Computational Auditory Induction by Missing-Data Non-Negative Matrix Factorization

Jonathan Le Roux[1,3], *Hirokazu Kameoka*[2], *Nobutaka Ono*[1], *Alain de Cheveigné*[3] *and Shigeki Sagayama*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[3]CNRS, Université Paris 5, and Ecole Normale Supérieure, France
{leroux,onono,sagayama}@hil.t.u-tokyo.ac.jp,
kameoka@eye.brl.ntt.co.jp, Alain.de.Cheveigne@ens.fr

Abstract

The human auditory system has the ability, known as auditory induction, to estimate the missing parts of a continuous auditory stream briefly covered by noise and perceptually resynthesize them. Humans are thus able to simultaneously analyze an auditory scene and reconstruct the underlying signal. In this article, we formulate this ability as a non-negative matrix factorization (NMF) problem with unobserved data, and show how to solve it using an auxiliary function method. We explain how this method can also be generally related to the EM algorithm, enabling the use of prior distributions on the parameters. We show how sparseness is a key to global feature extraction, and that our method is ideally able to extract patterns which never occur completely. We finally illustrate on an example how our method is able to simultaneously analyze a scene and interpolate the gaps into it.

Index Terms: Auditory induction, Bregman divergence, Missing-data, Auxiliary function, EM algorithm, Non-negative matrix factorization

1. Introduction

Computational Auditory Scene Analysis (CASA), whose main goal is to make computers able to imitate the human auditory segregation abilities, has been an area of intensive research in the recent years. A particular attention has been given to the so-called "cocktail party problem", the computational implementation of the "cocktail party effect", i.e., the ability of the human auditory system to focus on a single talker among a mixture of conversations and background noises, leading to the development of many methods for multi-pitch estimation, noise canceling or source separation [1]. Less emphasis has been put on the computational realization of another remarkable ability of the human auditory system, auditory induction. Humans are able, under certain conditions, to estimate the missing parts of a continuous acoustic stream briefly covered by noise, to perceptually resynthesize and clearly hear them [2]. Humans are thus able to simultaneously analyze an auditory scene, as in the cocktail party effect, in the presence of gaps, and to reconstruct the signal inside those gaps.

The development of an effective computational counterpart to this ability would lead to many important engineering applications, from polyphonic music recording analysis and restoration to mobile communications robust to both packet-loss and background noise.

This paper aims at developing such a computational counterpart to auditory induction, by simultaneously performing a decomposition of the magnitude wavelet spectrogram of a sound with missing or corrupted samples, and filling in the gaps into that spectrogram. Various approaches have emerged recently which attempt to analyze the structure of the spectrogram of an acoustical scene [3, 4], while on the other side gap interpolation techniques have been the subject of research for many years [5, 6]. However, only few models [7] so far try to deal with both issues. While [7] relies on local regularities of the spectrogram, the framework we introduce can use both local and global regularities.

We use here the non-negative matrix factor 2-D deconvolution algorithm (NMF2D) [3] as a representative method to extract global structures in audio spectrograms, and show how to extend it such that it can be used to analyze acoustic scenes with incomplete data. We also show how the introduced method can be interpreted in terms of the EM algorithm, enabling the use of prior distributions on the parameters which can enforce local smoothness regularities. Finally, we perform a short experimental evaluation to validate our method on a piece of computer generated music.

2. Audio inpainting

2.1. Problem setting

We consider the problem of interpolating gaps in audio signals by filling in the gaps in the magnitude spectrogram. We will not consider here the reconstruction of the phase. If the magnitude spectrogram can be accurately reconstructed, other methods could be used to obtain a phase consistent with it [8, 9]. We are interested in using local and global regularities in the spectrogram to simultaneously analyze the acoustical scene and fill in gaps that may have occurred into it. We believe that this is close to what is performed by humans in the auditory induction mechanism, when for example sounds actually missing from a speech signal can in certain conditions be perceptually synthesized by the brain and clearly heard [2]. Our goal is thus to "inpaint" the missing regions of the spectrogram based on global and local regularities, in the same spirit as is done for image inpainting [10], where diffusion-based (local) and exemplar-based (global) techniques are used [11].

As a method able to extract global patterns in non-negative data, we consider here the non-negative matrix factorization (NMF) framework. More precisely, we will use Schmidt and

Mørup's NMF2D algorithm [3, 12], which is well-suited to deal with audio signals. Local smoothness regularities can be enforced by adding prior distributions on the parameters, as explained in Section 3.

2.2. Overview of the NMF2D algorithm

The NMF2D algorithm is an extension of Smaragdis's non-negative matrix factor deconvolution (NMFD) [13], itself an extension of the original NMF [14]. NMF is a general tool which attempts to decompose a non-negative matrix $V \in \mathbb{R}^{\geq 0, M \times N}$ in the product of two usually lower-rank non-negative matrices $W \in \mathbb{R}^{\geq 0, M \times R}$ and $H \in \mathbb{R}^{\geq 0, R \times N}$,

$$V \approx WH.$$

In applications to audio, the horizontal and vertical dimensions of the matrices respectively represent time and frequency (or log-frequency). NMFD extends NMF by introducing a convolution in the time direction, and looks for a decomposition of V as

$$V \approx \Lambda = \sum_{\tau} W^{\tau} \overrightarrow{H}^{\tau} \tag{1}$$

where $\rightarrow \tau$ denotes the right shift operator which moves each element in a matrix τ columns to the right, and the superscript τ in W^{τ} is an index. NMFD thus enables the representation of time structure in the extracted patterns. NMF2D generalizes this approach to the frequency direction through a 2-D convolution. By using a log-frequency spectrogram, a pitch change corresponds to a shift on the frequency axis. Assuming that the spectral patterns to be modeled are roughly pitch-invariant, NMF2D can thus account for both time and frequency structures. Concretely, the NMF2D model is

$$V \approx \Lambda = \sum_{\tau} \sum_{\phi} \overset{\downarrow \phi}{W^{\tau}} \overrightarrow{H}^{\tau}{}^{\phi} \tag{2}$$

where $\downarrow \phi$ denotes the down shift operator which moves each element in a matrix ϕ lines down. Up shift and left shift operator are defined in the same way. Applying NMFD or NMF2D to audio signals implies making a sparseness assumption on the signal, as the additivity of magnitudes in the spectral domain is only true if the underlying components of the signal are sparse enough to minimize overlaps.

Lee and Seung [14] introduced efficient algorithms for computing the NMF of a matrix V based on both the least squares error and the $\mathcal{I}$-divergence, which have been extended by Smaragdis for NMFD [13] and Schmidt and Mørup for NMF2D [3, 12]. These algorithms are based on multiplicative updates. If Λ is defined as in (2), we define the objective function as $\mathcal{J}(W, H|V) = ||V - \Lambda||_F^2$ for the least squares error, where $|| \cdot ||_F$ denotes the Frobenius norm (sum of the square of all the elements), or $\mathcal{J}(W, H|V) = \sum_{i,j} V_{i,j} \log\left(\frac{V_{i,j}}{\Lambda_{i,j}}\right) - (V_{i,j} - \Lambda_{i,j})$ for the $\mathcal{I}$-divergence. For the least squares error, the updates can be written as

$$W^{\tau} \leftarrow W^{\tau} \bullet \frac{\sum_{\phi} \overset{\uparrow \phi}{V} \overrightarrow{H}^{\tau}{}^{\phi}}{\sum_{\phi} \overset{\uparrow \phi}{\Lambda} \overrightarrow{H}^{\tau}{}^{\phi}}, \; H^{\phi} \leftarrow H^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow \phi}{W^{\tau}}{}^{T} \overleftarrow{V}^{\tau}}{\sum_{\tau} \overset{\downarrow \phi}{W^{\tau}}{}^{T} \overleftarrow{\Lambda}^{\tau}} \tag{3}$$

while for the $\mathcal{I}$-divergence they become

$$W^{\tau} \leftarrow W^{\tau} \bullet \frac{\sum_{\phi} \overset{\uparrow \phi}{\left(\frac{V}{\Lambda}\right)} \overrightarrow{H}^{\tau}{}^{\phi}}{\sum_{\phi} 1 \cdot \overrightarrow{H}^{\tau}{}^{\phi}}, \; H^{\phi} \leftarrow H^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow \phi}{W^{\tau}}{}^{T} \overleftarrow{\left(\frac{V}{\Lambda}\right)}^{\tau}}{\sum_{\tau} \overset{\downarrow \phi}{W^{\tau}}{}^{T} \cdot 1}. \tag{4}$$

2.3. NMF2D on incomplete spectrograms

We consider the wavelet magnitude spectrogram of an acoustical scene represented as a non-negative matrix $V_{i,j}$, defined on a domain of definition $D = [\![1, M]\!] \times [\![1, N]\!]$ (corresponding for example to the time-frequency region $\{x, t \in \mathbb{R} \mid \Omega_0 \leqslant x \leqslant \Omega_1, T_0 \leqslant t \leqslant T_0 + T\}$). We assume in general that the spectro-temporal patterns to be modeled are roughly pitch-invariant, and that the signals are sparse enough such that the additivity assumption on the magnitude spectrograms holds.

We assume that regions of the magnitude spectrogram are degraded or missing and are interested in performing simultaneously an analysis of this acoustical scene with the NMF2D algorithm despite the presence of gaps, and a reconstruction of the missing parts.

If the data matrix V is incomplete, i.e., if the values $V_{i,j}$ are missing or considered not reliable for some indices $(i, j) \in J \subset D$ and only observed on $I = D \setminus J$, NMF2D cannot be used "as is" because the missing data, if for example assumed to be 0, would create a bias in the estimation of W and H due to the convolutive nature of the model. We thus need to develop a framework to make the algorithm usable in spite of the presence of unobserved data.

2.4. Auxiliary function method

We develop here a framework which can be used in general for adapting algorithms to missing-data problems, and is not restricted to NMF.

Suppose one wants to fit a parametric distribution to an observed contour which is incomplete, in the sense that its values are only known on a subset $I \subset D \subseteq \mathbb{R}^n$, where D is the domain of definition of the problem of interest. Suppose also that if the data were complete, the fitting could be performed (Gaussian distribution fitting, NMF, etc.). Then we show that using an iterative procedure based on the auxiliary function method, the fitting to the incomplete data can also be performed.

Let f be the observed contour, and $g(\cdot; \mathbf{\Theta})$ a model parameterized by $\mathbf{\Theta}$ such that the fitting of this model to an observed contour defined on the whole domain D can be performed.

We consider a distortion function $d : \mathcal{S} \times \mathcal{S} \rightarrow [0, +\infty)$ where $\mathcal{S} \subseteq \mathbb{R}^n$, such that $d(x, y) \geq 0, \forall x, y \in \mathcal{S}$ and equality holds if and only if $x = y$. As this function d is not required to respect the triangle inequality, it is not necessarily a distance, *sensu stricto*. For such a distortion function, we can introduce a measure of the distance between the observed data and the model by integrating d between f and $g(\cdot; \mathbf{\Theta})$ on the subset I:

$$\mathcal{L}(\mathbf{\Theta}) = \int_I d(f(x), g(x; \mathbf{\Theta}))dx. \tag{5}$$

In this kind of situation, it is often preferable, instead of defining an "incomplete model" whose estimation would be cumbersome, to try to fall back on a complete data estimation problem. This is what we do here by introducing an auxiliary function. For any function h taking values in $\mathcal{S}$ and defined on $D \setminus I$, let us define

$$\mathcal{L}^+(\mathbf{\Theta}, h) = \mathcal{L}(\mathbf{\Theta}) + \int_{D \setminus I} d(h(x), g(x; \mathbf{\Theta}))dx. \tag{6}$$

As the second term on the right-hand side is itself derived from the distortion measure, it is non-negative, and thus

$$\mathcal{L}(\mathbf{\Theta}) \leq \mathcal{L}^+(\mathbf{\Theta}, h), \; \forall h. \tag{7}$$

Moreover, there is equality in the inequality for $h = g(\cdot; \mathbf{\Theta})$.

The minimization procedure can now be described as follows. After initializing $\mathbf{\Theta}$ for example by performing the distribution fitting on the observed data completed by 0 on $D \setminus I$, one then iteratively performs the following updates:

Step 1 Estimate h such that $\mathcal{L}(\mathbf{\Theta}) = \mathcal{L}^+(\mathbf{\Theta}, h)$:

$$\hat{h} = g(\cdot; \mathbf{\Theta}). \tag{8}$$

Step 2 Update $\mathbf{\Theta}$ with $\hat{h}$ fixed:

$$\hat{\mathbf{\Theta}} = \operatorname{argmin} \mathcal{L}^+(\mathbf{\Theta}, \hat{h}). \tag{9}$$

Step 2 is simply the fitting of the model $g(\cdot; \mathbf{\Theta})$ on the complete data formed by f on I and $\hat{h}$ on $D \setminus I$. The optimization process is illustrated in Fig. 1.

2.5. Missing-data NMF2D with auxiliary function

Applying the method introduced above to NMF2D leads to the following algorithm, which can be used to analyze incomplete spectrograms, with both objective functions:

Step 1 $V_{i,j}^{(p+1)} = \begin{cases} V_{i,j} & \text{if } (i,j) \in I \\ \Lambda_{i,j}^{(p)} & \text{if } (i,j) \notin I \end{cases}$

Step 2 Update W and H through (3) or (4)

2.6. Sparseness as a key to global structure extraction

A sparseness term can be added to the NMF2D objective function, in the form of the L^1 norm of the matrix H, leading to the so-called Sparse NMF2D (SNMF2D) [12]. As pointed out by Mørup and Schmidt, there is an intrinsic ambiguity in the decomposition (2). The structure of a factor in H can to some extent be put into the signature of the same factor in W and vice versa. Imposing sparseness on H forces the structure into W and thus alleviates this ambiguity. In the case of spectrograms with gaps, this is even more critical, and sparseness becomes compulsory. Indeed, without a sparseness term, assuming that the spectral envelopes were time and pitch invariant (which is only approximately true), a perfect reconstruction of the spectrogram with gaps could be obtained with a single frame representing the spectral envelope template in W and the power envelope in the time direction (again, gaps included) in H. The role of sparseness is thus to ensure that global structures are extracted and used throughout the spectrogram, and it will be the key that will enable us to fill in the gaps in the spectrogram.

3. Probabilistic interpretation for Bregman divergences

3.1. Relation between Bregman divergence-based optimization and Maximum Likelihood estimation

Due to length considerations, we shall only give a rough overview of the concepts of exponential families and Bregman divergences, and refer to [15] for more details and rigorous derivations.

A Bregman divergence is a particular case of distortion measure defined as

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y); x - y \rangle,$$

where $\phi : \mathcal{S} \to \mathbb{R}$ is a strictly convex and differentiable function on an open convex set $\mathcal{S} \subseteq \mathbb{R}^d$, $\nabla\phi(y)$ is the gradient vector of ϕ evaluated at y and $\langle \cdot; \cdot \rangle$ the inner product. Bregman

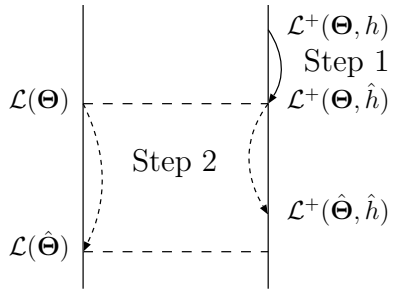


Figure 1: Optimization through the iterative procedure. During the step 1, the auxiliary parameter h is updated to $\hat{h}$ so that $\mathcal{L}(\mathbf{\Theta}) = \mathcal{L}^+(\mathbf{\Theta}, \hat{h})$. Then, during the step 2, $\mathcal{L}^+(\mathbf{\Theta}, \hat{h})$ is optimized w.r.t. $\mathbf{\Theta}$, ensuring that $\mathcal{L}(\hat{\mathbf{\Theta}}) \leq \mathcal{L}^+(\hat{\mathbf{\Theta}}, \hat{h}) < \mathcal{L}^+(\mathbf{\Theta}, \hat{h}) = \mathcal{L}(\mathbf{\Theta})$. The minimization of $\mathcal{L}(\mathbf{\Theta})$ can thus be performed through the minimization of the auxiliary function $\mathcal{L}^+(\mathbf{\Theta}, h)$ alternately w.r.t. h and $\mathbf{\Theta}$.

divergences include a large number of useful loss functions such as squared loss, KL-divergence, logistic loss, Mahalanobis distance, Itakura-Saito distance, and the $\mathcal{I}$-divergence. They are non-negative, and equal to zero if and only if $x = y$.

On the other hand, exponential families form a group of probability distributions which comprise many common families of probability distributions such as the normal, gamma, Dirichlet, binomial and Poisson distributions, among others. They can be represented in their so-called *mean-value parameterization*

$$P_{\phi,\mu}(X) = e^{\phi(\mu) + \langle \nabla\phi(\mu); X - \mu \rangle} r(X), \tag{10}$$

where r is a non-negative function. The mean of $P_{\phi,\mu}$ is μ, called the expectation parameter of the exponential family, itself denoted by $\mathcal{F}_\phi$.

Banerjee et al. [15] show that a correspondence exists between a wide subclass of Bregman divergences, including the loss functions mentioned above, and certain exponential families. More precisely, they show that, under certain conditions, there exist functions b_ϕ such that the relation

$$P_{\phi,\mu}(x) = e^{-d_\phi(x,\mu)} b_\phi(x) \tag{11}$$

holds. The point here is that although the set of all the instances for which this relation holds can be shown [15] to be included into the domain of definition $\mathrm{dom}(\phi)$ of ϕ, it may in some cases be strictly included. This is what happens for example for the $\mathcal{I}$-divergence (with $\phi(\mu) = \mu \log \mu - \mu$) for which $\mathrm{dom}(\phi) = \mathbb{R}^+$ (extending the definition of ϕ for $\mu = 0$). The corresponding exponential family is the Poisson family, for which the set of instances is only $\mathbb{N}$.

The relation (11) builds a bridge between optimization based on Bregman divergences and Maximum Likelihood (ML) estimation with exponential families. Trying to fit a model $g(\cdot; \mathbf{\Theta})$, defined on a domain D with parameter $\mathbf{\Theta}$, to an observed distribution f with a measure of distance between the two based on a Bregman divergence d_ϕ amounts to looking for $\mathbf{\Theta}$ minimizing $\int_D d_\phi(f(x), g(x; \mathbf{\Theta}))dx$. But according to (11), this is equivalent (up to some precautions regarding the domain of definition evoked above) to maximizing w.r.t. $\mathbf{\Theta}$ the log-likelihood $\int_D \log P_{\phi,g(x;\mathbf{\Theta})}(f(x))dx$ where the observed data points $f(x)$ at point x are assumed to have been independently generated from $P_{\phi,g(x;\mathbf{\Theta})}$.

3.2. Relation to the EM algorithm

We consider the framework of Section 2.4, with as distortion function a Bregman divergence d_ϕ associated to an exponential family $\mathcal{F}_\phi$. In the following, we will denote by $\nu_{x,\phi,\mathbf{\Theta}}(z)$ the density of a probability distribution from $\mathcal{F}_\phi$ with expectation parameter $g(x;\mathbf{\Theta})$. Following (11), we can write

$$\nu_{x,\phi,\mathbf{\Theta}}(z) = e^{-d_\phi(z,g(x;\mathbf{\Theta}))} b_\phi(z). \tag{12}$$

As in the ML counterpart to Bregman divergence based optimization the data are assumed to have been generated independently from probability distributions of $\mathcal{F}_\phi$, we notice that observed and unobserved data are independent conditionally to $\mathbf{\Theta}$. We can now derive the Q-function of the EM algorithm:

$$\begin{aligned}
Q(\mathbf{\Theta},\bar{\mathbf{\Theta}}) &= \mathbb{E}(\log P(h|\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} + \mathbb{E}(\log P(f|\mathbf{\Theta}))_{P(h|f,\bar{\mathbf{\Theta}})} \\
&= \int_{\mathbb{R}^n\setminus I} \int \nu_{x,\phi,\bar{\mathbf{\Theta}}}(z) \log \nu_{x,\phi,\mathbf{\Theta}}(z) dz\, dx \\
&\quad + \left(\int P(h|f,\bar{\mathbf{\Theta}})\right) \int_I \log P(f(x)|\mathbf{\Theta}) dx \\
&= \int_{\mathbb{R}^n\setminus I} \int \nu_{x,\phi,\bar{\mathbf{\Theta}}}(z) \Big(\log b_\phi(z) - d_\phi(z, g(x;\mathbf{\Theta}))\Big) dz\, dx \\
&\quad + \int_I \Big(\log b_\phi(f(x)) - d_\phi(f(x), g(x;\mathbf{\Theta}))\Big) dx \\
&= -\int_{\mathbb{R}^n\setminus I} \int \nu_{x,\phi,\bar{\mathbf{\Theta}}}(z) d_\phi(z, g(x;\mathbf{\Theta})) dz\, dx \\
&\quad - \int_I d_\phi(f(x), g(x;\mathbf{\Theta})) dx + C_1(f,\bar{\mathbf{\Theta}}),
\end{aligned} \tag{13}$$

where $C_1(f,\bar{\mathbf{\Theta}})$ does not depend on $\mathbf{\Theta}$. If we now rewrite $d_\phi(z, g(x;\mathbf{\Theta}))$ as

$$\begin{aligned}
d_\phi(z, g(x;\mathbf{\Theta})) &= d_\phi(g(x;\bar{\mathbf{\Theta}}), g(x;\mathbf{\Theta})) + \phi(z) - \phi(g(x;\bar{\mathbf{\Theta}})) \\
&\quad - \langle z - g(x;\bar{\mathbf{\Theta}}); \nabla\phi(g(x;\mathbf{\Theta}))\rangle,
\end{aligned} \tag{14}$$

we can simplify the first term in Eq. 13:

$$\begin{aligned}
&\int \nu_{x,\phi,\bar{\mathbf{\Theta}}}(z) d_\phi(z, g(x;\mathbf{\Theta})) dz \\
&\qquad = d_\phi(g(x;\bar{\mathbf{\Theta}}), g(x;\mathbf{\Theta})) + C_2(f,\bar{\mathbf{\Theta}}),
\end{aligned}$$

where $C_2(f,\bar{\mathbf{\Theta}})$ does not depend on $\mathbf{\Theta}$. To lead the calculation above, we used the fact that the mass of $\nu_{x,\phi,\bar{\mathbf{\Theta}}}$ is 1 and its mean is $g(x;\bar{\mathbf{\Theta}})$. We then obtain for the Q-function

$$\begin{aligned}
Q(\mathbf{\Theta},\bar{\mathbf{\Theta}}) &= -\int_{\mathbb{R}^n\setminus I} d_\phi(g(x;\bar{\mathbf{\Theta}}), g(x;\mathbf{\Theta})) dx \\
&\quad - \int_I d_\phi(f(x), g(x;\mathbf{\Theta})) dx + C(f,\bar{\mathbf{\Theta}}) \\
&= -\mathcal{L}^+(\mathbf{\Theta}, g(x;\bar{\mathbf{\Theta}})) + C(f,\bar{\mathbf{\Theta}}),
\end{aligned} \tag{15}$$

where $C(f,\bar{\mathbf{\Theta}})$ again does not depend on $\mathbf{\Theta}$.

Altogether, we find that there is a correspondence between the Q-function and the auxiliary function $\mathcal{L}^+$ that we introduced in 2.4. Computing the Q-function, i.e., the E-step of the EM algorithm, corresponds to computing the auxiliary function, which is done by replacing the unknown data by the model at the current step. Maximizing the Q-function w.r.t. $\mathbf{\Theta}$, i.e., the M-step of the EM algorithm, corresponds to minimizing the auxiliary function w.r.t. $\mathbf{\Theta}$. This shows how to derive the auxiliary function in an EM point of view, and enables us for example to consider prior distributions on the parameters and perform a MAP estimation.

We shall note however that one has to pay attention to the support of the probability distributions of the exponential family involved. Indeed, as noted earlier, it may happen that these distributions have a smaller support than the original set on which the Bregman divergence is defined. The formulation presented in Section 2.4 is thus more general than its EM counterpart, although it does not justify the use of penalty functions as prior distributions on the parameters. This is what happens with the $\mathcal{I}$-divergence, as noted above, although it is still possible to justify the use of the ML interpretation with continuous data in this particular case [16].

3.3. Use of prior distributions with SNMF2D

The NMF framework can be considered in a Bayesian way based on the correspondence between Bregman divergence based optimization and ML estimation either for the least squares error or the $\mathcal{I}$-divergence. Indeed, the NMF objective function can be converted into a log-likelihood [14, 17], to which prior constraints on the parameters can further be added.

Sparseness terms involving L^p norms of H can be considered as such, the L^1 norm sparseness term used here corresponding for example to a Laplace distribution.

But one can also introduce Markovian constraints on the parameters to ensure smooth solutions. Using Gamma chains on the coefficients of W and H in the time direction, one can show that analytical update equations can still be obtained and the objective function can be optimized based on the Expectation-Constrained Maximization (ECM) algorithm [18].

4. Experimental evaluation

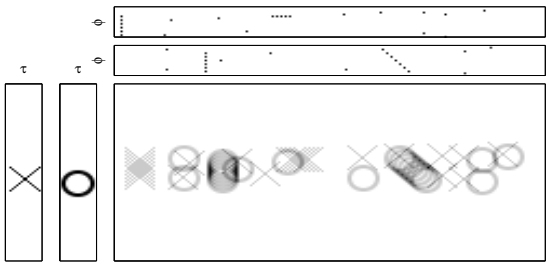
4.1. Toy example: reconstructing a 2D image

We first tested our algorithm on simulated data used by Mørup and Schmidt in [12]. The data, shown in Fig. 2 (a), were created with W consisting of one cross in the first factor and one circle in the second, convolved with H given in the top of the figure to yield the full data matrix V. The SNMF2D algorithm was used in the same conditions as in [12], with $\tau = \{0,\dots,16\}$ and $\phi = \{0,\dots,16\}$. The circle and cross templates span roughly 15 frames in both horizontal and vertical directions, while the whole data is 200 frames wide. To construct the incomplete data, we erased 3 frames horizontally and 2 frames every 10 frames vertically, as shown in Fig. 2 (b). Note that none of the occurrences of the structures (circle and cross) is fully available. However, in this ideal case where the original data is a strict convolution of the patterns, the proposed algorithm is able to extract the original patterns and their occurrences and to reconstruct the original data, as can be seen in Fig. 2 (c) for the $\mathcal{I}$-divergence update equations (similar results were obtained with the least squares updates). This shows that the reconstruction is based on global features of the data.

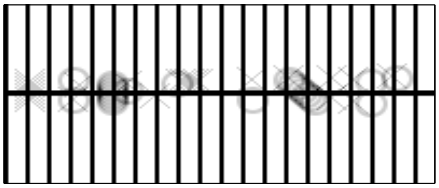
4.2. Audio example: reconstructing gaps in a sound

4.2.1. Experimental setting

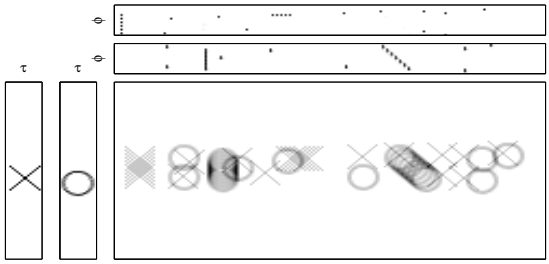
For auditory restoration experiments, contrary to what is done in [3], we did not use the short time Fourier transform afterwards converted into a log-frequency magnitude spectrogram, but a wavelet transform, which directly gives a log-frequency spectrogram. More precisely, the magnitude spectrogram was calculated from the input signals digitized at a 16kHz sampling



(a) Original simulated data.



(b) Incomplete data with erased regions in black.



(c) Reconstruction using the $\mathcal{I}$-divergence.

Figure 2: SNMF2D with missing-data on a toy example.

rate using a Gabor wavelet transform with a time resolution of 16ms for the lowest frequency subband. Higher subbands were downsampled to match the lowest subband resolution. The frequency range extended from 50Hz to 8kHz and was covered by 200 channels, for a frequency resolution of 44 cent.

We used a 4.8s piece of computer generated polyphonic music containing a trumpet and a piano, already used by Schmidt and Mørup in [3]. Its spectrogram can be seen in Fig. 3 (a). The incomplete waveform was built by erasing 80ms of signal every 416ms, leading to a signal with about 20% of data missing. Its spectrogram is shown in Fig. 3 (b).

The mask indicating the region I to inpaint was built according to the erased portions of the waveform. With a Gabor wavelet transform, the influence of a local modification of the signal theoretically spans the whole interval. However, as the windows are Gaussian, one can consider that the influence becomes almost null further than about three times the standard deviation. This standard deviation is inversely proportional with the frequency, and the influence should thus be considered to span a longer interval for lower frequencies. Although it leaves some unreliable portions of the spectrogram out of the mask in the lower frequencies, for simplicity, we did not consider here this dependence on frequency, and simply considered unreliable, for each 80ms portion of waveform erased, 6 whole spectrogram frames (corresponding to about 96ms of signal in the highest frequencies). The incomplete spectrogram is shown in Fig. 3 (c), with areas to inpaint in black.

The SNMF2D parameters were as follows. As in [3], we used two factors, $d = 2$, since we are analyzing a scene with two instruments, and the number of convolutive components in pitch was set to $\phi = \{0, \ldots, 11\}$, as the pitch of the notes in the data spans three whole notes. For the convolutive components in time, we used empirically $\tau = \{0, \ldots, 31\}$, for a time range of about 500ms, thus roughly spanning the length of the eighth notes in the music sample. The $\mathcal{I}$-divergence was used as the distortion measure, and the sparseness term coefficient set to 0.001. The algorithm was ran for 100 iterations.

Table 1: *Results of the reconstruction experiment*

	SNR		SSNR	
	in	out	in	out
MI / C	10.7	12.9	10.5	11.7
MI / I	-3.7	13.1	3.4	12.1
SC / C	13.1	13.0	12.3	11.9
SI / I	6.2	10.5	7.3	9.9
I / C	2.2	15.7	2.4	16.9

4.2.2. Results

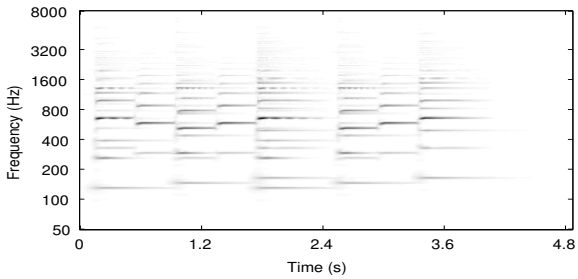
To evaluate the reconstruction accuracy of the spectrogram, we use two measures: Signal to Noise Ratio (SNR) and Segmental SNR (SNR) computed as the median of the individual SNRs of all the frames. We note that computing the SNR directly on the magnitude spectrogram amounts to assuming that the phase is perfectly reconstructed. The results are summarized in Table 1, where "in" refers to the measure computed inside the gaps (the inpainted part), "out" to the measure computed outside the gaps (the part more classically reconstructed based on observed data), "M" refers to the proposed Missing-data SNMF2D, "S" to SNMF2D, "C" to the magnitude spectrogram of the complete waveform, and "I" to the one of the incomplete waveform. Finally, "WX" refers to the spectrogram reconstructed by applying algorithm W on spectrogram X, and "Y/Z" to the comparison of spectrogram Y with spectrogram Z as a reference. For example, the SNR of "MI/C" is the SNR of the spectrogram reconstructed using our missing-data approach on the spectrogram of the incomplete data w.r.t. the spectrogram of the full waveform.

One can see through MI/I that the proposed algorithm correctly performs its task of reconstructing the observed data ("out"), which is not the case for SI/I, showing that the gaps hinder SNMF2D from performing well. The MI/C results show that the formerly erased regions ("in") are correctly inpainted, with a great improvement over the incomplete spectrogram, as seen in I/C, and that our method performs closely to SNMF2D applied on the complete spectrogram, as seen in SC/C.

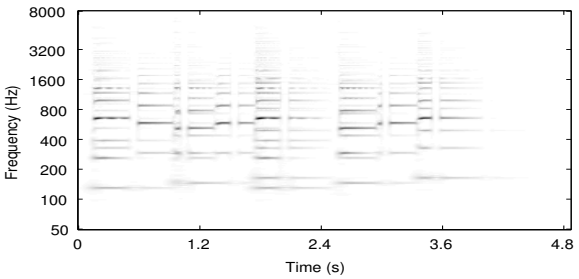
Graphical results are shown in Fig. 3 (d), (e), (f), where one can see in particular that the acoustic scene analysis is performed correctly, and that blind source separation is also performed in spite of the presence of gaps.

5. Conclusion

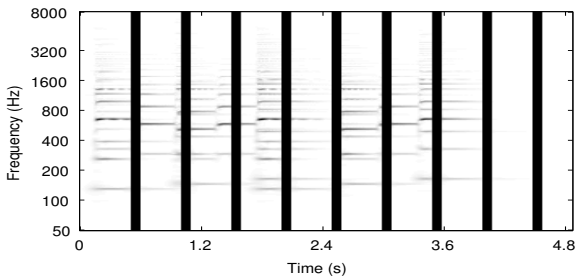
We presented a computational framework to model auditory induction, i.e. the human auditory system's ability to estimate the missing parts of a continuous auditory stream briefly covered by noise, by extending the SNMF2D algorithm to handle unobserved data. We related the method to the EM algorithm, enabling the use of priors on the parameters. We finally illustrated on a simple example how the proposed framework was able to simultaneously perform acoustic scene analysis and gap interpolation. Future works include a more thorough experimental evaluation of the method and the use of smoothing prior distributions.



(a) Spectrogram of the original waveform.

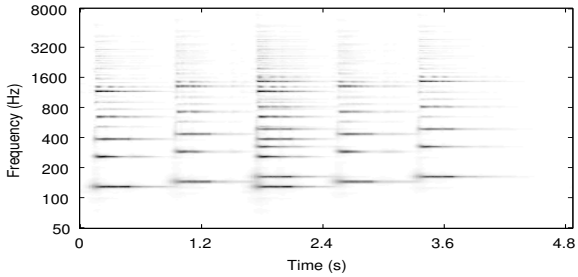


(b) Spectrogram of the incomplete waveform.

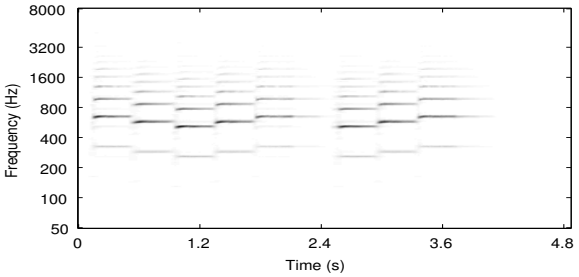


(c) Incomplete spectrogram with areas to inpaint in black.

(d) Spectrogram reconstructed using the $\mathcal{I}$-divergence.



(e) Reconstructed and separated spectrogram of the piano part.



(f) Reconstructed and separated spectrogram of the trumpet part.

Figure 3: SNMF2D with missing-data on the spectrogram of a waveform with missing samples.

6. References

[1] D.-L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley, 2006.

[2] M. Kashino, "Phonemic restoration: The brain creates missing speech sounds," *Acoust. Sci. & Tech.*, vol. 27, no. 6, pp. 318–321, 2006.

[3] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc. ICA*, Apr. 2006, pp. 700–707.

[4] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[5] P. A. A. Esquef and L. W. P. Biscainho, "An efficient model-based multirate method for reconstruction of audio signals across long gaps," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1391–1400, Jul. 2006.

[6] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration: A Statistical Model Based Approach*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1998.

[7] M. Reyes-Gomez, N. Jojic, and D. P. W. Ellis, "Towards single-channel unsupervised source separation of speech mixtures: the layered harmonics/formants separation-tracking model," in *Proc. SAPA*, Oct. 2004, pp. 25–30.

[8] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[9] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. SAPA*, Sep. 2008.

[10] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Siggraph 2000, Computer Graphics Proc.*, 2000, pp. 417–424.

[11] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[12] M. Mørup and M. N. Schmidt, "Sparse non-negative matrix factor 2-D deconvolution," Tech. Univ. of Denmark, Tech. Rep., 2006.

[13] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. ICA*, 2004, pp. 494–499.

[14] D. D. Lee and H. S. Seung, "Learning of the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[15] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[16] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "On the interpretation of I-divergence-based distribution-fitting as a maximum-likelihood estimation problem," The University of Tokyo, Tech. Rep. METR 2008-11, Mar. 2008.

[17] P. Sajda, S. Du, and L. C. Parra, "Recovery of constituent spectra using non-negative matrix factorization," in *Proc. SPIE Wavelets X*, Aug. 2003, pp. 321–331.

[18] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.