EUSIPCO 2013 1569739113

MULTIPLICATIVE UPDATES FOR MODELING MIXTURES OF NON-STATIONARY SIGNALS IN THE TIME-FREQUENCY DOMAIN

Roland Badeau[*]

Institut Mines-Telecom,
Telecom ParisTech, CNRS LTCI

Alexey Ozerov

Technicolor
Research & Innovation

ABSTRACT

We recently introduced the high-resolution nonnegative matrix factorization (HR-NMF) model for representing mixtures of non-stationary signals in the time-frequency domain, and we highlighted its capability to both reach a high spectral resolution and reconstruct high quality audio signals. An expectation-maximization (EM) algorithm was also proposed for estimating its parameters. In this paper, we replace the maximization step by multiplicative update rules (MUR), in order to improve the convergence rate. We also introduce general MUR that are not limited to nonnegative parameters, and we propose a new insight into the EM algorithm, which shows that MUR and EM actually belong to the same family. We thus introduce a continuum of algorithms between them. Experiments confirm that the proposed approach permits to overcome the convergence rate of the EM algorithm.

***Index Terms*—** Nonnegative Matrix Factorization, High Resolution methods, Expectation-Maximization algorithm, Multiplicative update rules.

1. INTRODUCTION

Nonnegative Matrix Factorization (NMF) [1] is a powerful tool for decomposing mixtures of non-stationary signals in the Time-Frequency (TF) domain. However, unlike the High Resolution (HR) methods [2] dedicated to mixtures of complex exponentials, its spectral resolution is limited by that of the underlying TF representation. Following previous works which aimed at providing a probabilistic framework for NMF [3–5], we introduced in [6,7] a unified probabilistic model called HR-NMF, that permits to overcome this limit by taking both phases and local correlations in each frequency band into account. It can be used with both complex-valued and real-valued TF representations, like the short-time Fourier transform (STFT) or the modified discrete cosine transform (MDCT). Moreover, we showed that HR-NMF generalizes some very popular models: the Itakura-Saito NMF model (IS-NMF) [5], autoregressive (AR) processes, and the Exponential Sinusoidal Model (ESM), commonly used in HR spectral analysis of time series [2]. In [6, 7], HR-NMF was estimated with the Expectation-Maximization (EM) algorithm. In this paper, we propose to replace the maximization step by MUR derived by following the strategy presented in [8]. Indeed, MUR are very popular in the NMF community [1, 9], and recent works showed that they tend to converge faster than EM-related algorithms [5]. We also introduce general MUR that are not limited to nonnegative parameters, and we propose a new insight into the EM algorithm, which shows that MUR and EM actually belong to the same family of algorithms. We thus introduce a parametric continuum of algorithms between those two ones. Finally, as a by-product we show that the expectation step can be used as a simple way of computing the gradient of the log-likelihood function. This general result is not limited to MUR and it applies to any kind of minorize-maximization (MM) optimization strategy. This paper is organized as follows: the principle of MUR is summarized in section 2, a new insight into the EM algorithm is proposed in section 3, the HR-NMF model is presented in section 4, and the MUR for estimating HR-NMF are introduced in section 5. Section 6 is devoted to experimental results, and conclusions are drawn in section 7. The following notation will be used throughout the paper:

- $\mathbf{0}$: vector whose entries are all equal to 0,
- $\mathbf{1}$: vector whose entries are all equal to 1,
- $\boldsymbol{M}^*$: conjugate of matrix (or vector) $\boldsymbol{M}$,
- $\boldsymbol{M}^\top$: transpose of matrix (or vector) $\boldsymbol{M}$,
- $\boldsymbol{M}^H$: conjugate transpose of matrix (or vector) $\boldsymbol{M}$,
- $[\boldsymbol{M};\boldsymbol{N}]$: vertical concatenation of $\boldsymbol{M}$ and $\boldsymbol{N}$,
- $\mathcal{N}_{\mathbb{F}}(\boldsymbol{\mu},\boldsymbol{R})$: real (if $\mathbb{F}$=$\mathbb{R}$) or circular complex (if $\mathbb{F}$=$\mathbb{C}$) normal distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{R}$.

2. MULTIPLICATIVE UPDATE RULES

2.1. Scalar MUR with nonnegative constraints

When minimizing a criterion $f(\boldsymbol{\theta}) \in \mathbb{R}$ with $\boldsymbol{\theta} \in \mathbb{R}_+^d$, suppose that the gradient of f is of the form $\boldsymbol{\nabla} f(\boldsymbol{\theta}) = \boldsymbol{p}(\boldsymbol{\theta}) - \boldsymbol{m}(\boldsymbol{\theta})$, where the entries of $\boldsymbol{p}(\boldsymbol{\theta})$ and $\boldsymbol{m}(\boldsymbol{\theta})$ are nonnegative.

[*]This work is supported by the French National Research Agency (ANR) as a part of the DReaM project (ANR-09-CORD-006-03) and partly supported by the Quaero Program, funded by OSEO.

Then the MUR for minimizing $f(\boldsymbol{\theta})$ is defined as:

$$\boldsymbol{\theta} \longleftarrow \left(\frac{\boldsymbol{m}(\boldsymbol{\theta})}{\boldsymbol{p}(\boldsymbol{\theta})}\right)^{\eta} \boldsymbol{\theta}, \tag{1}$$

where all mathematical operations are entrywise, and $\eta > 0$ is a step size parameter (the application to NMF is presented *e.g.* in [1, chap. 3]). In [8], we proved the local convergence of (1) to a local minimum of function f under mild assumptions[1].

2.2. Vector MUR without nonnegative constraints

When minimizing a criterion $f(\boldsymbol{\theta}) \in \mathbb{R}$ with $\boldsymbol{\theta} \in \mathbb{C}^d$, suppose that the (complex) gradient of f can be written in the form $\boldsymbol{\nabla} f(\boldsymbol{\theta}) = \boldsymbol{R}_p(\boldsymbol{\theta})\,\boldsymbol{\theta} - \boldsymbol{R}_m(\boldsymbol{\theta})\,\boldsymbol{\theta}$, where $\boldsymbol{R}_p(\boldsymbol{\theta})$ and $\boldsymbol{R}_m(\boldsymbol{\theta})$ are two positive definite matrices. Following our previous works [11] in this domain, we propose the following multiplicative update rule for minimizing $f(\boldsymbol{\theta})$:

$$\boldsymbol{\theta} \longleftarrow \left(\boldsymbol{R}_p(\boldsymbol{\theta})^{-1}\boldsymbol{R}_m(\boldsymbol{\theta})\right)^{\eta} \boldsymbol{\theta}, \tag{2}$$

where $\eta > 0$ is a step size parameter. Note that in equation (2), matrix $\boldsymbol{R}_p(\boldsymbol{\theta})^{-1}\boldsymbol{R}_m(\boldsymbol{\theta})$ is generally not Hermitian, but all its eigenvalues are always positive[2]. Therefore the exponentiation in equation (2) must be understood in the following way: given the eigenvalue decomposition $\boldsymbol{R}_p(\boldsymbol{\theta})^{-1}\boldsymbol{R}_m(\boldsymbol{\theta}) = \boldsymbol{G}\boldsymbol{\Lambda}\boldsymbol{G}^{-1}$, where $\boldsymbol{G}$ is non-singular and $\boldsymbol{\Lambda}$ is diagonal with positive diagonal coefficients, we define $\left(\boldsymbol{R}_p(\boldsymbol{\theta})^{-1}\boldsymbol{R}_m(\boldsymbol{\theta})\right)^{\eta} = \boldsymbol{G}\boldsymbol{\Lambda}^{\eta}\boldsymbol{G}^{-1}$, where $\boldsymbol{\Lambda}^{\eta}$ also has positive diagonal coefficients. Besides, when this MUR algorithm converges, its limit point is necessarily a fixed point of (2) (which can be proved[3] to be a stationary point of f).

3. A NEW INSIGHT INTO THE EM ALGORITHM

The EM algorithm is an iterative method for finding the maximum likelihood estimate of the parameter $\boldsymbol{\theta}$ of a probabilistic model involving both observed variables x and latent variables c. It consists of two steps called Expectation (E-step) and Maximization (M-step):

- E-step: evaluate the a posteriori distribution $p(c|x;\boldsymbol{\theta})$;
- M-step: $\boldsymbol{\theta} \longleftarrow \underset{\boldsymbol{\theta}'}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}', \boldsymbol{\theta})$, where

$$Q(\boldsymbol{\theta}', \boldsymbol{\theta}) = \int \ln(p(x, c; \boldsymbol{\theta}'))p(c|x;\boldsymbol{\theta})dc. \tag{3}$$

The EM algorithm is proved to increase the log-likelihood $L(\boldsymbol{\theta}) = \ln(p(x;\boldsymbol{\theta}))$ of the observed data x at each iteration. If parameter $\boldsymbol{\theta}'$ belongs to a Hilbert space $\mathcal{H}$, and if both functions $L(\boldsymbol{\theta}')$ and $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$ are Gâteaux differentiable w.r.t. $\boldsymbol{\theta}'$, Proposition 1 shows how the E-step can be used for efficiently computing the gradient of L, for any value of the parameter $\boldsymbol{\theta}$. It has indeed to be applied to $f(\boldsymbol{\theta}') = -L(\boldsymbol{\theta}')$, and $g(\boldsymbol{\theta}', \boldsymbol{\theta}) = -L(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}) - Q(\boldsymbol{\theta}', \boldsymbol{\theta})$.

Proposition 1. *Let $f(.)$ and $g(.,.)$ be two real-valued functions defined in a Hilbert space $\mathcal{H}$, such that for a given $\boldsymbol{\theta} \in \mathcal{H}$, $f(\boldsymbol{\theta}')$ and $g(\boldsymbol{\theta}', \boldsymbol{\theta})$ are Gâteaux differentiable w.r.t. $\boldsymbol{\theta}' \in \mathcal{H}$, $\forall \boldsymbol{\theta}' \in \mathcal{H}$, $g(\boldsymbol{\theta}', \boldsymbol{\theta}) \geq f(\boldsymbol{\theta}')$, and $g(\boldsymbol{\theta}, \boldsymbol{\theta}) = f(\boldsymbol{\theta})$. Then the gradients w.r.t. $\boldsymbol{\theta}'$ are such that $\boldsymbol{\nabla} g(\boldsymbol{\theta}, \boldsymbol{\theta}) = \boldsymbol{\nabla} f(\boldsymbol{\theta})$.*

Proposition 1 is proved by comparing the first order Taylor series expansions of functions f and g, and it applies to any kind of minorize-maximization (MM) optimization strategy. Considering the particular framework of the EM algorithm, we conclude that $\forall \boldsymbol{\theta} \in \mathcal{H}$, the gradient $\boldsymbol{\nabla} L(\boldsymbol{\theta})$ can be numerically computed as $\boldsymbol{\nabla} Q(\boldsymbol{\theta}, \boldsymbol{\theta})$ (whose closed-form expression is generally easier to obtain). This observation paves the way for gradient-based optimization techniques (such as the MUR) for directly maximizing L, which will replace the M-step in the EM algorithm.

4. HR-NMF TIME-FREQUENCY MIXTURE MODEL

4.1. Definition

The HR-NMF mixture model of TF data $x(f,t) \in \mathbb{F}$ (where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) is defined for all frequencies $1 \leq f \leq F$ and times $1 \leq t \leq T$ as the sum of K latent components $c_k(f,t) \in \mathbb{F}$ plus a white noise $n(f,t) \sim \mathcal{N}_{\mathbb{F}}(0, \sigma^2)$:

$$x(f,t) = n(f,t) + \sum_{k=1}^{K} c_k(f,t), \tag{4}$$

where $c_k(f,t) = \sum_{p=1}^{P(k,f)} a(p,k,f)\, c_k(f, t-p) + b_k(f,t)$ is obtained by autoregressive filtering of a non-stationary signal $b_k(f,t) \in \mathbb{F}$ (where $a(p,k,f) \in \mathbb{F}$ and $P(k,f) \in \mathbb{N}$ is such that $a(P(k,f),k,f) \neq 0$), $b_k(f,t) \sim \mathcal{N}_{\mathbb{F}}(0, v_k(f,t))$ where $v_k(f,t)$ is defined as $v_k(f,t) = w(k,f)h(k,t)$ with $w(k,f) \geq 0$, $h(k,t) \geq 0$, and processes n and $b_1 \dots b_K$ are mutually independent. Moreover, $\forall k, f$, the random vectors $\boldsymbol{c}_k(f,0) = [c_k(f,0); \dots; c_k(f, -P(k,f)+1)]$ are assumed to be independent and distributed according to the prior distribution $\boldsymbol{c}_k(f,0) \sim \mathcal{N}_{\mathbb{F}}(\boldsymbol{\mu}_k(f), \boldsymbol{Q}_k(f)^{-1})$, where the mean $\boldsymbol{\mu}_k(f)$ and the precision matrix $\boldsymbol{Q}_k(f)$ are fixed parameters[4]. Lastly, we assume that $\forall f \in \{1 \dots F\}$, $\forall t \leq 0$, $x(f,t)$ are unobserved. The parameters $\boldsymbol{\theta}$ to be estimated are σ^2, $a(p,k,f)$, $w(k,f)$, and $h(k,t)$.

[1] Note that a number of convergence analyses presented in the literature only focus on the decrease of the cost function f (see *e.g.* [10]). However, this property is not sufficient (nor necessary) to prove the convergence of the algorithm to a local minimum of f. In [8], we proved the local convergence of (1), and we also provided some examples where the cost function is not decreasing, but the algorithm still converges to a local minimum.

[2] Indeed, if λ and $\boldsymbol{u} \neq \boldsymbol{0}$ are such that $\boldsymbol{R}_p(\boldsymbol{\theta})^{-1}\boldsymbol{R}_m(\boldsymbol{\theta})\boldsymbol{u} = \lambda\boldsymbol{u}$, then $\boldsymbol{R}_m(\boldsymbol{\theta})\boldsymbol{u} = \lambda\boldsymbol{R}_p(\boldsymbol{\theta})\boldsymbol{u}$ and $\lambda = \frac{\boldsymbol{u}^H\boldsymbol{R}_m(\boldsymbol{\theta})\boldsymbol{u}}{\boldsymbol{u}^H\boldsymbol{R}_p(\boldsymbol{\theta})\boldsymbol{u}} > 0$.

[3] If $\boldsymbol{\theta}$ is such a fixed point, then equation (2) shows that either $\boldsymbol{\theta} = \boldsymbol{0}$, or $\boldsymbol{\theta} \neq \boldsymbol{0}$ is an eigenvector of matrix $\left(\boldsymbol{R}_p(\boldsymbol{\theta})^{-1}\boldsymbol{R}_m(\boldsymbol{\theta})\right)^{\eta}$ with eigenvalue 1, thus it is also an eigenvector of matrix $\boldsymbol{R}_p(\boldsymbol{\theta})^{-1}\boldsymbol{R}_m(\boldsymbol{\theta})$ with eigenvalue 1, therefore $\boldsymbol{R}_m(\boldsymbol{\theta})\,\boldsymbol{\theta} = \boldsymbol{R}_p(\boldsymbol{\theta})\,\boldsymbol{\theta}$. In both cases, $\boldsymbol{\nabla} f(\boldsymbol{\theta}) = \boldsymbol{0}$.

[4] In practice we choose $\boldsymbol{\mu}_k(f) = \boldsymbol{0}$ and $\boldsymbol{Q}_k(f)^{-1} = \xi\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix and ξ is small relative to 1, in order to both enforce the causality of the latent components and avoid singular matrices.

It can be noticed that if $\sigma^2 = 0$ and $\forall k, f, P(k,f) = 0$, equation(4) becomes $x(f,t) = \sum_{k=1}^{K} b_k(f,t)$, thus $x(f,t) \sim \mathcal{N}_{\mathbb{F}}(0, \widehat{V}_{ft})$, where $\widehat{\boldsymbol{V}}$ is defined by the NMF $\widehat{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{H}$ with $W_{fk} = w(k,f)$ and $H_{kt} = h(k,t)$. The maximum likelihood estimation of $\boldsymbol{W}$ and $\boldsymbol{H}$ is then equivalent to the minimization of the IS-divergence between the matrix model $\widehat{\boldsymbol{V}}$ and the spectrogram $\boldsymbol{V}$ (where $V_{ft} = |x(f,t)|^2$). We conclude that this IS-NMF model [5] is a particular case of HR-NMF.

4.2. EM algorithm for estimating the HR-NMF model

In order to estimate the HR-NMF model parameters, the EM algorithm is applied to the observed data x and the latent components $c = \{c_1 \dots c_K\}$ (here the complete data is $\{x, c\}$). In order to handle the case of missing data, we define $\delta(f,t) = 1$ if $x(f,t)$ is observed, and $\delta(f,t) = 0$ else.

4.2.1. Definitions and useful inequalities

We first introduce some definitions that will be used in sections 4.2.3 and 5. Let $\boldsymbol{a}(k,f) = [1; -a(1,k,f) \dots - a(P(k,f),k,f)]$, and for any function $f(\boldsymbol{c})$, let $\langle f(\boldsymbol{c})\rangle$ denote the a posteriori mean of $f(\boldsymbol{c})$:

- $\boldsymbol{\Gamma}_k(f,t)$ is the prior covariance matrix of $\boldsymbol{c}_k(f,t) = [c_k(f,t); \dots; c_k(f,t-P(k,f)+1)]$, $\boldsymbol{m}_k(f,t) = \langle \boldsymbol{c}_k(f,t)\rangle$ is its posterior mean, $\tilde{\boldsymbol{\Gamma}}_k(f,t)$ is its posterior covariance matrix, while $\tilde{\boldsymbol{\Gamma}}(f,t)$ is the posterior covariance matrix of vector $\boldsymbol{c}(f,t) = [c_1(f,t); \dots; c_K(f,t)]$;
- $v_k(f,t) = w(k,f)h(k,t)$ is the prior variance of $b_k(f,t)$, $\langle b_k(f,t)\rangle = \boldsymbol{m}_k(f,t)^\top \boldsymbol{a}(k,f)$ is its posterior mean, and $\tilde{v}_k(f,t) = \boldsymbol{a}(k,f)^H \tilde{\boldsymbol{\Gamma}}_k(f,t)^* \boldsymbol{a}(k,f)$ is its posterior variance;
- σ^2 is the prior variance of $n(f,t)$, $\langle n(f,t)\rangle = x(f,t) - \mathbf{1}^\top \langle \boldsymbol{c}(f,t)\rangle$ is its posterior mean, and $\tilde{\sigma}(f,t)^2 = \mathbf{1}^\top \tilde{\boldsymbol{\Gamma}}(f,t)\mathbf{1}$ is its posterior variance.

We can now introduce the following inequalities[5][6], that permit to guarantee the positiveness of the terms involved in the MUR that will be introduced in section 5:

$$\tilde{\boldsymbol{\Gamma}}_k(f,t) \leq \boldsymbol{\Gamma}_k(f,t), \tag{5}$$

$$\tilde{v}_k(f,t) \leq v_k(f,t), \tag{6}$$

$$\tilde{\sigma}(f,t)^2 \leq \sigma^2. \tag{7}$$

4.2.2. Expectation step

The purpose of the E-step is to determine $p(c|x)$. A recursive implementation was presented in [6,7], based on Kalman filtering and smoothing. Its overall computational complexity is $O(FTK^3(1+P)^3)$, where $P = \max_{k,f} P(k,f)$.

4.2.3. Maximization step

The criterion Q defined in equation (3) can be written in the form $Q = \langle \ln(p(x,c))\rangle = \langle \ln(p(x|c_1 \dots c_K))\rangle + \sum_{k=1}^{K} \langle \ln(p(c_k))\rangle$. It is thus equal, up to an additive and a multiplicative constant, to $Q_0 + \sum_{k=1}^{K} Q_k$, with

$$Q_0 = -\sum_{f=1}^{F}\sum_{t=1}^{T} \delta(f,t)\ln(\sigma^2) + e(f,t)/\sigma^2, \tag{8}$$

$$Q_k = -\sum_{f=1}^{F}\sum_{t=1}^{T} \ln(w(k,f)h(k,t)) + \frac{\boldsymbol{a}(k,f)^H \boldsymbol{S}(k,f,t)\,\boldsymbol{a}(k,f)}{w(k,f)h(k,t)}, \tag{9}$$

where

$$e(f,t) = \delta(f,t)\left(|\langle n(f,t)\rangle|^2 + \tilde{\sigma}(f,t)^2\right), \tag{10}$$

$$\boldsymbol{S}(k,f,t) = \boldsymbol{m}_k(f,t)^* \boldsymbol{m}_k(f,t)^\top + \tilde{\boldsymbol{\Gamma}}_k(f,t)^*, \tag{11}$$

$$\boldsymbol{a}(k,f)^H \boldsymbol{S}(k,f,t) \boldsymbol{a}(k,f) = |\langle b_k(f,t)\rangle|^2 + \tilde{v}_k(f,t). \tag{12}$$

Maximizing Q is thus equivalent to independently maximizing Q_0 w.r.t. σ^2 and each Q_k w.r.t. $h(k,t)$, $w(k,f)$ and $\boldsymbol{a}(k,f)$. The full mathematical derivation of the M-step has been provided in [6,7]. Its complexity is $O(FTK(1+P)^3)$.

5. MUR FOR HR-NMF

The structure of the proposed MUR algorithm is the same as that of the algorithm used for initializing EM in [7]:

repeat
- E-step (as in section 4.2.2)
- Multiplicative update of σ^2 (equation (13))
- $\forall k, f$, multiplicative update of $w(k,f)$ (equation (15))
- $\forall k, f$, multiplicative update of $\boldsymbol{a}(k,f)$ (equation (16))
- E-step (as in section 4.2.2)
- Multiplicative update of σ^2 (equation (13))
- $\forall k, t$, multiplicative update of $h(k,t)$ (equation (14))
- Normalization of the NMF [7]

until some stopping criterion is satisfied

In the MUR introduced below in this section, we use a step size $\eta > 0$ and a parameter $\varepsilon \geq 0$. The case $\varepsilon = 1$ and $\eta = 1$ will lead to the same update rules as those involved in the M-step of the EM algorithm, resulting in a generalized EM (GEM) algorithm[8]. However, the case $\varepsilon = 0$, $\eta = 1$, $\sigma^2 = 0$, $\forall k, f, t$, $P(k,f) = 0$ and $\delta(f,t) = 1$ will lead to standard IS-NMF MUR [1,5,12]. In practice, parameters ε and η have to be chosen empirically. Parameter ε must be nonnegative in order to guarantee the positiveness of the terms involved in the MUR; we will show in section 6 that the lower

[5] These inequalities rely on the fact that a posteriori variances are always lower than or equal to a priori variances.

[6] Equation (5) means that $\boldsymbol{\Gamma}_k(f,t) - \tilde{\boldsymbol{\Gamma}}_k(f,t)$ is positive semidefinite.

[7] As in any NMF MUR algorithm, w and h are normalized in order to remove the trivial scale indeterminacies of the NMF (*cf.* [5–7]).

[8] The resulting algorithm is not an exact EM algorithm, because the MUR algorithm performs two E-steps instead of one per iteration, and the EM algorithm in [6,7] performs several updates of $w(k,f)$, $\boldsymbol{a}(k,f)$ and $h(k,t)$ instead of (at most) one per iteration.

ε, the faster the convergence, but a too small value of ε yields numerical stability problems. Parameter η must lie in an interval of the form $]0, \eta_0[$ in order to guarantee the convergence of the MUR. In the case of IS-NMF, we have shown in [8] that $\eta_0 = 2$. It is possible to use different values of ε and η for the various parameters, and to make ε and η vary over the iterations.

5.1. Update of σ^2, $w(k,f)$ and $h(k,t)$

With a particular ε-parametrized decomposition of the gradient of (8), equations (1), (8) and (10) yield

$$\sigma^2 \longleftarrow \sigma^2 \left(\frac{\sum_{f=1}^{F} \sum_{t=1}^{T} \delta(f,t) \left(|\langle n(f,t) \rangle|^2 + \varepsilon \tilde{\sigma}(f,t)^2 \right)}{\sum_{f=1}^{F} \sum_{t=1}^{T} \delta(f,t) \left(\sigma^2 - \tilde{\sigma}(f,t)^2 + \varepsilon \tilde{\sigma}(f,t)^2 \right)} \right)^{\eta}, \tag{13}$$

whose non-negativity is guaranteed by equation (7).

In the same way, equations (1), (9) and (12) yield

$$h(k,t) \longleftarrow h(k,t) \left(\frac{\frac{1}{F} \sum_{f=1}^{F} \frac{|\langle b_k(f,t) \rangle|^2}{w(k,f)} + \varepsilon \tilde{h}(k,t)}{h(k,t) - \tilde{h}(k,t) + \varepsilon \tilde{h}(k,t)} \right)^{\eta}, \tag{14}$$

$$w(k,f) \longleftarrow w(k,f) \left(\frac{\frac{1}{T} \sum_{t=1}^{T} \frac{|\langle b_k(f,t) \rangle|^2}{h(k,t)} + \varepsilon \tilde{w}(k,f)}{w(k,f) - \tilde{w}(k,f) + \varepsilon \tilde{w}(k,f)} \right)^{\eta}, \tag{15}$$

where $\tilde{h}(k,t) = \frac{1}{F} \sum_{f=1}^{F} \frac{\tilde{v}_k(f,t)}{w(k,f)}$ and $\tilde{w}(k,f) = \frac{1}{T} \sum_{t=1}^{T} \frac{\tilde{v}_k(f,t)}{h(k,t)}$. The non-negativity of (14) and (15) is guaranteed by equation (6), which shows that $\tilde{h}(k,t) \leq h(k,t)$ and $\tilde{w}(k,f) \leq w(k,f)$. We recover two well-known particular cases: If $\varepsilon = 1$ and $\eta = 1$, equations (13), (14), and (15) become the M-step updates (16), (17) and (18) in [7]. If $\varepsilon = 0, \eta = 1$, and $\forall k, f$, $P(k,f) = 0$, then equations (13), (14), and (15), become the same as the MUR (59), (60) and (61) in [7], which further reduce to the standard MUR for IS-NMF [1, 5, 12] when $\sigma^2 = 0$ and $\forall f, t, \delta(f,t) = 1$ [9].

5.2. Update of $a(k,f)$

Let $\mathbf{\Gamma}_k(f) = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbf{\Gamma}_k(f,t)}{h(k,t)}$ and $\tilde{\mathbf{\Gamma}}_k(f) = \frac{1}{T} \sum_{t=1}^{T} \frac{\tilde{\mathbf{\Gamma}}_k(f,t)}{h(k,t)}$. By applying equation (2) to the parameter $\boldsymbol{\theta} = \frac{\boldsymbol{a}(k,f)}{\sqrt{w(k,f)}}$, straightforward calculations show that equations (9) and (11) yield

$$\boldsymbol{a}(k,f) \propto \left(\left(\frac{1}{T} \sum_{t=1}^{T} \frac{\boldsymbol{m}_k(f,t)^* \boldsymbol{m}_k(f,t)^\top}{h(k,t)} + \varepsilon \tilde{\mathbf{\Gamma}}_k(f)^* \right)^{-1} \left(\mathbf{\Gamma}_k(f) - \tilde{\mathbf{\Gamma}}_k(f) + \varepsilon \tilde{\mathbf{\Gamma}}_k(f) \right)^* \right)^{\eta} \boldsymbol{a}(k,f), \tag{16}$$

where the positive definiteness is guaranteed by equation (5). In equation (16), the symbol $\propto$ means that $\boldsymbol{a}(k,f)$ has to be multiplied by a complex number so that its first coefficient becomes 1. If $\varepsilon = 1$ and $\eta = 1$, equation (16) leads to the M-step update of $\boldsymbol{a}(k,f)$ as described in [6]. Note that this algorithm requires computing $\mathbf{\Gamma}_k(f)$. If $\eta = 1$, this calculation can be avoided since $\mathbf{\Gamma}_k(f)^* \boldsymbol{a}(k,f) = w(k,f)\boldsymbol{e}$, where $\boldsymbol{e}$ is the first column of the identity matrix. Otherwise $\mathbf{\Gamma}_k(f)$ can be computed during the E-step, within the Kalman filtering (by means of the predict phase, and by skipping both the update and the smoothing phases in [6,7]).

6. SIMULATION RESULTS

6.1. Synthetic data

We first evaluate the proposed algorithm on synthetic data generated from the HR-NMF model itself. For each of 20 independent trials the following was done: the model parameters (*i.e.* σ^2, h, w, and a) were first randomly drawn, with $K = 2, F = 3, T = 20, \delta(f,t) = 1$ and $P(k,f) = 3$. Given these model parameters, an observation sequence x was generated from the corresponding HR-NMF model. 200 iterations of four versions of the proposed MUR were then run with $\eta = 1$ [10]. Two versions were implemented with a constant value of ε, respectively equal to 1 (which corresponds to the EM algorithm) and 0.2 [11]. Two other versions were implemented with ε varying linearly from 0.2 to 1 and from 1 to 0.2. The resulting likelihoods averaged over the trials are plotted in Fig. 1 as functions of the iteration number. We see that during the first 100 - 120 iterations, the MUR with $\varepsilon = 0.2$ converge faster than the EM algorithm. Nevertheless, the final likelihood value for these MUR is smaller than that of EM. We believe that this can be explained by the fact that MUR with a small ε have a very aggressive behavior (*e.g.* by setting some parameters to zero), which allows increasing the likelihood very fast, but leads at the end to a suboptimal local maximum. Thus, this strategy can be interesting if the objective is to increase the likelihood in only a few iterations. However we see that starting with $\varepsilon = 1$ and decreasing it to $\varepsilon = 0.2$ leads systematically to a higher likelihood than EM.

[9] In this case the E-step amounts to updating the NMF $v(f,t) = \sum_{k=1}^{K} w(k,f)h(k,t)$ (more precisely, we get $\langle b_k(f,t) \rangle = \frac{v_k(f,t)}{v(f,t)} x(f,t)$ and $\tilde{v}_k(f,t) = \frac{v_k(f,t)(v(f,t) - v_k(f,t))}{v(f,t)}$).

[10] During some preliminary experiments over a range of ε values we have found that $\eta = 1$ is a good compromise between convergence speed and algorithmic stability.

[11] While the proposed implementation is mathematically strictly equivalent to standard MUR for IS-NMF when $\varepsilon = 0$, and $P(k,f) = 0$, we observed that it actually exhibits some numerical instabilities for small values of ε. This is why we have chosen $\varepsilon = 0.2$ in this experiment.

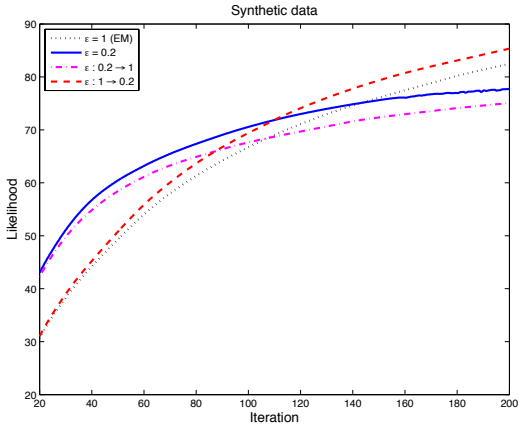

Fig. 1. Comparison of convergence rates with synthetic data.

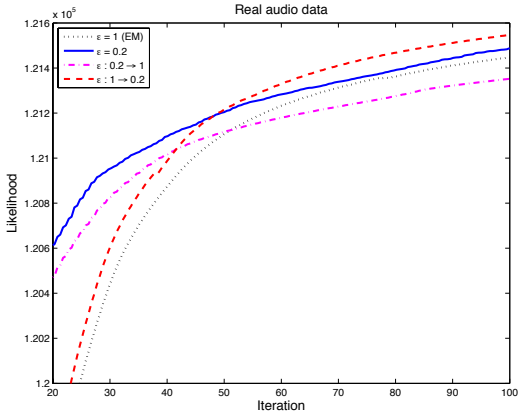

Fig. 2. Comparison of convergence rates with real audio data.

6.2. Real audio data

We then performed a similar experiment on real audio data. We took the same audio example as that presented in [6, 7]: a mixture of two piano notes played in octave relationship, and we applied the proposed MUR to the same fully observed STFT $x(f,t)$, with the same model dimensions and $K = 2$. In a first stage, parameters h and w were initialized randomly and estimated from the data using 30 iterations of the MUR algorithm described in [7]. Then the AR filters a were initialized to identity, and all parameters were jointly estimated by the same four versions of MUR as those tested for the synthetic data. The results plotted in figure 2 show the same qualitative behavior of the algorithms as in the synthetic data case.

7. CONCLUSIONS

In this paper, we introduced some novel results regarding MUR and the EM algorithm: we proposed general MUR for vector parameters without nonnegative constraints, and we presented a new insight into the EM algorithm, designed for estimating probabilistic models involving both observed and latent variables. In particular, we showed that the E-step of the EM algorithm permits to easily implement gradient-based methods, such as MUR, for directly maximizing the log-likelihood. We then applied this approach to the recently introduced HR-NMF model, which generalizes the popular IS-NMF model. We thus introduced a new family of algorithms, which encompasses both the GEM algorithm and MUR. Further research will include a theoretical study of the convergence and monotonicity of the proposed MUR, and the introduction of similar algorithms for other models, *e.g.* for multichannel NMF [9].

8. REFERENCES

[1] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, Sept. 2009.

[2] M. H. Hayes, *Statistical Digital Signal Processing And Modeling*, Wiley, Aug. 2009.

[3] P. Smaragdis, "Probabilistic decompositions of spectra for sound separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds., pp. 365–386. Springer, 2007.

[4] T. Virtanen, A. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. of IEEE ICASSP*, Apr. 2008, pp. 1825–1828.

[5] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[6] R. Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *Proc. of IEEE WASPAA*, Oct. 2011, pp. 253–256.

[7] R. Badeau, "High resolution NMF for modeling mixtures of non-stationary signals in the time-frequency domain," Tech. Rep. 2012D004, Télécom ParisTech, Paris, France, July 2012.

[8] R. Badeau, N. Bertin, and E. Vincent, "Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 21, no. 12, pp. 1869–1881, Dec. 2010.

[9] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "New formulations and efficient algorithms for multichannel NMF," in *Proc. of IEEE WASPAA*, Oct. 2011, pp. 153–156.

[10] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. of IEEE MLSP*, Aug. 2010, pp. 283–288.

[11] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non-stationary audio events," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 744–753, May 2011.

[12] P. Eggermont and V. LaRiccia, "On EM-like algorithms for minimum distance estimation," Mathematical sciences, University of Delaware, Mar. 1998.