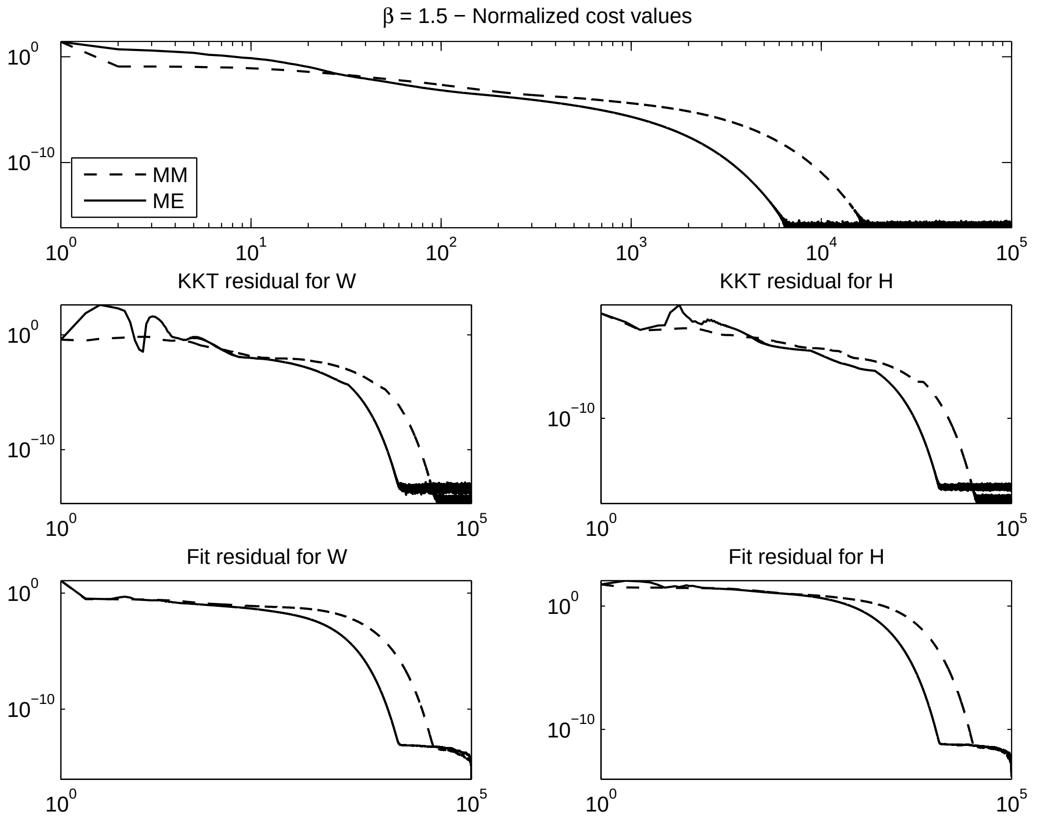


Figure 5: One run of the ME and MM algorithms on synthetic data with $\beta = 1.5$. Logarithmic scales for both x- and y- axes.

than with ME, and it was found that MM is generally slower than the other two algorithms. These results suggest a mixed update of the form of Eq. (61) where the mixture parameter θ could be made iteration-dependent so as to give more weight to the heuristic update in the early iterations and then to ME.

5.3 Face data decompostion

Finally, in this section we consider decomposition of face data using β-NMF. We use the Olivetti dataset, composed of 10 grayscale 8 bits 64×64 face images of 40 people. We retrieved the data in MATLAB format from `http://cs.nyu.edu/~roweis/data.html`. The images are vectorized and form the columns of $\mathbf{V}$, with dimensions $F = 4096$ and $N = 400$. Fig. 8 displays the objective functions of one run of the ME and MM algorithms for $\beta \in \{1.5, 2\}$, and illustrate the faster convergence of ME. Other runs led to sensibly similar plots.

As stated in the introduction, β-NMF is popular in audio signal processing where the value of β can be controlled so as to improve transcription or separation accuracy. The idea of tuning the value of β so as to optimize performance applies to any NMF-based method for any type of data. As such, to motivate the use of β-NMF in a non-audio setting we propose an image interpolation exemple, inspired by (Cichocki et al., 2008; Cemgil, 2009), where we show the influence of β on the reconstruction of missing data. We discard 25 % of the Olivetti data randomly and produce NMF decompositions using the available data for $\beta \in \{-1, 0, 1, 2, 3\}$ and $K \in \{50, 100, 200\}$. Accounting for the missing data requires minor modifications in the algorithms, basically multiplying $\mathbf{V}$ and its approximate $\mathbf{WH}$ with a binary mask in which zeroes indicate missing pixels, see (Ho, 2008; Cichocki et al., 2008; Le Roux et al., 2008; Cemgil, 2009; Smaragdis et al., 2009) for similar setups. For simplicity we only considered the MM algorithm, as it is consistently defined for all values of β. It was run from 5 different initializations for every combination (K, β), and the factorization yielding lowest end cost value was selected. Fig. 9 displays the original image, missing pixels and reconstructions for two of the images in the dataset. We have also computed the Peak Signal to Noise Ratio (PSNR) between original and reconstructed images.[5] The maximum mean PSNR value

[5] PSNR is a standard evaluation criterion in image reconstruction, defined as $20 \log_{10}(FP/\|\mathbf{v} - \hat{\mathbf{v}}\|_2)$, where $\mathbf{v}$ and $\hat{\mathbf{v}}$ denote the vectorized original and reconstructed images, and P is the maximum pixel possible value ($P = 255$ in our case).

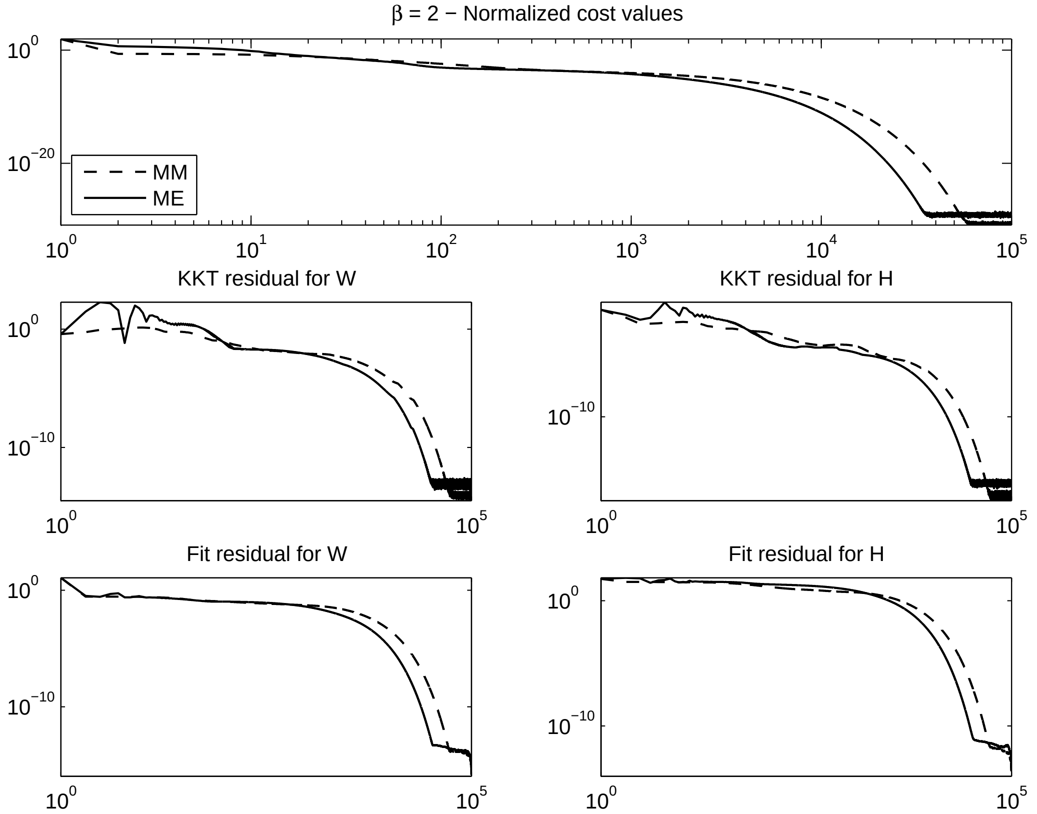


Figure 6: One run of the ME and MM algorithms on synthetic data with $\beta = 2$. Logarithmic scales for both x- and y- axes.

(averaged over all 400 images) is obtained for $\beta = 2$ and $K = 200$. However, since the PSNR value is equivalent to the Euclidean distance between the original and reconstructed image, it is expected that the optimal value of β is biased towards the metric used to assess the quality of reconstruction. Perceptually, we often found the reconstruction obtained with $\beta = 3$ to be more satisfying than with $\beta = 2$.

6 Variants of β-NMF

In this section we briefly discuss how some common variants of NMF, penalized NMF and convex-NMF, can be handled under the β-divergence.

Penalized β-NMF Supplementary functions of $\mathbf{W}$ and/or $\mathbf{H}$ are often added to the cost function (3) so as to induce some sort of regularization of the factor estimates or so as to reflect prior belief, e.g., in Bayesian maximum a posteriori (MAP) estimation. When such penalty terms are separable in the columns of $\mathbf{H}$ or in the rows of $\mathbf{W}$, penalized NMF essentially amounts to solving the following optimization problem:

$$\min_{\mathbf{h}} C_P(\mathbf{h}) \overset{\text{def}}{=} D(\mathbf{v}|\mathbf{W}\mathbf{h}) + L(\mathbf{h}) \text{ subject to } \mathbf{h} \geq 0 \tag{71}$$

where $L(\mathbf{h})$ is the penalty term. An auxiliary function to $C_P(\mathbf{h})$ is readily given by

$$G_P(\mathbf{h}|\tilde{\mathbf{h}}) \overset{\text{def}}{=} G(\mathbf{h}|\tilde{\mathbf{h}}) + L(\mathbf{h}) \tag{72}$$

where $G(\mathbf{h}|\tilde{\mathbf{h}})$ is any auxiliary function to $C(\mathbf{h}) = D(\mathbf{v}|\mathbf{W}\mathbf{h})$. MM or ME algorithms can then be designed on a case-by-case basis. Let us consider a short example for illustration: ℓ_1-norm regularization. In that case we have

$$L(\mathbf{h}) = \lambda \sum_k h_k \tag{73}$$

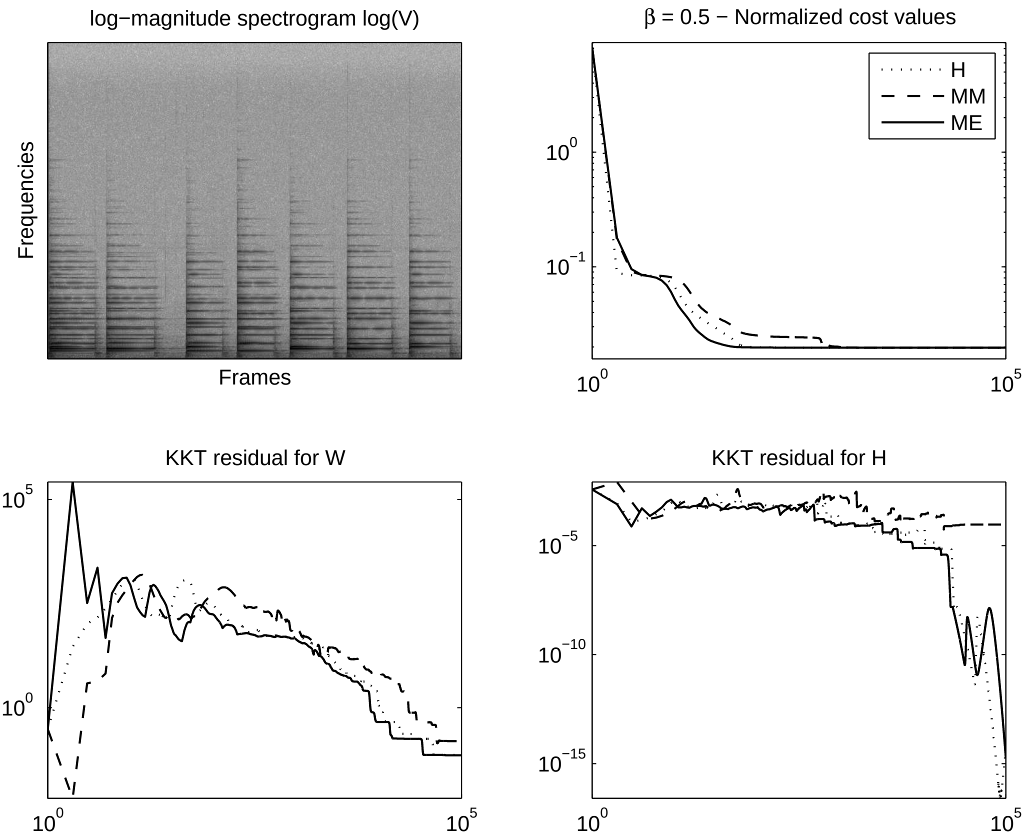


Figure 7: One run of the heuristic (H), MM and ME algorithms on the piano magnitude spectrogram with $\beta = 0.5$. Logarithmic scales for both x- and y- axes.

where λ is a positive weight parameter. Using the auxiliary function designed in Section 3.2 and Eq. (36), the gradient of the penalized auxiliary function writes

$$\nabla_{h_k} G_L(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f w_{fk} \left[\breve{d}' \left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) + \overset{\frown}{d}'(v_f|\tilde{v}_f) \right] + \lambda.$$

The MM algorithm for ℓ_1-regularized β-NMF takes a very simple form for $\beta \leq 1$, such that

$$h_k = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1} + \lambda} \right)^{\gamma(\beta)}. \tag{74}$$

This in particular leads to ℓ_1-regularized NMF algorithms for KL-NMF and IS-NMF with proven monotonicity. An update similar to Eq. (74) is obtained for $\beta \geq 2$ but the λ term appears through its sign opposite at the numerator, instead of appearing at the denominator. Hence the nonnegativity constraint may become active and must be treated carefully; in that case our result coincides with similar findings of Pauca et al. (2006); Mørup and Clemmensen (2007) for the specific case of ℓ_1-regularized NMF with the Euclidean distance ($\beta = 2$). In the case $\beta \in (1, 2)$ the MM algorithm does not come up with a simple closed-form update, which supports the fact in the penalized case handy algorithms may only come on a case-by-case basis. This is similar to Expectation-Maximization (EM) procedures for MAP estimation, in which the E-step is essentially unchanged but where the M-step might become intractable because of the penalty term. ME algorithms can also be designed for the ℓ_1-regularized problem and as a matter of fact it can be shown that the results of Table 3 (i.e., the values of β for which a closed-form update exists) still hold in that case.

Convex β-NMF In some recent NMF-related works the dictionary $\mathbf{W}$ is constrained to belong to a known subspace $\mathbf{S} \in \mathbb{R}_+^{F \times M}$ such that

$$\mathbf{W} = \mathbf{SL} \tag{75}$$

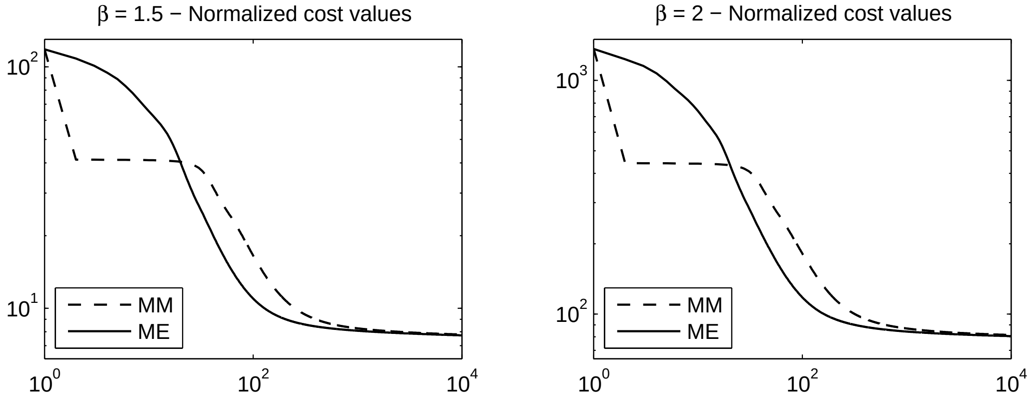
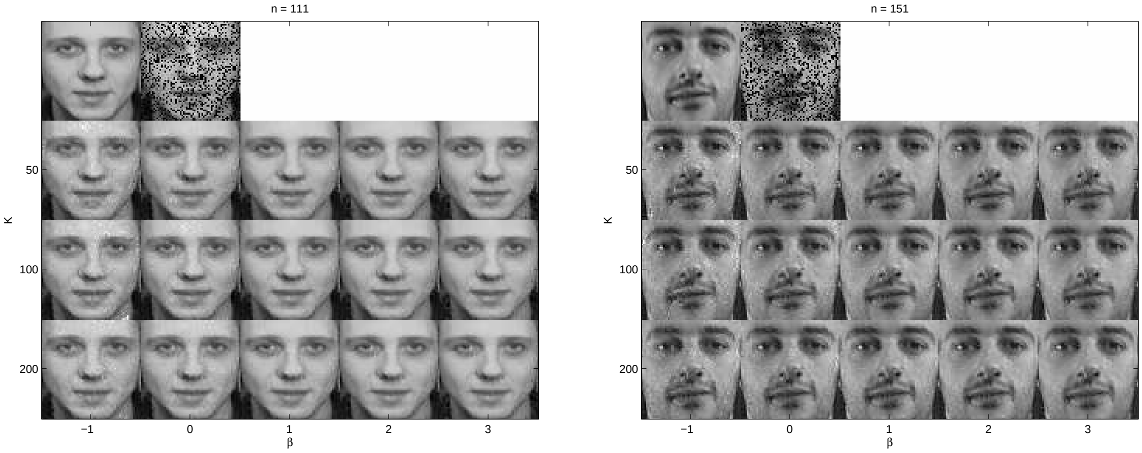


Figure 8: One run of the MM and ME algorithms on the Olivetti dataset with $\beta = 1.5$ (left) and $\beta = 2$ (right). Logarithmic scales for both x- and y- axes.



K / β	-1	0	1	2	3
50	28.3	30.7	32.2	31.8	31.2
100	26.6	30.6	32.5	33.2	32.3
200	29.1	31.2	32.4	32.7	32.1

K / β	-1	0	1	2	3
50	26.3	27.9	29.6	29.5	28.8
100	26.2	28.3	29.3	29.8	29.8
200	27.4	28.3	29.0	30.0	30.0

Figure 9: Interpolation results with the Olivetti dataset. Original and corrupted data are shown top left of each plot. Below are the reconstructions obtained for $K \in \{50, 100, 200\}$ and $\beta \in \{-1, 0, 1, 2, 3\}$. Tables report PSNRs (in dB) of the reconstructions.

where $\mathbf{L} \in \mathbb{R}_+^{M \times K}$. For example Ding et al. (2010) assume the columns of $\mathbf{W}$ to be linear combinations (with unknown expansion coefficients) of data points (columns of $\mathbf{V}$), so as to enforce the dictionary to be composed of *data centroids*, while Vincent et al. (2010) assume the dictionary element to be linear combinations of narrow band spectra, so as to enforce harmonicity and smoothness of the dictionary. The term "convex-NMF" was introduced by Ding et al. (2010) to express the idea that $\mathbf{W}$ belongs to the convex set of all nonnegative linear combinations of elements of $\mathbf{S}$, but this does not make the optimization problem convex in itself, in the general case.

In this setting, the dictionary update is tantamount to solving

$$\min_{\mathbf{L}} C_{cv}(\mathbf{L}) \overset{\text{def}}{=} D(\mathbf{V}|\mathbf{SLH}) = \sum_{fn} d\left(v_{fn}|\sum_{mk} s_{fm} l_{mk} h_{kn}\right) \quad \text{subject to } \mathbf{L} \geq 0. \tag{76}$$

As a matter of fact, this matricial optimization problem can be turned into vectorial nonnegative linear regression so that the results of Section 4 holds. Given some mappings $(f, n) \in \{1, F\} \times \{1, N\} \to p \in \{1, FN\}$ and $(m, k) \in \{1, M\} \times \{1, K\} \to q \in \{1, MK\}$ let us introduce the following variables: $\mathbf{T}$ is the matrix of dimension $FN \times MK$ with coefficients $t_{pq} = s_{fm} h_{kn}$, $\underline{\mathbf{v}}$ is the column vector of size FN with coefficients $\underline{v}_p = v_{fn}$, $\underline{\mathbf{l}}$ is the column vector of size MK with coefficients $\underline{l}_q = l_{mk}$. Then we have

$$D(\mathbf{V}|\mathbf{SLH}) = \sum_p d\left(\underline{v}_p | \sum_q t_{pq}\, \underline{l}_q\right) \tag{77}$$

and thus the estimation of $\mathbf{L}$ amounts to the approximation $\underline{\mathbf{v}} \approx \mathbf{T}\,\underline{\mathbf{l}}$. As such, any of the algorithms described in Section 4 can be employed for this task. As before, the resulting vectorial updates can be turned into matricial updates, leading to simple and efficient implementations. For example, the MM update reads

$$\mathbf{L} \leftarrow \mathbf{L}.\left(\frac{\mathbf{S}^T\left[(\mathbf{SLH})^{\cdot(\beta-2)}.\mathbf{V}\right]\mathbf{H}^T}{\mathbf{S}^T\left[(\mathbf{SLH})^{\cdot(\beta-1)}\right]\mathbf{H}^T}\right)^{\cdot\gamma(\beta)}. \tag{78}$$

This result proves the monotonicity of some of the algorithms derived heuristically in (Vincent et al., 2010) and also extends the results of (Ding et al., 2010) for convex NMF with the Euclidean distance to the more general β-divergence.[6]

7 Conclusions

This paper has addressed NMF with the β-divergence. The problem may be reduced to a mere nonnegative linear regression problem and our approach is based on the construction of an auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ which majorizes the objective function $C(\mathbf{h})$ everywhere and is tight for $\mathbf{h} = \tilde{\mathbf{h}}$. The auxiliary function unifies existing auxiliary functions for the Euclidean distance and the KL divergence (Lee and Seung, 2001), for the "generalized divergence" of Kompass (2007) (in essence the β-divergence on its convex part, i.e., $\beta \in [1, 2]$) and for the IS divergence (Cao et al., 1999). Various descent algorithms, free of tuning parameters, may then be derived from this auxiliary function. As such, the findings of this paper may be summarized as follows.

- The MM algorithm based on the described auxiliary function is shown to yield multiplicative algorithms for $\beta \in \mathbb{R}$, as described by Eq. (38). For $\beta \in [1, 2]$ (interval of values for which the β-divergence is convex), the MM algorithm coincides with the heuristic algorithm given by Eq. (43), as already known from Kompass (2007).

- In Section 4.2, we prove the monotonicity of the heuristic algorithm for $\beta \in (0, 1)$ by proving the inequality $G(\mathbf{h}^{\mathrm{H}}|\tilde{\mathbf{h}}) \leq G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}})$. Hence, aggregating the existing monotonicity results for $\beta = 0$ and $\beta \in [1, 2]$, it can now be claimed that the heuristic algorithm is monotone for $\beta \in [0, 2]$, which is the range of values of practical interest that has been considered in the literature.

- In Section 4.3, we introduced the concept of maximization-equalization (ME) algorithms. Such algorithms are exhibited for specific values of β, in particular for $\beta \in \{0, 0.5, 1.5, 2\}$ which are values of practical interest. For $\beta = 0$ (IS divergence) the ME algorithm coincides with the heuristic algorithm, whose monotonicity already holds from (Cao et al., 1999). For other values of β the ME algorithms are nonmultiplicative. For $\beta \in \{0.5, 1.5, 2\}$ they amount to solving polynomial equations of order 1 or 2. The result section has illustrated the faster convergence of the ME approach w.r.t to MM or heuristic, with equivalent complexity.

- Finally, in Section 6 we have considered variants of NMF with the β-divergence. We have explained how penalty terms may be handled in the auxiliary function setting; in particular we have presented simple multiplicative algorithms for ℓ_1 regularized KL or IS NMF. Then, we have shown how the algorithms constructed for plain NMF holds for convex-NMF, generalizing and proving the monotonicity of existing algorithms.

As for perspectives, the present work leaves two important questions unanswered. The first one is the monotonicity of the heuristic algorithm for $\beta \notin [0, 2]$. The monotonicity is observed in practice but we have not been able to come up with proofs in the presented setting. Either other approaches need to be followed or a different type of auxiliary functions than the one presented here needs to be envisaged. As suggested in Section 2.1, the convex-concave decomposition of the β-divergence is not unique and decompositions other than the "natural" one employed in this paper may lead to auxiliary functions that

[6] More precisely, Ding et al. (2010) consider a "semi"-NMF version where $\mathbf{S} = \mathbf{V}$ and the data is allowed to be real-valued while the nonnegativity constraint is solely imposed on $\mathbf{L}$ and $\mathbf{H}$; our results do not apply to this more general framework but only to the special case where $\mathbf{V}$ in nonnegative.

more closely fit to the criterion. The second, probably more ambitious question is the convergence of the sequence of iterates produced by the proposed algorithms a stationary point. Partial results exist for Euclidean NMF (Lin, 2007a), convergence of multiplicative rules for nonnegative linear regression (i.e., when only one of the two matrices is updated) has been studied in a few cases, see, e.g., (Titterington, 1987; De Pierro, 1993; Eggermont and LaRiccia, 1998), but general results for NMF with the β-divergence are still lacking. A noteworthy attempt has recently been made by Badeau et al. (2010), which points difficulties in the convergence study due to the inherent scale ambiguity of factorization models.

Finally, another relevant perspective is the design of new types of β-NMF algorithms. In the Euclidean case, projected gradient methods (Lin, 2007b), second-order active sets methods (Kim et al., 2008), block-coordinate descent methods (Mairal et al., 2010) have recently been shown to outperform standard multiplicative updates, see also (Mørup and Hansen, 2009) for a comparison of a selection of algorithms. As such it would be interesting to study how these approaches may extend to the more general β-NMF framework.

Acknowledgements

The authors would like to thank Francis Bach, Henri Lantéri, Augustin Lefèvre, Cédric Richard and Céline Theys for inspiring discussions related to optimization in NMF. Many thanks to the reviewers for helpful comments. This work is supported by project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization).

References

R. Badeau, N. Bertin, and E. Vincent. Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization. *IEEE Transactions on Neural Networks*, 21(12):1869–1881, Dec. 2010.

A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, Sep. 1998.

M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1): 155–173, Sep. 2007.

N. Bertin, C. Févotte, and R. Badeau. A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1545 –1548, Taipei, Taiwan, Apr. 2009. doi: 10.1109/ICASSP.2009.4959891. URL `http://www.tsi.enst.fr/~fevotte/Proceedings/icassp09b.pdf`.

J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. In *Proceedings of the National Academy of Sciences*, pages 4164–4169, Mar. 2004.

Y. Cao, P. P. B. Eggermont, and S. Terebey. Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing*, 8(2):286–292, Feb. 1999. doi: 10.1109/83.743861.

A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009(Article ID 785152):17 pages, 2009. doi:10.1155/2009/785152.

A. Cichocki and S. Amari. Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, June 2010.

A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, pages 32–39, Charleston SC, USA, Mar. 2006.

A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29(9):1433–1440, July 2008.

M. Daube-Witherspoon and G. Muehllehner. An iterative image space reconstruction algorthm suitable for volume ECT. *IEEE Transactions on Medical Imaging*, 5(5):61 – 66, 1986. doi: 10.1109/TMI.1986.4307748.

A. R. De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Medical Imaging*, 12(2):328–333, 1993. doi: 10.1109/42.232263.

A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proc. 11th International Society for Music Information Retrieval Conference (ISMIR'2010)*, 2010.

I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. *Advances in Neural Information Processing Systems (NIPS)*, 19, 2005.

C. H. Q. Ding, Tao Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45 – 55, 2010. doi: 10.1109/TPAMI.2008.277.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

K. Drakakis, S. Rickard, R. de Frein, and A. Cichocki. Analysis of financial data using non-negative matrix factorization. *International Journal of Mathematical Sciences*, 6(2), June 2007.

P. P B. Eggermont and V. N. LaRiccia. On EM-like algorithms for minimum distance estimation. http://www.udel.edu/FREC/eggermont/Preprints/emlike.pdf, 1998.

S. Eguchi and Y. Kano. Robustifying maximum likelihood estimation. Technical report, Institute of Statistical Mathematics, June 2001. URL `http://www.ism.ac.jp/$\sim$eguchi/pdf/Robustify_MLE.pdf`. Research Memo. 802.

C. Févotte and A. T. Cemgil. Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 1913–1917, Glasgow, Scotland, Aug. 2009. URL `http://www.tsi.enst.fr/~fevotte/Proceedings/eusipco09_1.pdf`.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, Mar. 2009. doi: 10.1162/neco.2008.04-08-771. URL `http://www.tsi.enst.fr/~fevotte/Journals/neco09_is-nmf.pdf`.

D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Proc. Irish Signals and Systems Conference*, 2009.

Y. Gao and G. Church. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21:3970–3975, 2005. doi: doi:10.1093/bioinformatics/bti653.

E. F. Gonzalez and Y. Zhang. Accelerating the Lee-Seung algorithm for non-negative matrix factorizations. Technical report, Rice University, 2005. URL `http://www.caam.rice.edu/caam/trs/2005/TR05-02.ps`.

D. Greene, G. Cagney, N. Krogan, and P. Cunningham. Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics*, 24(15):1722–1728, 2008.

R. Hennequin, R. Badeau, and B. David. NMF with time-frequency activations to model non stationary audio events. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10)*, pages 445–448, 2010. doi: 10.1109/ICASSP.2010.5495733.

Ngoc-Diep Ho. *Nonnegative matrix factorization algorithms and applications*. PhD thesis, Université Catholique de Louvain, 2008. URL `www.inma.ucl.ac.be/~vdooren/ThesisHo.pdf`.

D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58:30 – 37, 2004.

D. Kim, S. Sra, and I. S. Dhillon. Fast projection-based methods for the least squares nonnegative matrix approximation problem. *Statistical Analysis and Data Mining*, 1:38–51, 2008.

R. Kompass. A generalized divergence measure fon nonnegative matrix factorization. *Neural Computation*, 19(3):780–791, 2007.

H. Lantéri, M. Roche, O. Cuevas, and C. Aime. A general method to devise maximum-likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Processing*, 81(5):945–974, May 2001.

H. Lantéri, C. Theys, C. Richard, and C. Févotte. Split gradient method for nonnegative matrix factorization. In *Proc. 18th European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, Aug. 2010. URL `http://perso.telecom-paristech.fr/~fevotte/Proceedings/eusipco10.pdf`.

H. Laurberg, M. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. Theorems on positive data: On the uniqueness of NMF. *Computational Intelligence and Neuroscience*, Article ID 764206, 2008.

J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama. Computational auditory induction by missing-data non-negative matrix factorization. In *Proc. ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, pages 1–6, Sep. 2008.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

C.-J. Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18:1589–1596, 2007a.

C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19: 2756–2779, 2007b.

L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal*, 79:745–754, 1974. doi: 10.1086/111605.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:10–60, 2010.

M. Minami and S. Eguchi. Robust blind source separation by beta-divergence. *Neural Computation*, 14: 1859–1886, 2002.

M. Mørup and L. H. Clemmensen. Multiplicative updates for the LASSO. In *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP'07)*, 2007.

M. Mørup and L. K. Hansen. Tuning pruning in sparse non-negative matrix factorization. In *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*, Glasgow, Scotland, Aug. 2009.

M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama. Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP'2010)*, Sep. 2010.

P. D. O'Grady. *Sparse separation of under-determined speech mixtures.* PhD thesis, National University of Ireland Maynooth, 2007.

P. D. O'Grady and B. A. Pearlmutter. Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. *Neurocomputing*, 72(1-3):88 – 101, 2008.

V. P. Pauca, J. Piper, and R. J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and its Applications*, 416:29–47, 2006.

W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62:55–59, 1972.

R. Salakhutdinov and S. Roweis. Adaptive overrelaxed bound optimization methods. In *Proc. International Conference on Machine Learning (ICML)*, pages 664–671, 2003.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, Oct. 2003.

P. Smaragdis, B. Raj, and M. Shashanka. Missing data imputation for spectral audio signals. In *IEEE international workshop on Machine Learning for Signal Processing (MLSP)*, Grenoble, France, Sep. 2009.

D. M. Titterington. On the iterative image space reconstruction algorthm for ECT. *IEEE Trans. Medical Imaging*, 6(1):52–56, 1987. doi: 10.1109/TMI.1987.4307797.

E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech and Language Processing*, 18:528 – 537, 2010. doi: 10.1109/TASL.2009.2034186.