The editorial staff at IGI introduced important errors in the published version of this chapter; in particular one was figure has been split in two parts, another figure is missing and references to figures and tables have been messed. This document is produced from the original LaTex source and should be close to error-free.

Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition.

Cédric Févotte
CNRS LTCI; Télécom ParisTech
Email: fevotte@telecom-paristech.fr

Abstract

Nonnegative matrix factorization (NMF) is a popular linear regression technique in the fields of machine learning and signal/image processing. Much research about this topic has been driven by applications in audio. NMF has been for example applied with success to automatic music transcription and audio source separation, where the data is usually taken as the *magnitude* spectrogram of the sound signal, and the Euclidean distance or Kullback-Leibler divergence are used as measures of fit between the original spectrogram and its approximate factorization. In this chapter we give evidence of the relevance of considering factorization of the *power* spectrogram, with the Itakura-Saito (IS) divergence. Indeed, IS-NMF is shown to be connected to maximum likelihood inference of variance parameters in a well-defined statistical model of superimposed Gaussian components and this model is in turn shown to be well suited to audio. Furthermore, the statistical setting opens doors to Bayesian approaches and to a variety of computational inference techniques. We discuss in particular model order selection strategies and Markov regularization of the activation matrix, to account for time-persistence in audio. This chapter also discusses extensions of NMF to the multichannel case, in both instantaneous or convolutive recordings, possibly underdetermined. We present in particular audio source separation results of a real stereo musical excerpt.

1 Introduction

Nonnegative matrix factorization (NMF) is a is a linear regression technique, employed for non-subtractive, part-based representation of nonnegative data. Given a data matrix $\mathbf{V}$ of dimensions $F \times N$ with nonnegative entries, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

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 $F\,K + K\,N << F\,N$, hence reducing the data dimension. Early works about NMF include (Paatero, 1997) and (Lee and Seung, 1999), the latter in particular prove very influential. NMF has been applied to diverse problems (such as pattern recognition, clustering, data mining, source separation, collaborative filtering) in many areas (such as text processing, bioinformatics, signal/image processing, finance). Much research about NMF has been driven by applications in audio, namely automatic music transcription (Smaragdis and Brown, 2003; Abdallah and Plumbley, 2004) and source separation (Virtanen, 2007; Smaragdis, 2007), where the data $\mathbf{V}$ is usually taken as the magnitude spectrogram of the audio signal.

Along Vector Quantization (VQ), Principal Component Analysis (PCA) or Independent Component Analysis (ICA), NMF provides an unsupervised linear representation of data, in the sense that a data point $\mathbf{v}_n$ (n^{th} column of $\mathbf{V}$) is approximated as a linear combination of salient features :

$$\begin{array}{ccccc} \mathbf{v}_n & \approx & \mathbf{W} & & \mathbf{h}_n \\ & & \text{“explanatory variables”} & & \text{“regressors”} \\ & & \text{“basis”, “dictionary”} & & \text{“expansion coefficients”} \\ & & \text{“patterns”} & & \text{“activation coefficients”} \end{array}$$

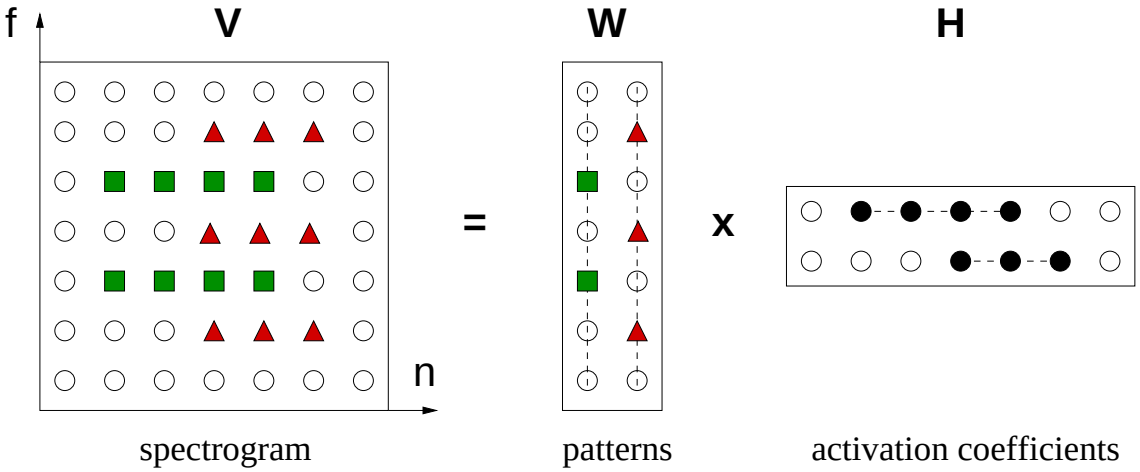

Figure 1: Illustration of the effect of NMF on an audio spectrogram.

A distinctive feature of NMF with respect to VQ, PCA or ICA is that it keeps $\mathbf{W}$ and $\mathbf{h}_n$ nonnegative, hence improving the interpretability of the learnt dictionary and of the activation coefficients when the data is nonnegative. More precisely, the nonnegativity restriction on $\mathbf{W}$ allows the learnt features to belong to the same space than data, while the nonnegative restriction on $\mathbf{H}$ favors part-based decomposition as subtractive combination are forbidden, i.e, the data has to be "assembled" from the elementary building blocks in $\mathbf{W}$. As such, in their seminal paper Lee and Seung (1999) show how parts of faces (noise, eyes, cheeks, etc.) can be learnt from a training set composed of faces, when the bases returned by PCA or VQ are more "holistic" in the sense that each feature attempts to generalize as much as possible the entire dataset. The effect of NMF onto an audio spectrogram is illustrated on figure 1. It can be seen that the NMF model is well suited to the composite structure of music in the sense that the factorization can be expected to separate mingled patterns in the spectrogram; the patterns may correspond to spectra of elementary musical objects such as notes or percussions or, as we shall see later, higher level structures.

In the literature, the factorization (1) is usually achieved through minimization of a measure of fit defined by

$$D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}) \tag{2}$$

where $d(x|y)$ is a scalar cost function, typically a positive function with a single minimum 0 for $x = y$. The minimization, with respect to $\mathbf{W}$ and $\mathbf{H}$, is subject to nonnegativity constraints on the coefficients of both factors. Popular cost functions are the Euclidean distance, here defined as

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2 \tag{3}$$

and the (generalized) Kullback-Leibler (KL) divergence, also referred to as I-divergence, defined by

$$d_{KL}(x|y) = x\,\log\frac{x}{y} - x + y\,. \tag{4}$$

More general families of cost functions have also been considered for NMF, such as Csiszár and Bregman divergences (Cichocki et al., 2006; Dhillon and Sra, 2005). The choice of the NMF cost function should be driven by the problem setup and type of data, and if a good deal of literature is devoted to improving performance of algorithms given a cost function, little literature has been devoted to how to choose a cost function with respect to a particular type of data and application.

In this chapter we are specifically interested in NMF with the Itakura-Saito (IS) divergence, for its relevance to the decomposition of audio power spectrograms, as we intend to show. The expression of the IS divergence is given by

$$d_{IS}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1 \ . \tag{5}$$

This divergence was obtained by Itakura and Saito (1968) from the maximum likelihood (ML) estimation of short-time speech spectra under autoregressive modeling. It was presented as "a measure of the goodness of fit between two spectra" and became popular in the speech community during the seventies. It was in particular praised for the good perceptual properties of the reconstructed signals it led to (Gray et al., 1980).

As we shall see, this divergence has interesting properties. It is in particular scale-invariant, meaning that low energy components of $\mathbf{V}$ bear the same relative importance as high energy ones. This is relevant to situations in which the coefficients of $\mathbf{V}$ have a large dynamic range, such as in audio short-term spectra. The IS divergence also leads to desirable statistical interpretations of the NMF problem. Indeed, IS-NMF can be recast as ML estimation of $\mathbf{W}$ and $\mathbf{H}$ in a model of superimposed Gaussian components. Equivalently, IS-NMF can be interpreted as ML of $\mathbf{W}$ and $\mathbf{H}$ in multiplicative Gamma noise. As will shall see, this statistical framework open doors to Bayesian approaches to NMF, that can account for regularization constraints on either $\mathbf{W}$ or $\mathbf{H}$, model order selection as well as a variety of computational inference techniques.

This chapter is organized as follows. Section 2 is devoted to the general presentation of NMF with the IS divergence. We discuss its scale invariance, its nonconvexity, the latent statistical model and describe algorithms for achieving IS-NMF. Decomposition results of a simple case-study piano sequence are presented; they illustrate the relevance of IS-NMF of the power spectrogram as compared to the more standard approach consisting of KL-NMF of the magnitude spectrogram. Section 3 describe Bayesian extensions of IS-NMF. We present in particular a model order selection strategy based on *automatic relevance determination* that allows to determine an "efficient" number of components K. We also describe an IS-NMF algorithm with a Markov model for $\mathbf{H}$ that promotes smoothness of the activation coefficients. We also mention advanced computational techniques for NMF, based on Markov chain Monte Carlo (MCMC). While Sections 2 and 3 inherently assume single-channel data, Section 4 describes extensions to multichannel data : we present nonnegative *tensor* factorization techniques for either instantaneously or convolutively mixed data, that allow joint processing of the channel spectrograms. Finally, Section 5 draws conclusions and perspectives of this work. This chapter intends to describe and discuss in a unified framework parts of recent contributions (Févotte et al., 2009; Ozerov and Févotte, 2010; Tan and Févotte, 2009; Ozerov et al., 2009; Févotte and Cemgil, 2009; Durrieu et al., 2009, 2010).

2 NMF with the Itakura-Saito divergence

This section is devoted to a general presentation of NMF with the IS divergence. In the following, the entries of matrices $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$ are denoted v_{fn}, w_{fk} and h_{kn} respectively. Lower case bold letters will in general denote columns, such that $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, while lower case plain letters with a single index denote rows, such that $\mathbf{H} = [h_1^T, \ldots, h_K^T]^T$. We also define the matrix $\hat{\mathbf{V}} = \mathbf{WH}$, whose entries are denoted $\hat{v}_{fn}$. Where these conventions clash, the intended meaning should be clear from the context.

2.1 Properties

2.1.1 Scale invariance and nonconvexity

An interesting property of the IS divergence is scale-invariance, i.e,

$$d_{IS}(\lambda x|\lambda y) = d_{IS}(x|y). \tag{6}$$

This property is not shared by the popular Euclidean and KL cost functions. As such we have $d_{EUC}(\lambda x|\lambda y) = \lambda^2 \, d_{EUC}(x|y)$ and $d_{KL}(\lambda x|\lambda y) = \lambda \, d_{KL}(x|y)$. The scale invariance means that

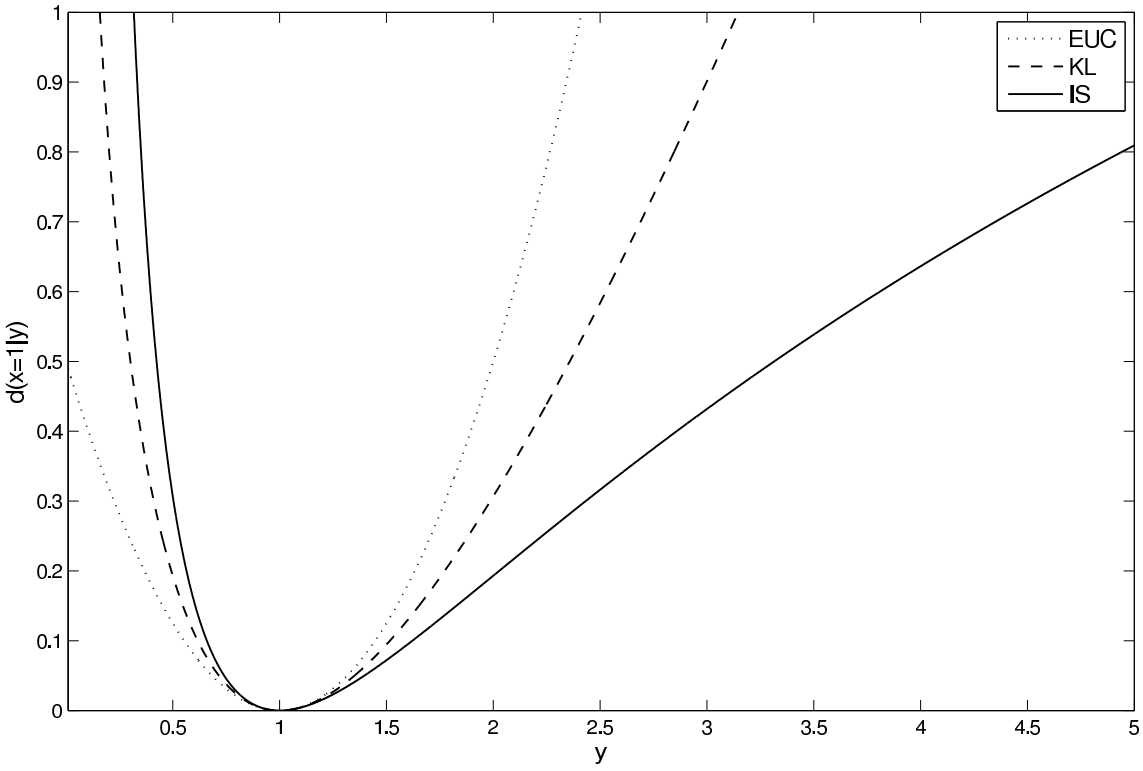

Figure 2: Euclidean, KL and IS costs $d(x|y)$ as a function of y and for $x = 1$. The Euclidean and KL divergences are convex on $(0, \infty)$. The IS divergence is convex on $(0, 2x]$ and concave on $[2x, \infty)$.

same relative weight is given to small and large coefficients of $\mathbf{V}$ in cost function (2), in the sense that a bad fit of the factorization for a low-power coefficient v_{fn} will cost as much as a bad fit for a higher power coefficient $v_{f'n'}$. In contrast, factorizations obtained with the Euclidean distance or the KL divergence will rely more heavily on the largest coefficients and less precision is to be expected in the estimation of the low-power components. The scale invariance of the IS divergence is relevant to the 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.

A property shared by the Euclidean and KL costs $d_{EUC}(x|y)$ and $d_{KL}(x|y)$ is convexity with respect to y. This means that in their cases, the cost function $D(\mathbf{V}|\mathbf{WH})$ is at least convex with respect to either $\mathbf{W}$ or $\mathbf{H}$. In contrast, the IS divergence $d_{IS}(x|y)$ is, as a function of y, convex on $(0, 2x]$ and concave on $[2x, \infty)$, see figure 2, which, we observed in practice, makes it more prone to local minima.

2.1.2 Statistical interpretations

Very interestingly, IS-NMF can be viewed as maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in the statistical composite model described next. Let $\mathbf{x}_n \in \mathbb{C}^{F\times 1}$ be the STFT at frame n of some time-domain signal x. Consider the generative model defined by, $\forall n = 1, \ldots, N$

$$\mathbf{x}_n = \sum_{k=1}^{K} \mathbf{c}_{k,n} \tag{7}$$

where $\mathbf{c}_{k,n}$ belongs to $\mathbb{C}^{F\times 1}$ and

$$\mathbf{c}_{k,n} \sim \mathcal{N}_c(0, h_{kn}\,\mathrm{diag}\,(\mathbf{w}_k)), \tag{8}$$

where $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the proper multivariate complex Gaussian distribution, defined in Appendix A. The variable $\mathbf{c}_{k,n}$, which we refer to as *component* in the following, is basically characterized by a spectral signature $\mathbf{w}_k$, amplitude-modulated in time by the frame-dependent coefficient h_{kn}, which accounts for nonstationarity. Assume each component sequence of frames $\{\mathbf{c}_{k,1}, \ldots, \mathbf{c}_{k,N}\}$ to be independently distributed and the components $\mathbf{c}_{1,n}, \ldots, \mathbf{c}_{K,n}$ to be mutually independent at each frame. Then, the likelihood criterion to optimize writes

$$\begin{aligned} C_{ML}(\mathbf{W}, \mathbf{H}) &\stackrel{\text{def}}{=} -\log p(\mathbf{X}|\mathbf{W}, \mathbf{H}) && (9) \\ &= -\sum_{n=1}^{N}\sum_{f=1}^{F} \log \mathcal{N}_c\left(x_{fn}|0, \sum_k w_{fk}\, h_{kn}\right) && (10) \\ &\stackrel{\text{c}}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} d_{IS}\left(|x_{fn}|^2 \mid \sum_k w_{fk}\, h_{kn}\right) && (11) \end{aligned}$$

where $\stackrel{\text{c}}{=}$ denotes equality up to additive constant terms. Taking $\mathbf{V}$ as the matrix with entries $v_{fn} = |x_{fn}|^2$, ML estimation of $\mathbf{W}$ and $\mathbf{H}$ hence amounts to the NMF $\mathbf{V} \approx \mathbf{WH}$, with the IS divergence.

The generative model (7)-(8) was considered by Benaroya et al. (2003, 2006a) for single-channel audio source separation, but there the dictionary $\mathbf{W}$ is trained separately with VQ and the link with IS-NMF is not fully exploited. Another equivalent way to look at the statistical structure is to assume that $\mathbf{V}$ is a noisy observation of $\hat{\mathbf{V}}$, with *multiplicative* independently and identically Gamma distributed noise (with mean 1). This is how Abdallah and Plumbley (2004) derive a "statistically motivated error measure", which coincides with the IS divergence, in the very similar context of nonnegative sparse coding. Parry and Essa (2007) derive a "phase-aware nonnegative spectrogram factorization" method which also amounts to NMF of the power spectogram with the IS divergence, based on a statistical model very similar to (7)-(8). In contrast Euclidean and KL NMF respectively underlay additive Gaussian and Poisson noise, see, e.g., (Févotte and Cemgil, 2009) and below.

2.1.3 Phase and component reconstruction

The IS-NMF composite model (7)-(8) sets no informative assumption about the phase of each component; the proper complex Gaussian assumption amounts to modeling the phases as uniform random variables. Given estimates of $\mathbf{W}$ and $\mathbf{H}$, minimum mean square error (MMSE) estimates can be obtained through Wiener filtering, such that

$$\hat{c}_{k,fn} = \frac{w_{fk}\, h_{kn}}{\sum_{l=1}^{K} w_{fl}\, h_{ln}} x_{fn}. \tag{12}$$

A consequence of Wiener reconstruction is that the phase of all components $\hat{c}_{k,fn}$ is equal to the phase of x_{fn}. Note that with the Wiener gains summing up to 1 for a fixed entry (f, n), the decomposition is conservative, i.e,

$$\mathbf{x}_n = \sum_{k=1}^{K} \hat{\mathbf{c}}_{k,n}. \tag{13}$$

We would like to contrast the IS-NMF model (7) with the more common approach consisting of factorizing the *magnitude* spectrogram with the KL divergence, see e.g, (Smaragdis and Brown, 2003; Smaragdis, 2007; Virtanen, 2007; Cemgil, 2009). In that case, the assumed latent generative model is

$$|\mathbf{x}_n| = \sum_{k=1}^{K} |\mathbf{c}_{k,n}| \tag{14}$$

with

$$|c_{k,fn}| \sim \mathcal{P}(h_{kn} w_{fk}), \tag{15}$$

where $\mathcal{P}(\lambda)$ denotes the Poisson distribution, defined in Appendix A. The Poisson distribution being closed under summation (like the Gaussian distribution), one obtains $|x_{fn}| \sim \mathcal{P}(\sum_{k=1}^{K} w_{fk} h_{kn})$ and it can easily be derived that the likelihood $-\log p(\ |\mathbf{X}|\ |\mathbf{W},\mathbf{H})$ is up to a constant equal to $D_{KL}(\ |\mathbf{X}|\ |\ \mathbf{W}\,\mathbf{H})$. Given estimates of $\mathbf{W}$ and $\mathbf{H}$, MMSE estimates of the magnitude-components are obtained by

$$\widehat{|c_{k,fn}|} = \frac{w_{fk}\, h_{kn}}{\sum_{l=1}^{K} w_{fl}\, h_{ln}} |x_{fn}|. \tag{16}$$

A major difference between the approaches consisting of using KL-NMF on the magnitude spectrogram and IS-NMF on the power spectrogram is that the generative model (14)-(15) implied by the former simply discards the phase of the observation, and of the components. In contrast, the phase is taken into account in the other approach, even though in a noninformative way. When it comes to reconstruct time domain components from equation (16) through inverse-STFT, it is common practice to set the phase of the individual components to the phase of x_{fn}, i.e,

$$\hat{c}_{k,fn} = \widehat{|c_{k,fn}|}\ \arg(x_{fn}), \tag{17}$$

where $\arg(x)$ denotes the phase of complex scalar x. So one can argue that in the end, component estimates have same phase than observation in both cases - but it is important to understand that in the case of IS-NMF of the power spectrogram this is a natural consequence of the modeling, while it is a somehow ad-hoc choice for KL-NMF of the magnitude spectrogram. Another criticism about model (14) & (15) concerns the Poisson modeling of $|c_{k,fn}|$, which is formerly only defined for integers and thus impairs rigorous statistical interpretation of KL-NMF on non-countable data such as audio spectra.

Finally, we want to stress that applying IS-NMF to the *power* spectrogram or KL-NMF to the *magnitude* spectrogram can be motivated by the equivalence with ML estimation of $\mathbf{W}$ and $\mathbf{H}$ in an assumed generative model of respectively $\mathbf{x}_n$ or $|\mathbf{x}_n|$. In contrast applying IS-NMF to the magnitude spectrogram (as considered in Virtanen (2007)) or KL-NMF to the power spectrogram (as considered in Févotte et al. (2009)) pertains to an ad-hoc choice that cannot be motivated by a sound statistical model of the observed data. Then, the results presented in Section 2.3 tend to show that the IS-NMF model described by equations (7) and (8) is a more suitable model than the KL-NMF model described by equations (14) & (15) because it leads to better decompositions, in the sense that the semantics revealed by the factors $\mathbf{W}$ and $\mathbf{H}$ and the reconstructed components is closer to our own comprehension of sound.

2.2 Algorithms for IS-NMF

MATLAB implementations of the algorithms described next are available from `http://www.tsi.enst.fr/~fevotte`.

2.2.1 Multiplicative updates

A very popular optimization strategy in NMF is based on multiplicative updates. In this iterative scheme $\mathbf{W}$ and $\mathbf{H}$ are optimized alternatively. Each matrix is updated through a gradient descent, where the step size is analytically chosen so that the update becomes multiplicative. More precisely, the approach is equivalent to updating each coefficient θ of $\mathbf{W}$ or $\mathbf{H}$ by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion with respect to this parameter, namely $\theta \leftarrow \theta.[\nabla f(\theta)]_-/[\nabla f(\theta)]_+$, where $\nabla f(\theta) = [\nabla f(\theta)]_+ - [\nabla f(\theta)]_-$ and the summands are both nonnegative. This ensures nonnegativity of the parameter updates, provided initialization with a nonnegative value. A fixed point θ^* of the algorithm implies either $\nabla f(\theta^*) = 0$ or $\theta^* = 0$. In the IS case, the gradients of $D(\mathbf{V}|\mathbf{WH})$ write

$$\nabla_{\mathbf{H}}\, D_{IS}(\mathbf{V}|\mathbf{WH}) = \mathbf{W}^T \left((\mathbf{WH})^{\cdot[-2]} . (\mathbf{WH} - \mathbf{V}) \right) \tag{18}$$

$$\nabla_{\mathbf{W}}\, D_{IS}(\mathbf{V}|\mathbf{WH}) = \left((\mathbf{WH})^{\cdot[-2]} . (\mathbf{WH} - \mathbf{V}) \right) \mathbf{H}^T \tag{19}$$

Algorithm 1 IS-NMF/MU

Input : nonnegative matrix $\mathbf{V}$
Output : nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{V} \approx \mathbf{WH}$
Initialize $\mathbf{W}$ and $\mathbf{H}$ with nonnegative values
for $l = 1 : n_{iter}$ **do**
$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T((\mathbf{WH})^{\cdot[-2]}.\mathbf{V})}{\mathbf{W}^T(\mathbf{WH})^{\cdot[-1]}}$
$\mathbf{W} \leftarrow \mathbf{W}.\frac{((\mathbf{WH})^{\cdot[-2]}.\mathbf{V})\,\mathbf{H}^T}{(\mathbf{WH})^{\cdot[-1]}\,\mathbf{H}^T}$
Normalize $\mathbf{W}$ and $\mathbf{H}$
end for

where . denotes Hadamard entrywise product and $\mathbf{A}^{\cdot[n]}$ denotes the matrix with entries $[\mathbf{A}]_{ij}^n$. This leads to Algorithm 1, in which $\frac{\mathbf{A}}{\mathbf{B}}$ denotes the matrix $\mathbf{A}.\mathbf{B}^{\cdot[-1]}$. This algorithm includes a normalization step at every iteration, which eliminates trivial scale indeterminacies leaving the cost function unchanged. We impose $\|\mathbf{w}_k\|_1 = 1$ and scale h_k accordingly.

The multiplicative approach to NMF was proposed by Lee and Seung (2001) for the Euclidean and KL costs, and the simplicity of the algorithm structure played a key role in the popularization of NMF. Using the convexity of $d_{EUC}(x|y)$ and $d_{KL}(x|y)$ as functions of y, the criterion $D(\mathbf{V}|\mathbf{WH})$ can be shown non-decreasing under the multiplicative rules (Lee and Seung, 2001; Kompass, 2007; Févotte and Cemgil, 2009). The proof, based on the introduction of an auxiliary function, does not however hold for the IS divergence because it is not convex, but in practice we always observed monotonicity of the criterion under Algorithm 1.

2.2.2 Expectation maximization

The statistical composite model underlying IS-NMF can readily serve for the construction of an EM algorithm for the maximization of the likelihood of $\mathbf{W}$ and $\mathbf{H}$, or equivalently the minimization of $D_{IS}(|\mathbf{X}|^2|\mathbf{WH})$. Denoting $\mathbf{C}_k$ the $F \times N$ matrix with coefficients $c_{k,fn}$ and $\mathbf{C}$ the tensor with slices $\mathbf{C}_k$, the latent components $\mathbf{C}$ are a natural choice of complete data. More precisely, each component $\mathbf{C}_k$ can act as a complete data space for the subset of parameters $\boldsymbol{\theta}_k = \{\mathbf{w}_k, h_k\}$. The EM functional to minimize iteratively writes

$$Q^{ML}(\boldsymbol{\theta}|\boldsymbol{\theta}') \stackrel{\text{def}}{=} -\int_{\mathbf{C}} \log p(\mathbf{C}|\boldsymbol{\theta})\, p(\mathbf{C}|\mathbf{X}, \boldsymbol{\theta}')\, d\mathbf{C}. \tag{20}$$

Using conditional independence

$$p(\mathbf{C}|\boldsymbol{\theta}) = \prod_k p(\mathbf{C}_k|\boldsymbol{\theta}_k) \tag{21}$$

the EM functional can be written

$$Q^{ML}(\boldsymbol{\theta}|\boldsymbol{\theta}') \quad = \quad \sum_k Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}'), \tag{22}$$

where

$$Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \stackrel{\text{def}}{=} -\int_{\mathbf{C}_k} \log p(\mathbf{C}_k|\boldsymbol{\theta}_k)\, p(\mathbf{C}_k|\mathbf{X}, \boldsymbol{\theta}')\, d\mathbf{C}_k. \tag{23}$$

Finally, under stated independent assumptions the functional is further reduced to

$$Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') = -\sum_{n,f} \int_{c_{k,fn}} \log p(c_{k,fn}|\boldsymbol{\theta}_k)\, p(c_{k,fn}|x_{fn}, \boldsymbol{\theta}')\, d\, c_{k,fn}. \tag{24}$$

The minus log-likelihood of the complete data writes

$$-\log p(c_{k,fn}|\boldsymbol{\theta}_k) \quad \stackrel{\text{c}}{=} \quad \log(w_{fk}h_{kn}) + \frac{|c_{k,fn}|^2}{w_{fk}h_{kn}} \tag{25}$$

and its posterior distribution writes

$$p(c_{k,fn}|x_{fn}, \boldsymbol{\theta}) = \mathcal{N}(c_{k,fn}|\mu_{k,fn}^{post}, \lambda_{k,fn}^{post}) \tag{26}$$

Algorithm 2 IS-NMF/EM

Input : nonnegative matrix $\mathbf{V}$
Output : nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{V} \approx \mathbf{WH}$
Initialize $\mathbf{W}$ and $\mathbf{H}$ with nonnegative values
for $l = 1 : n_{iter}$ **do**
  for $k = 1 : K$ **do**
    Compute $\mathbf{G}_k = \frac{\mathbf{w}_k h_k}{\mathbf{WH}}$ `% Wiener gain`
    Compute $\mathbf{V}_k = \mathbf{G}_k^{.[2]}.\mathbf{V} + (1 - \mathbf{G}_k).(\mathbf{w}_k h_k)$ `% Posterior power of` $\mathbf{C}_k$
    $h_k \leftarrow \frac{1}{F}(\mathbf{w}_k^{.[-1]})^T \mathbf{V}_k$ `% Update row` k `of` $\mathbf{H}$
    $\mathbf{w}_k \leftarrow \frac{1}{N}\mathbf{V}_k \, (h_k^{.[-1]})^T$ `% Update column` k `of` $\mathbf{W}$
    Normalize $\mathbf{w}_k$ and h_k
  end for
end for

`% Note that` $\mathbf{WH}$ `needs to be computed only once, at initialization, and be subsequently updated as` $\mathbf{WH} - \mathbf{w}_k^{old} h_k^{old} + \mathbf{w}_k^{new} h_k^{new}$.

with

$$\mu_{k,fn}^{post} = \frac{w_{fk}\, h_{kn}}{\sum_l w_{fl}\, h_{ln}}\, x_{fn}, \quad \lambda_{k,fn}^{post} = \frac{w_{fk}\, h_{kn}}{\sum_l w_{fl}\, h_{ln}} \sum_{l \neq k} w_{fl}\, h_{ln}. \tag{27}$$

Plugging equation (25) into equation (24), computing and solving the gradients $Q_k^{ML}(\mathbf{w}_k, h_k|\boldsymbol{\theta}')$ finally leads to

$$h_{kn} = \frac{1}{F} \sum_f \frac{v'_{k,fn}}{w_{fk}} \tag{28}$$

$$w_{fk} = \frac{1}{N} \sum_n \frac{v'_{k,fn}}{h_{kn}}, \tag{29}$$

where $v_{k,fn} \overset{\text{def}}{=} |\mu_{k,fn}^{post}|^2 + \lambda_{k,fn}^{post}$ is the posterior power of $c_{k,fn}$ and the apostrophe refers to parameter values from previous iteration. Note that these update equations differ from the multiplicative updates given in Algorithm 1. In practice the subsets $\boldsymbol{\theta}_k$ are updated sequentially and the set of all parameters $\boldsymbol{\theta}$ is refreshed hereafter, leading to the Space-Alternating Generalized Expectation-Maximization (SAGE) algorithm described in (Févotte et al., 2009). The general EM algorithm is summarized in Algorithm 2 and its convergence to a stationary point of $D_{IS}(\mathbf{V}|\mathbf{WH})$ is granted by property of EM.

2.3 Case study

In this section we wish to establish the relevance of decomposing audio with IS-NMF of the power spectrogram as compared with KL-NMF of the magnitude spectrogram. To that purpose we consider a well structured simple piano sequence. The sequence, played from score given in figure 3 on a Yamaha DisKlavier MX100A upright piano, was recorded in a small size room by a Schoeps omnidirectional microphone, placed about 15 cm (6 in) above the opened body of the piano. The sequence is composed of 4 notes, played all at once in the first measure and then played by pairs in all possible combinations in the subsequent measures. The 15.6 seconds long recorded signal was downsampled to $\nu_s = 22050$ Hz, yielding $T = 339501$ samples. A STFT $\mathbf{X}$ of x was computed using a sinebell analysis window of length $L = 1024$ (46 ms) with 50 % overlap between two frames, leading to $N = 674$ frames and $F = 513$ frequency bins. The time-domain signal x and its log-power spectrogram are represented on figure 3.

NMF of the power spectrogram $|\mathbf{X}|^2$ of this piano sequence with various model orders K and with either Euclidean, IS or KL cost function was thoroughly studied in (Févotte et al.,

2009). Here, given the fixed number of components $K = 8$, we rather compare the decomposition results of IS-NMF of $|\mathbf{X}|^2$ and KL-NMF of $|\mathbf{X}|$; our motivation is to compare the decomposition results obtained from the two statistical composite models discussed in Section 2.1.3, based either on the Poisson modeling of $|\mathbf{c}_{k,n}|$ or the Gaussian modelling of $\mathbf{c}_{k,n}$. In each case we run the corresponding multiplicative NMF algorithm for 5000 iterations. To reduce the impact of initialization and the odds of obtaining a local solution, we run the algorithms 10 times in both case, from random nonnegative initializations of $\mathbf{W}$ and $\mathbf{H}$ and select the factorization with lowest final cost value. The learnt factor matrices $\mathbf{W}$ and $\mathbf{H}$ as well as the reconstructed time components $\hat{c}_k(t)$ are shown in figure 4 and 5; the results are displayed by decreasing variance of $\hat{c}_k(t)$. The pitch estimator described in (Vincent et al., 2007; Févotte et al., 2009) was applied to the elements of the dictionary in order to inspect correctness of pitch with ground truth. In our implementation the MIDI range was sampled from 20.6 to 108.4 with step 0.2. An arbitrary pitch value of 0 is given to unpitched dictionary elements; as a matter of fact the pitch estimator returns maximum value 108.4 for these elements and the unpitchedness is confirmed by looking at the dictionary and listening to the reconstructed components. The pitch estimates are reported in the captions of figure 4 and 5. The reconstructed components can be listened to online at `http://www.tsi.enst.fr/~fevotte/Samples/machine-audition/`.

The decomposition produced by IS-NMF of $|\mathbf{X}|^2$ is as follows. The first four components each capture one of the four notes. The fifth component captures the sound produced by the hammer hitting the strings, the sixth component captures the sound produced by the sustain pedal when it is released, seventh and eighth components contain inaudible residual noise. The pitch estimates perfectly correspond to the ground truth.

The decomposition produced by KL-NMF of $|\mathbf{X}|$ is as follows. The first four components each capture one of the four notes. The fifth component captures both the sound of hammer hits and pedal releases; the hammer hits are followed by an unnatural tremolo effect, and are thus not as well localized in time as with the IS decomposition. The following two components are less interpretable. The sixth component is pitched and seems to capture the decay part of all 4 notes. The seventh component is made of "breath" sound localized at the hammer hits and pedal release occurrences. The eighth component contains faintly audible residual noise. The pitch estimates of the first four components are close to the ground truth values, but not exactly so.

This experimental study shows that the nature of the decomposition obtained with IS-NMF of the power spectrogram of this well structured example is in accord with our own comprehension of sound. Entities with well-defined semantics emerge from the decomposition (individual notes, hammer hits, pedal releases, residual noise) while the decomposition obtained from KL-NMF of the magnitude spectrogram is less interpretable from this perspective. More experiments with the same dataset are described in (Févotte et al., 2009) and further illustrate the relevance of the IS-NMF model for sound decomposition.

3 Bayesian extensions to Itakura-Saito NMF

3.1 Bayesian NMF

The ML likelihood framework presented above inherently assumes $\mathbf{W}$ and $\mathbf{H}$ to be deterministic parameters with no prior information available. In this section we turn to a Bayesian setting where the parameters are given prior distributions $p(\mathbf{W})$ and $p(\mathbf{H})$, reflecting prior beliefs such as smoothness, sparsity, structure, etc. Bayesian inference revolves around the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ of the set of all unknown parameters : information about $\boldsymbol{\theta}$ or subsets of $\boldsymbol{\theta}$ is inferred from the data through manipulation of the posterior. As such, typical point estimates are the maximum a posteriori (MAP) estimate $\hat{\boldsymbol{\theta}}_{MAP} = \text{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathbf{X})$ and the MMSE estimate $\hat{\boldsymbol{\theta}}_{MMSE} = \mathbb{E}\{\boldsymbol{\theta}|\mathbf{X}\} = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{X})\, d\boldsymbol{\theta}$.

While we present in the following Bayesian extensions of IS-NMF, it is worth mentioning related Bayesian treatments of KL-NMF. In particular sparse priors have been considered in (Shashanka

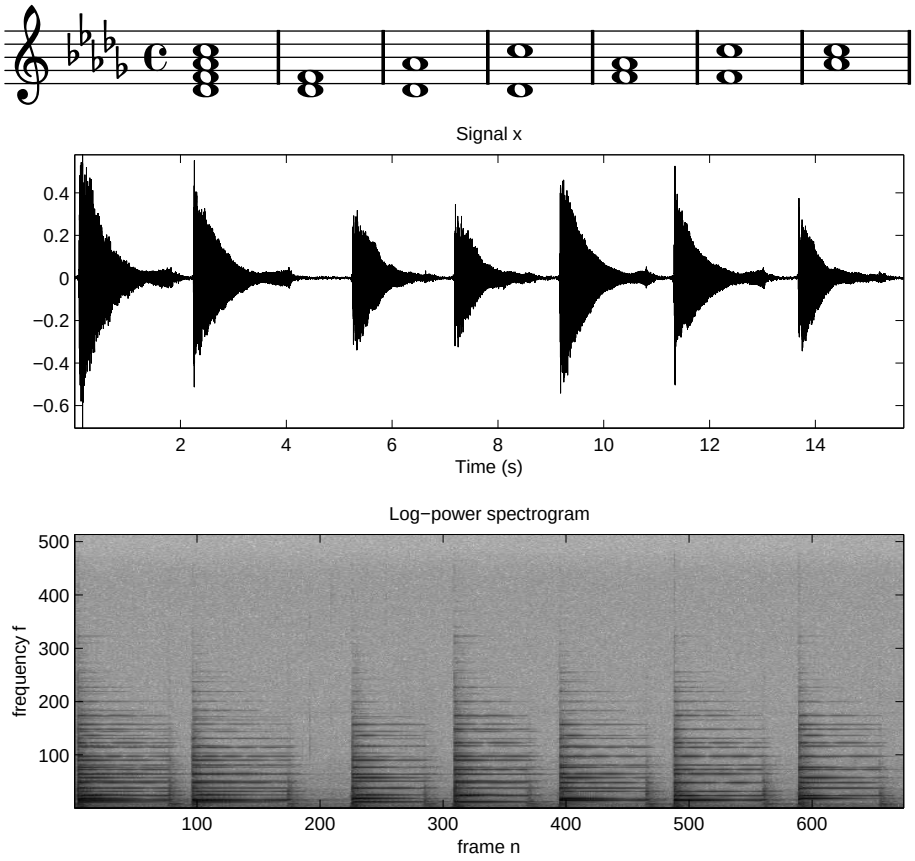

Figure 3: Three representations of data; (top): original score, (middle): time-domain recorded signal x, (bottom): log-power spectrogram $\log|\mathbf{X}|^{\cdot[2]}$. The four notes read D_4^b (pitch 61), F_4 (pitch 65), A_4^b (pitch 68) and C_5 (pitch 72). They all together form a D^b major seventh chord. In the recorded interpretation the third chord is slightly out of tempo. Figure reproduced from (Févotte et al., 2009).

et al., 2008; Cemgil, 2009) and regularization constraints enforcing smoothness of the rows of $\mathbf{H}$ have been considered in (Virtanen et al., 2008).

3.1.1 MAP estimation

In our NMF setting, MAP estimates of $\mathbf{W}$ and $\mathbf{H}$ are sought through minimization of

$$C_{MAP}(\mathbf{W},\mathbf{H}) \stackrel{\text{def}}{=} -\log p(\mathbf{W},\mathbf{H}|\mathbf{X}) \tag{30}$$

$$\stackrel{\text{c}}{=} D_{IS}(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{W}) - \log p(\mathbf{H}) \tag{31}$$

When the priors $p(\mathbf{W})$ and $p(\mathbf{H})$ depend themselves on parameters, so-called *hyperparameters*, they can be included in the MAP criterion and optimized over as well; this will be considered in one of the examples below.

Multiplicative updates Standard NMF multiplicative updates may be attempted provided the gradient of the priors can be expressed as the difference of two positive functions. In that case the resulting updates for $\mathbf{H}$ take the following form

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{[\nabla_{\mathbf{H}} D_{IS}(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{H})]_-}{[\nabla_{\mathbf{H}} D_{IS}(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{H})]_+} \tag{32}$$

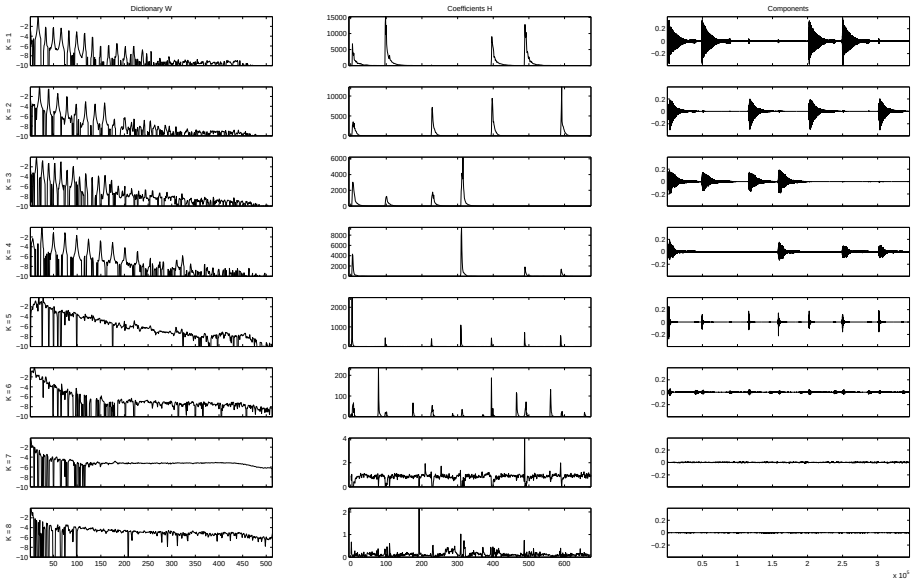

Figure 4: IS-NMF of $|\mathbf{X}|^2$ with $K = 8$. Left : columns of $\mathbf{W}$ ($\log_{10}$ scale). Middle : rows of $\mathbf{H}$. Right : Component reconstruction with equation (12). Pitch estimates : [65.0 68.0 61.0 72.0 0 0 0 0]. Top to down display by decreasing variance of the reconstructed components.

and similar updates are obtained for $\mathbf{W}$. While this scheme was reported to decrease the MAP criterion for various pairs of cost function and priors, see e.g (Cichocki et al., 2006) and Section 3.2, there is yet no theoretical guarantee of such a property.

EM optimization In contrast, the structure of the EM algorithm can accommodate MAP estimation, with guaranteed convergence. In the MAP setting, the functional to optimize iteratively writes

$$Q^{MAP}(\boldsymbol{\theta}|\boldsymbol{\theta}') \stackrel{\text{def}}{=} -\int_{\mathbf{C}} \log p(\boldsymbol{\theta}|\mathbf{C})\, p(\mathbf{C}|\mathbf{X}, \boldsymbol{\theta}')\, d\mathbf{C} \tag{33}$$

$$\stackrel{\text{c}}{=} Q^{ML}(\boldsymbol{\theta}|\boldsymbol{\theta}') - \log p(\mathbf{W}) - \log p(\mathbf{H}) \tag{34}$$

Hence, if we assume independent priors such that $-\log p(\mathbf{H}) = -\sum_k \log p(h_k)$ and $-\log p(\mathbf{W}) = -\sum_k \log p(\mathbf{w}_k)$, the functional writes

$$Q^{MAP}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \sum_k Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \tag{35}$$

where

$$Q_k^{MAP}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \stackrel{\text{def}}{=} Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') - \log p(\mathbf{w}_k) - \log p(h_k) \tag{36}$$

Hence, the E-step still amounts to computing $Q_k^{ML}(\mathbf{w}_k, h_k|\boldsymbol{\theta}')$, as done in Section 2.2.2, and only the M-step is changed by the regularization constraints $-\log p(\mathbf{w}_k)$ and $-\log p(h_k)$ which now need to be taken into account. However, the addition of the prior terms may render the M-step more complex, with in some cases no analytical solutions. Hence, a compromise may need to be made between the complexity of the prior structure and the optimization it leads to, so as to keep inference tractable.

3.1.2 Sampling the posterior distribution

The above EM algorithm for MAP estimation can be easily adapted so as to produce a MCMC sampling algorithm. Sampling the posterior distribution, i.e, generating realizations of $p(\boldsymbol{\theta}|\mathbf{X})$, is

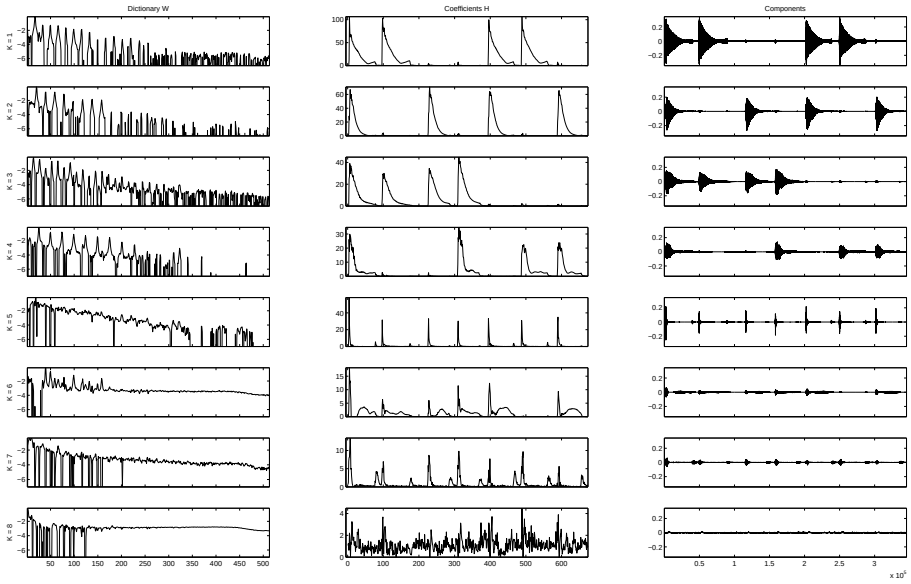

Figure 5: KL-NMF of $|\mathbf{X}|$ with $K = 8$. Left : columns of $\mathbf{W}$ ($\log_{10}$ scale). Middle : rows of $\mathbf{H}$. Right : Component reconstruction with equation (17). Pitch estimates : [65.2 68.2 61.0 72.2 0 56.2 0 0]. Top to down display by decreasing variance of the reconstructed components.

interesting because it yields a more detailed characterization of the posterior distribution than mere point estimates. In particular, confidence intervals and error variances can readily be computed from the samples. In most settings sampling directly from the posterior distribution is difficult and a common strategy consists of producing a Markov chain whose stationary distribution is $p(\boldsymbol{\theta}|\mathbf{X})$, yielding the family of so-called Markov chain Monte Carlo (MCMC) algorithms. Parameter values $\boldsymbol{\theta}^{(l)}$ are generated iteratively according to a Markov chain kernel $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ until convergence is achieved. The problem then boils down to how to generate a Markov chain having the desired stationary distribution $p(\boldsymbol{\theta}|\mathbf{X})$.

In our case, a Space Alternating Data Augmentation (SADA) algorithm can readily be employed. SADA is a stochastic variant of SAGE, described in (Doucet et al., 2005). Instead of computing the expectation of the log parameter posterior with respect to the hidden data posterior, SADA generates, at each iteration (l) of the sampler and sequentially over k, a realization $\mathbf{C}_k^{(l)}$ from the (instrumental) distribution $p(\mathbf{C}_k|\mathbf{X}, \boldsymbol{\theta}^{(l-1)})$, and then update the parameters by generating a realization $\boldsymbol{\theta}_k^{(l)}$ from $p(\boldsymbol{\theta}_k|\mathbf{C}_k^{(l)})$. Following a "burn in" period, the realizations of $\boldsymbol{\theta}$ obtained from the sampler are drawn from the desired posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$.

MCMC algorithms are more computationally demanding than their EM counterpart. The cost per iteration of each algorithm may be similar as sampling variables from their conditional posterior (MCMC) is somehow as expensive as computing the value of the mode (EM), provided the distribution is easy to sample. But the convergence of MCMC algorithms is usually attained in many more iterations than EM, as the former thoroughly explores the posterior space while EM points at nearest local minimum. As such, a very important characteristic of MCMC algorithms is that they are theoretically less prone to local convergence : over an infinitely large number of iterations, the MCMC algorithm is guaranteed to explore the full posterior while EM will stay trapped in the nearest local minimum. In practice however, MCMC convergence may be slow and the sampler may get stuck in local modes. A challenging aspect in the design of MCMC algorithms is to produce samplers that "mix" well, i.e, move fast in the posterior space. Other works related to NMF using MCMC techniques can be found in (Moussaoui et al., 2006; Schmidt et al., 2009; Cemgil, 2009; Févotte and Cemgil, 2009) .

3.2 Automatic relevance determination in NMF

3.2.1 Model order selection

A very challenging problem in NMF, as in most decomposition algorithms, is how to determine a suitable number of components K ? There might not be a ground truth value, but one would like to at least be able to select an "efficient" or "relevant number". This a model order selection problem. Usual criteria such as the Bayesian information criterion (BIC) or Akaike's criterion (AIC) - see, e.g., (Stoica and Selén, 2004) - cannot be directly applied to NMF, because the number of parameters $(F\,K + K\,N)$ is not constant with respect to the number of observations N. This feature breaks the validity of the assumptions in which these criteria have been designed. What would be ideally required is ML estimation over $p(\mathbf{V}|\mathbf{W})$ instead of $p(\mathbf{V}|\mathbf{W},\mathbf{H})$, treating $\mathbf{H}$ as a latent variable, like in independent component analysis (MacKay, 1996; Lewicki and Sejnowski, 2000) or in some dictionary learning methods (Kreutz-Delgado et al., 2003). But this is nontrivial as this likelihood function is not always easily computed, in particular in our case.

The Bayesian setting offers ways to handle the model order selection problem. In essence, the BIC and AIC criteria form approximations of the model marginal likelihood. Denoting $\mathcal{M}_K$ the model consisting of assuming that the dimensionality of the dictionary $\mathbf{W}$ is K, the marginal likelihood, sometimes also referred to as *evidence*, writes

$$p(\mathbf{X}|\mathcal{M}_K) = \int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta},\mathcal{M}_K)p(\boldsymbol{\theta}|\mathcal{M}_K)d\boldsymbol{\theta}. \tag{37}$$

Model order selection can be envisaged by computing the marginal likelihood of each considered value of K and then form a decision based on the largest value of $\{p(\mathbf{X}|\mathcal{M}_K)\}_K$. However, for most models the integral involved in equation (37) is difficult to compute and various approximations or numerical optimization strategies have been considered for this task. One of them, the Chibbs method (Chib, 1995), allows to compute (37) numerically using the output of a MCMC sampler. This method was in particular considered for KL-NMF in (Cemgil, 2009).

3.2.2 Automatic relevance determination

Hence, we now turn to another yet more simpler Bayesian method based on *automatic relevance determination* (ARD). ARD was introduced by Mackay (1995) in the context of regression to assess a relevant number of explanatory variables. It was then considered by Bishop (1999) to determine an "efficient" number of components in PCA and by Tipping (2001) with similar motives in sparse Bayesian learning. Our idea, originally described in (Tan and Févotte, 2009), is to place priors, dependent on variance-like "relevance" parameters $\boldsymbol{\phi} = [\phi_1,\ldots,\phi_K]$, on both the columns of $\mathbf{W}$ and the rows of $\mathbf{H}$. More precisely, we tie $\mathbf{w}_k$ and h_k together through the following priors :

$$p(\mathbf{w}_k|\phi_k) = \prod_f \mathcal{HN}(w_{fk}|\phi_k), \tag{38}$$

$$p(h_k|\phi_k) = \prod_n \mathcal{HN}(h_{kn}|\phi_k), \tag{39}$$

where $\mathcal{HN}(x,\sigma^2)$ is the half-normal distribution defined in Appendix A. Note that this prior is not overconstraining the scales, because of the scale indeterminacy between $\mathbf{w}_k$ and h_k. We also consider a conjugate inverse-Gamma prior $\mathcal{IG}(\phi_k|\alpha,\beta)$ for the relevance parameters, whose influence will be discussed later. Then we seek MAP estimates of $\mathbf{W}$, $\mathbf{H}$ and $\boldsymbol{\phi}$ through minimization of

$$\begin{aligned} C_{MAP}(\mathbf{W},\mathbf{H},\boldsymbol{\phi}) &\overset{\text{def}}{=} -\log p(\mathbf{W},\mathbf{H},\boldsymbol{\phi}|\mathbf{V}), \\ &\overset{\text{c}}{=} -\log p(\mathbf{V}|\mathbf{W},\mathbf{H}) - \log p(\mathbf{W}|\boldsymbol{\phi}) - \log p(\mathbf{H}|\boldsymbol{\phi}) - \log p(\boldsymbol{\phi}), \\ &\overset{\text{c}}{=} D_{IS}(\mathbf{V}|\mathbf{WH}) + \sum_k \left[\frac{1}{2}\left(\sum_f w_{fk}^2 + \sum_n h_{kn}^2\right) + \beta\right]\frac{1}{\phi_k} \\ &\quad + \left(\frac{F+N}{2} + \alpha + 1\right)\log\phi_k \end{aligned} \tag{40}$$

Algorithm 3 Automatic relevance determination for IS-NMF

Input : Nonnegative matrix $\mathbf{V}$, fixed hyperparameters α, β.
Output : Nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W}\,\mathbf{H}$, nonnegative vector $\boldsymbol{\phi}$.
Initialize $\mathbf{W}$ and $\mathbf{H}$ with nonnegative values.
for $l = 1 : n_{iter}$ **do**
$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T\,((\mathbf{WH})^{\cdot[-2]}.\mathbf{V})}{\mathbf{W}^T\,(\mathbf{WH})^{\cdot[-1]}+\mathrm{diag}(\boldsymbol{\phi}^{\cdot[-1]})\,\mathbf{H}}$
$\mathbf{W} \leftarrow \mathbf{W}.\frac{((\mathbf{WH})^{\cdot[-2]}.\mathbf{V})\,\mathbf{H}^T}{(\mathbf{WH})^{\cdot[-1]}\,\mathbf{H}^T+\mathbf{W}\,\mathrm{diag}(\boldsymbol{\phi}^{\cdot[-1]})}$
$\phi_k \leftarrow \frac{\sum_f w_{fk}^2+\sum_n h_{kn}^2+2\beta}{F+N+2(\alpha+1)}, \quad k = 1, \ldots, K.$
end for

As a result of inference, a subset of the relevance parameters $\boldsymbol{\phi}$ corresponding to irrelevant components is driven to a small lower bound, with the corresponding columns of $\mathbf{W}$ and rows of $\mathbf{H}$ driven to small values, by equations (38) and (39). The value of the lower bound depends on the hyperparameters α and β, see Algorithm 3, which have to be user-defined. As can be seen from the MAP function given at equation (40) setting these variables to low values leads to a noninformative prior (i.e, the data is let to speak for itself).

The gradients of cost function (40) leads to

$$\nabla_{\mathbf{W}} C_{MAP}(\mathbf{W},\mathbf{H},\boldsymbol{\phi}) = \Big((\mathbf{WH})^{\cdot[-2]}.(\mathbf{WH}-\mathbf{V})\Big)\,\mathbf{H}^T + \mathbf{W}\mathrm{diag}(\boldsymbol{\phi}^{\cdot[-1]}), \quad (41)$$

$$\nabla_{\mathbf{H}} C_{MAP}(\mathbf{W},\mathbf{H},\boldsymbol{\phi}) = \mathbf{W}^T\,\Big((\mathbf{WH})^{\cdot[-2]}.(\mathbf{WH}-\mathbf{V})\Big) + \mathrm{diag}(\boldsymbol{\phi}^{\cdot[-1]})\mathbf{H}, \quad (42)$$

$$\nabla_{\phi_k} C_{MAP}(\mathbf{W},\mathbf{H},\boldsymbol{\phi}) = -\left[\frac{1}{2}\left(\sum_f w_{fk}^2 + \sum_n h_{kn}^2\right)+\beta\right]\frac{1}{\phi_k^2} + \left(\frac{F+N}{2}+\alpha+1\right)\frac{1}{\phi_k} \quad (43)$$

and the multiplicative strategy described in Section 3.1.1 leads to Algorithm 3.

3.2.3 Results

We applied Algorithm 3 to the piano data considered in Section 2.3. We set the number of iterations to 5000 and as before, in order to marginalize the influence of local minima, we run the ARD algorithm from different random initializations and selected the factorization with lowest final MAP criterion value. We set the initial number of components K to 25 and we set $\alpha = 0$ and $\beta = 0.001$, yielding a rather noninformative prior. The convergence and final values of the relevance parameters ϕ_k are displayed on figure 6. We see that only 6 components emerge. Inspection of these six components (not shown here) reveals that they essentially correspond to the first six components displayed in figure 4, i.e, the four notes, hammer hits and pedal releases. We wish to point out that the problem of convergence to a local solution seemed even more severe here, probably due to the large dimension $K = 25$ of the hidden space, and that we experienced numerical difficulties with Algorithm 3 for too low values of β. However the factorizations given by the 10 runs all agreed upon a number of relevant components in the 5-7 range, so that a solution of practical value to avoid the enhanced problem of local convergence can consist in applying the ARD algorithm first, so as to obtain an approximation of the relevant number of components, and then re-run the straight IS-NMF algorithm (possibly from various initializations) to improve the correctness of the decomposition.

3.3 IS-NMF with Markov prior

If a reasonable amount of literature has been devoted to NMF with sparse constraints (either on $\mathbf{W}$ or $\mathbf{H}$), little literature is dealing with persistence constraints. However if $\mathbf{V}$ is indexed by time (such as time-frequency spectra), it is expected that $\mathbf{h}_n$ is correlated with $\mathbf{h}_{n-1}$ and this information should be integrated in the NMF model. In (Févotte et al., 2009), we have described a

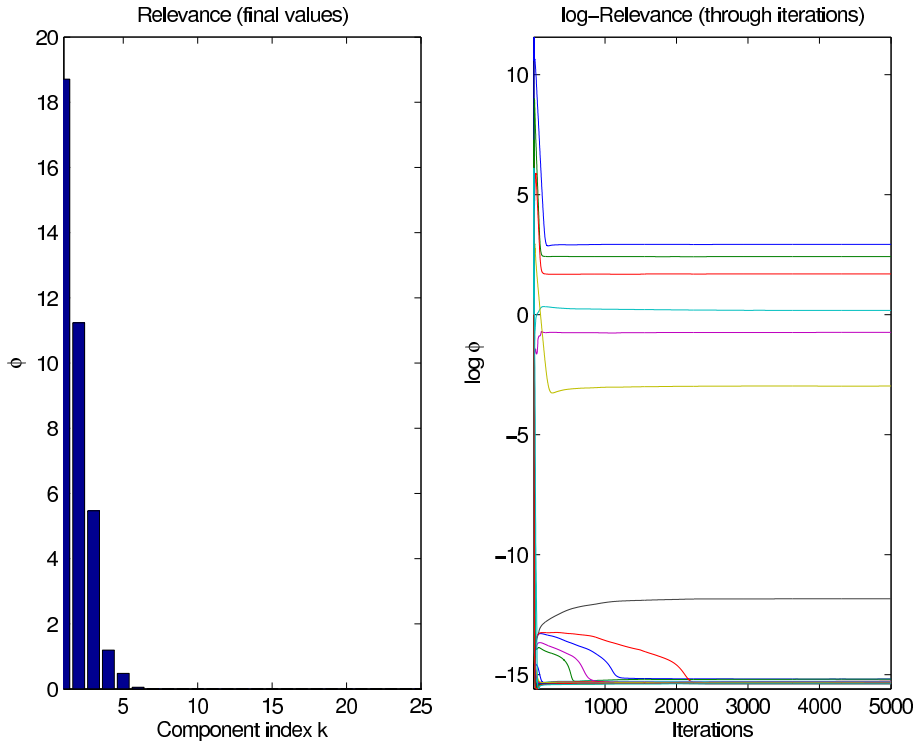

Figure 6: Values of the relevance parameters $\phi_1, \ldots, \phi_K$, of the $K = 25$ components, by decreasing order. The six most relevant components correspond to the first six components on figure 4, but in a different order.

IS-NMF algorithm accounting for Markov structure of the rows of $\mathbf{H}$. In this setting the following prior structure is assumed for h_k :

$$p(h_k) = \prod_{n=2}^{N} p(h_{kn}|h_{k(n-1)})\; p(h_{k1}), \tag{44}$$

where the Markov kernel $p(h_{kn}|h_{k(n-1)})$ is a pdf with mode $h_{k(n-1)}$. The motivation behind this prior is to constrain h_{kn} not to differ significantly from its value at entry $n-1$, hence favoring smoothness of the estimate. A possible pdf choice is, for $n = 2, \ldots, N$,

$$p(h_{kn}|h_{k(n-1)}) = \mathcal{IG}(h_{kn}|\alpha, (\alpha+1)\, h_{k(n-1)}) \tag{45}$$

where $\mathcal{IG}(x|\alpha, \beta)$ is the inverse-Gamma pdf defined in Appendix A, with mode $\beta/(\alpha+1)$, and variance $\beta^2/[(\alpha-1)^2(\alpha-2)](for \alpha > 2)$. The prior is constructed so that its mode is obtained for $h_{kn} = h_{k(n-1)}$. α is a shape parameter that controls the sharpness of the prior around its mode. A high value of α will increase sharpness and will thus accentuate smoothness of h_k while a low value of α will render the prior more diffuse and thus less constraining. More details as well as presentation of Gamma Markov chains can be found in (Févotte et al., 2009). In the following, h_{k1} is assigned the scale-invariant Jeffreys noninformative prior $p(h_{k1}) \propto 1/h_{k1}$.

3.3.1 EM algorithm

Under prior structure (44), the derivative of $Q_k^{MAP}(\mathbf{w}_k, h_k|\boldsymbol{\theta}')$ with respect to h_{kn} writes, $\forall n = 2, \ldots, N-1$,

$$\begin{aligned}\nabla_{h_{kn}} Q_k^{MAP}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') = \\ \nabla_{h_{kn}} Q_k^{ML}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') - \nabla_{h_{kn}} \log p(h_{k(n+1)}|h_{kn}) - \nabla_{h_{kn}} \log p(h_{kn}|h_{k(n-1)})\end{aligned} \tag{46}$$

inverse-Gamma Markov chain			
	p_2	p_1	p_0
h_{k1}	$(\alpha+1)/h_{k2}$	$F-\alpha+1$	$-F\,\hat{h}_{k1}^{ML}$
h_{kn}	$(\alpha+1)/h_{k(n+1)}$	$F+1$	$-F\,\hat{h}_{kn}^{ML}-(\alpha+1)\,h_{k(n-1)}$
h_{kN}	0	$F+\alpha+1$	$-F\,\hat{h}_{kN}^{ML}-(\alpha+1)\,h_{k(N-1)}$

Table 1: Coefficients of the order 2 polynomial to solve in order to update h_{kn} in Bayesian IS-NMF with an inverse-Gamma Markov chain prior. $\hat{h}_{kn}^{ML}$ denotes the ML update, given by equation (28).

Algorithm 4 IS-NMF with inverse-Gamma Markov prior for $\mathbf{H}$

Input : nonnegative matrix $\mathbf{V}$
Output : nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{V} \approx \mathbf{WH}$
Initialize $\mathbf{W}$ and $\mathbf{H}$ with nonnegative values
for $l = 1 : n_{iter}$ **do**
  for $k = 1 : K$ **do**
    Compute $\mathbf{G}_k = \frac{\mathbf{w}_k h_k}{\mathbf{WH}}$ `% Wiener gain`
    Compute $\mathbf{V}_k = \mathbf{G}_k^{.[2]}.\mathbf{V} + (1-\mathbf{G}_k).(\mathbf{w}_k h_k)$ `% Posterior power of` $\mathbf{C}_k$

    `% Update of` h_k
    Compute $\hat{h}_k^{ML} = \frac{1}{F}(\mathbf{w}_k^{.[-1]})^T\,\mathbf{V}_k$ `% ML estimate of` h_k
    for $n = 1 : N$ **do**
      Compute p_0, p_1, p_2 as in Table 1
      $h_{kn} \leftarrow (\sqrt{p_1^2 - 4\,p_2\,p_0} - p_1)/(2\,p_2)$ `% Smooth ML estimate : MAP estimate of` h_k
    end for

    `% Update of` $\mathbf{w}_k$
    $\mathbf{w}_k \leftarrow \frac{1}{N}\mathbf{V}_k\,(h_k^{.[-1]})^T$ `% ML estimate of` $\mathbf{w}_k$
    Normalize $\mathbf{w}_k$ and h_k
  end for
end for

`% Note that` **WH** `needs to be computed only once, at initialization, and be subsequently updated as` $\mathbf{WH} - \mathbf{w}_k^{old}h_k^{old} + \mathbf{w}_k^{new}h_k^{new}$.

This is shown to be equal to

$$\nabla_{h_{kn}} Q_k^{MAP}(\mathbf{w}_k, h_k|\boldsymbol{\theta}') = \frac{1}{h_{kn}^2}(p_2\,h_{kn}^2 + p_1\,h_{kn} + p_0) \tag{47}$$

where the values of p_0, p_1 and p_2 are given in Table 1. Updating h_{kn} then simply amounts to solving an order 2 polynomial. The polynomial has only one nonnegative root, given by

$$h_{kn} = \frac{\sqrt{p_1^2 - 4\,p_2\,p_0} - p_1}{2\,p_2}. \tag{48}$$

The coefficients h_{k1} and h_{kN} at the borders of the Markov chain require specific updates, but they also only require solving polynomials of order either 2 or 1, with coefficients given in Table 1 as well. The overall IS-NMF approach using an inverse-Gamma Markov prior for the rows of $\mathbf{H}$ is described in Algorithm 4. A MATLAB implementation of this algorithm is available from `http://www.tsi.enst.fr/~fevotte`.

3.3.2 Results

We applied the previous algorithm with an inverse-Gamma Markov chain prior on the rows of $\mathbf{H}$, with $\alpha = 50$, to the piano data presented in Section 2.3. Figure 7 displays on the same graph a zoom

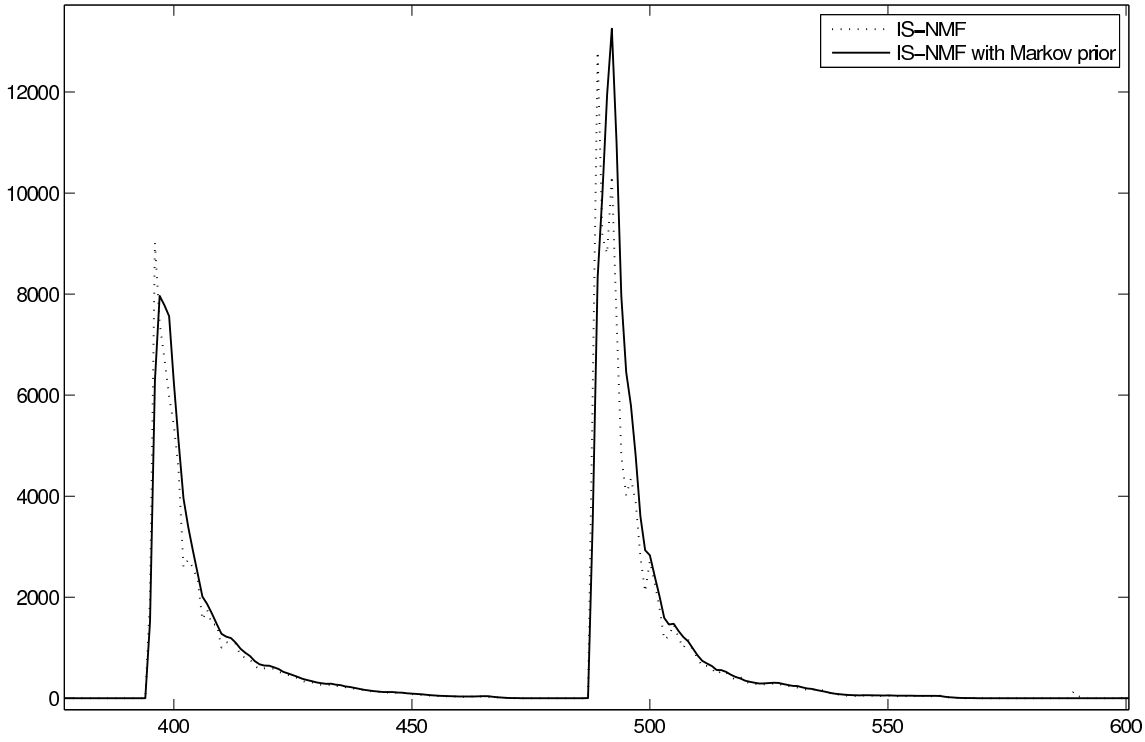

Figure 7: Segment of row h_1 ($\approx 400 - 650$ frame range) as obtained from IS-NMF with the multiplicative nonregularized algorithm (dotted line) and from IS-NMF with a Markov regularization (solid line).

on the row h_1 obtained with the multiplicative algorithm (see figure 4) and with the regularized version. The difference in smoothness is apparent. In (Févotte et al., 2009), both algorithms were considered for the decomposition of a real early jazz music piece and the regularized version of IS-NMF was shown to lead to source estimates much more pleasant to listen to.

3.4 IS-NMF and beyond

The IS-NMF composite model can be extended in many ways in order to either give a more general model to every component, or singularize a particular component. Recall that model (7) takes the audio data x to be a sum of components each characterized by a spectral signature. This model is well suited for polyphonic music, where each frame is a sum of sounds such as notes. This model may however not be entirely relevant for certain sources, such as voice, either spoken or sung. As a matter of fact, voice is often relegated into a specific component, see, e.g, (Benaroya et al., 2006b; Ozerov et al., 2007; Blouet et al., 2008; Durrieu et al., 2010). Voice is not a sum of elementary components but rather a single component with many states, each representative of an elementary sound (phoneme, note). As such, it is more appropriate to model voice with a mixture of distributions, such as a Gaussian Mixture Model (GMM). A relevant voice + music model is then

$$\mathbf{x}_n = \mathbf{s}_{V,n} + \mathbf{s}_{M,n} \tag{49}$$

where $\mathbf{s}_{V,n}$ stands for the voiced component and $\mathbf{s}_{M,n}$ stands for the musical accompaniment. From there, the accompaniment may be given a IS-NMF model $\{\mathbf{W}^M, \mathbf{H}^M\}$ as of equations (7)-(8) :

$$p(\mathbf{s}_{M,n}|\mathbf{W}^M, \mathbf{H}^M) = \mathcal{N}\left(\mathbf{s}_{M,n}|0, \sum_{k=1}^{K} h_{kn}^M \mathrm{diag}(\mathbf{w}_k^M)\right) \tag{50}$$

while the voice component may be given a Gaussian scaled mixture model $\{\mathbf{W}^V, \mathbf{H}^V\}$ (Benaroya et al., 2006b), described by

$$p(\mathbf{s}_{V,n}|\mathbf{W}^V, \mathbf{H}^V) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{s}_{V,n}|0, h_{kn}^V \mathrm{diag}(\mathbf{w}_k^V)) \tag{51}$$

Pay attention that in the accompaniment model a summation occurs on the frame domain, while it occurs on the pdfs of the frames in the voice model. Inference in the model defined by equations (49), (50) and (51) can be carried with an EM algorithm similar to the one described in Section 2.2.2, but where the sequence of states of the voice component is added to the complete dataset, as routinely done in EM estimation of GMMs. The resulting algorithm is described in (Ozerov et al., 2009). The latter paper reports speech + music separation results for which both IS-NMF and GSMM are envisaged for modeling the voice component, while the music component is modeled with IS-NMF. A Source to Distortion Ratio improvement is systematically observed with the GSMM model, corroborating the relevance of the latter model for voice.

Note also that pitchness of the voice component (especially when sung) can further be enforced in the model, as proposed in (Durrieu et al., 2009, 2010). This is done through $\mathbf{W}^V$, which is given a excitation/filter model. More precisely, given a fixed dictionary of harmonic combs $\mathbf{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$ (the "excitations") and a series of unknown envelopes $\mathbf{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_Q\}$ (the "filters"), the dictionary $\mathbf{W}^V$ is made of all excitation-filters combinations $\{\mathbf{e}_m.\mathbf{f}_q\}_{q,m}$, mimicking the human voice production system. Note that in that setting, only the spectral envelopes $\mathbf{F}$ are learnt from the data, and not the entire dictionary $\mathbf{W}^V$. Durrieu et al. (2010) describe both EM and multiplicative updates for inference in generative model (49),(50),(51) with the latter voice production model. The paper reports in particular excellent voice separation results, some of them available from the Signal Separation Evaluation Campaign (SiSEC 2008) webpage.[1]

Finally, we would like to mention an interesting extension of the plain NMF model that has been considered in the literature, using the Euclidean or KL cost functions, and that could readily be considered with the IS divergence. The NMF model (1) basically assumes that the spectrogram (either in power or in magnitude) is decomposed as a linear combination of elementary patterns $\mathbf{w}_k$. These patterns are in essence a spectral characterization of sound objects with time support equal to the STFT window size. While this simple assumption can lead to satisfying decompositions, as shown in this chapter, a more realistic assumption would consist of assuming that the spectrogram is a linear combination of elementary "patches", i.e, patterns with frame-duration larger than one, or equivalently sound objects with support spread over several STFT windows. This framework was coined "convolutive NMF" in (Smaragdis, 2007) and was shown to adequately learn speech phones from mixtures of speech signals. It was furthermore shown to improve NMF-based onset detection in (Wang et al., 2009).

4 Multichannel IS-NMF

The work described in the previous sections inherently assumes the audio data to be single-channel : the spectrogram $\mathbf{V}$ of some unidimensional observation x is computed, factorized and a time-domain decomposition of x is then reconstructed from the factorization. However, most musical data is now available in multichannel format, especially stereo. Each channel is typically a mixture of mutual sources, with different mixing parameters. One approach to the decomposition of such multichannel recordings is to decompose each channel individually, but this is not optimal as the redundancy between the channel is not used. In this section we describe generalizations of NMF to the multichannel case, allowing for joint processing of the channels.

[1] `http://sisec.wiki.irisa.fr/`

4.1 Instantaneous mixing

4.1.1 Generative model

Assume a multichannel recording with I channels $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ ($I = 2$ corresponds to the stereo case). The data has typically been produced as a mixture of a certain number of instrumental sources. In the simplest case the observations are a linear instantaneous mixture of *point-source* signals $s_1(t), \ldots, s_J(t)$, such that

$$x_i(t) = \sum_{j=1}^{J} a_{ij} s_j(t), \tag{52}$$

where the coefficients a_{ij} are real; the instantaneous mixing corresponds to elementary "pan pot" mixing. By linearity of STFT, the multichannel model reads similarly in the time-frequency domain

$$\mathbf{x}_{i,n} = \sum_{j=1}^{J} a_{ij}\, \mathbf{s}_{j,n}, \tag{53}$$

where $\mathbf{x}_{i,n}$ and $\mathbf{s}_{j,n}$ denote the complex-valued STFTs of time-domain signals $x_i(t)$ and $s_j(t)$. In the following we note :

- $\mathbf{X}_i$ the $F \times N$ STFT matrix of channel i, with coefficients $\{x_{i,fn}\}_{fn}$,
- $\mathbf{V}_i$ the $F \times N$ power spectrogram matrix of channel i, with coefficients $\{v_{i,fn}\}_{fn}$ defined as $v_{i,fn} = |x_{i,fn}|^2$,
- $\mathbf{X}$ the $I \times F \times N$ STFT tensor with coefficients $\{x_{i,fn}\}_{ifn}$,
- $\mathbf{S}_i$ the $F \times N$ STFT matrix of source j, with coefficients $\{s_{j,fn}\}_{fn}$,
- $\mathbf{S}$ the $J \times F \times N$ STFT tensor with coefficients $\{s_{j,fn}\}_{jfn}$,
- $\mathbf{A}$ the mixing matrix with coefficients $\{a_{ij}\}_{ij}$, and with rows $\{a_i\}_i$ and columns $\{\mathbf{a}_j\}_j$,
- $\mathbf{Q}$ the "power" mixing matrix with coefficients $\{q_{ij}\}_{ij}$ defined as $q_{ij} \overset{\text{def}}{=} |a_{ij}|^2$.

Now assume a IS-NMF model for each source STFT, i.e,

$$\mathbf{s}_{j,n} = \sum_{k \in \mathcal{K}_j} \mathbf{c}_{k,n} \quad \text{with} \quad \mathbf{c}_{k,n} \sim \mathcal{N}_c(0, h_{kn}\, \mathrm{diag}\,(\mathbf{w}_k)) \tag{54}$$

where $\{\mathcal{K}_1, \ldots, \mathcal{K}_J\}$ denotes a nontrivial partition of $\{1, \ldots, K\}$ and $K \geq J$ is the total number of components used to describe the multichannel data. In the following we note :

- $\mathbf{W}$ and $\mathbf{H}$ the dictionary and activation matrices of sizes, respectively, $F \times K$ and $K \times N$, with coefficients $\{w_{fk}\}$ and $\{h_{kn}\}$, characterizing the components of all sources,
- $\mathbf{W}_j$ and $\mathbf{H}_j$ the dictionary and activation matrices of sizes, respectively, $F \times \#\mathcal{K}_j$ and $\#\mathcal{K}_j \times N$, with coefficients $\{w_{fk}\}_{f,k \in \mathcal{K}_j}$ and $\{h_{kn}\}_{k \in \mathcal{K}_j, n}$, characterizing the components of source j,
- $\mathbf{C}_k$ the $F \times N$ matrix with coefficients $\{c_{k,fn}\}_{fn}$ and $\mathbf{C}$ the $K \times F \times N$ matrix with coefficients $\{c_{k,fn}\}_{kfn}$.

The task envisaged in this section is the estimation of $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{A}$ from $\mathbf{X}$, and then the reconstruction of the components $\mathbf{C}$ and/or sources $\mathbf{S}$. The problem may be coined "multichannel NMF" because we are attempting to estimate the nonnegative factors $\mathbf{W}_j$ and $\mathbf{H}_j$, but not from the sources themselves (because not available) but from mixtures of these sources. The mixing system is so far assumed linear instantaneous, the convolutive case being addressed in the next section.

Under the assumed models, the expectations of the source and channel power spectrograms are given by

$$\mathrm{E}\{|\mathbf{S}_j|^2\} = \mathbf{W}_j\mathbf{H}_j \quad (55)$$

$$\mathrm{E}\{|\mathbf{X}_i|^2\} = \sum_j q_{ij}\,\mathbf{W}_j\mathbf{H}_j \quad (56)$$

$$\stackrel{\text{def}}{=} \mathbf{V}_i \quad (57)$$

Given independence assumptions between all components, the likelihood of *each* channel writes

$$-\log p(\mathbf{X}_i|\mathbf{W},\mathbf{H},a_i) \stackrel{\text{c}}{=} D_{IS}(\mathbf{V}_i|\hat{\mathbf{V}}_i) \quad (58)$$

The approach we propose here to process the channels jointly is to minimize the sum of the individual channel likelihoods, i.e,

$$C(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_i D_{IS}(\mathbf{V}_i|\hat{\mathbf{V}}_i) \quad (59)$$

An algorithm based on multiplicative updates for the optimization of this criterion will be described in Section 4.2.

4.1.2 Statistical implications

We stress that the proposed approach is *not* ML estimation. Indeed our criterion is not equivalent to the full joint likelihood of the observations $-\log p(\mathbf{X}|\boldsymbol{\theta})$, but instead to the sum of the individual likelihoods of each channel $\sum_i -\log p(\mathbf{X}_i|\boldsymbol{\theta})$. Our approach is hence suboptimal, given that the generative data model is true. Optimizing the joint likelihood is somehow more complex than optimizing the mere sum of individual channel likelihoods. A multiplicative algorithm for the latter is described in the next section. An EM algorithm for optimizing the joint likelihood is described in (Ozerov and Févotte, 2010), it is in essence a generalization of the EM algorithm presented at Section 2.2.2 and is based on same complete dataset $\mathbf{C}$. As noted in (Durrieu et al., 2009), $C(\boldsymbol{\theta})$ is actually the real likelihood corresponding to the model where the channel STFTs are assumed mutually independent. To evidence this, note that model (53)-(54) can be rewritten as

$$\mathbf{x}_{i,n} = \sum_{k=1}^{K} a_{ij_k}\mathbf{c}_{k,n} \quad \text{with} \quad \mathbf{c}_{k,n} \sim \mathcal{N}_c(0, h_{kn}\,\mathrm{diag}\,(\mathbf{w}_k)) \quad (60)$$

where j_k indexes the set $\mathcal{K}_j$ to which k belongs. Now if we assume that the contributions of component k to each channel is not the exact same signal $\mathbf{c}_{k,n}$ but different realizations $\{\mathbf{c}_{ik,n}\}_{i=1\ldots I}$ of the same random process $\mathcal{N}_c(0, h_{kn}\,\mathrm{diag}\,(\mathbf{w}_k))$, our model writes

$$\mathbf{x}_{i,n} = \sum_{k=1}^{K} a_{ij_k}\mathbf{c}_{ik,n} \quad \text{with} \quad \mathbf{c}_{ik,n} \sim \mathcal{N}_c(0, h_{kn}\,\mathrm{diag}\,(\mathbf{w}_k)) \quad (61)$$

and the joint likelihood is now equivalent to (59).

Optimizing the sum of likelihoods $\sum_i -\log p(\mathbf{X}_i|\boldsymbol{\theta})$ under the assumption that model (60) is true is suboptimal as compared to optimizing the joint likelihood. However, it may be something sensible to do when equation (60) fails to represent the data correctly, in particular when the point-source assumption fails. See further discussion in (Ozerov and Févotte, 2010).

4.1.3 Reconstruction of source images

Given that optimizing $C(\boldsymbol{\theta})$ is equivalent to maximizing the joint likelihood for model (61), component contributions to each channel may for example be reconstructed via MMSE estimates given by

$$\hat{\mathbf{c}}_{ik,n} = \mathrm{E}\{\mathbf{c}_{ik,n}|\mathbf{x}_i, \hat{\boldsymbol{\theta}}\} \quad (62)$$

Under our Gaussian model assumptions this leads to the following Wiener filter

$$\hat{c}_{ik,fn} = \frac{q_{ij_k} w_{fk} h_{kn}}{\hat{v}_{i,fn}} x_{i,fn}. \tag{63}$$

Note at this point that though criterion $C(\boldsymbol{\theta})$ is blind to the sign of a_{ij}, only the modulus of the mixing coefficients intervenes in this component reconstruction formula. The decomposition is conservative in the sense that it satifies

$$\mathbf{x}_{i,n} = \sum_k \hat{\mathbf{c}}_{ik,n} \tag{64}$$

Inverse-STFT of the $F \times N$ matrices $\{\hat{c}_{ik,fn}\}_{fn}$ for all i and k leads to a set of time-domain component "images" $\{\hat{\mathbf{c}}_1(t), \ldots, \hat{\mathbf{c}}_K(t)\}$, with

$$\hat{\mathbf{c}}_k(t) = \begin{bmatrix} \hat{c}_{1k}(t) \\ \vdots \\ \hat{c}_{Ik}(t) \end{bmatrix} \tag{65}$$

With equation (64) and with the linearity of the inverse-STFT, the time-domain decomposition is conservative as well, i.e,

$$\mathbf{x}(t) = \sum_k \hat{\mathbf{c}}_k(t) \tag{66}$$

where we recall that $\mathbf{x}(t)$ is the time-domain original multichannel data. Source image estimates may be obtained as

$$\hat{\mathbf{s}}_j(t) = \sum_{k \in \mathcal{K}_j} \hat{\mathbf{c}}_k(t). \tag{67}$$

Note that we are thus not estimating the point-source signal $s_j(t)$ given in equation (52) but rather yield a "spatial" multichannel image estimate $\hat{\mathbf{s}}_j(t)$ reflecting the contribution of $s_j(t)$ to the mix.

4.1.4 Relation to PARAFAC-NTF

Our model has links with the PARAFAC-NTF approach employed in (FitzGerald et al., 2005; Parry and Essa, 2006) for stereo source separation. PARAFAC refers to the parallel factor analysis tensor model which essentially consists of approximating a tensor as a sum of rank-1 tensors, see, e.g, (Bro, 1997). The approximate data structure is

$$\hat{v}^{NTF}_{i,fn} = \sum_k q^{NTF}_{ik} \, w_{fk} \, h_{kn}. \tag{68}$$

It is only a sum of $I \times F \times N$ rank-1 tensors and amounts to assuming that the variance model for channel i $\hat{\mathbf{V}}^{NTF}_i = [\hat{v}^{NTF}_{i,fn}]_{fn}$ is a linear combination of $F \times N$ time-frequency patterns $\mathbf{w}_k h_k$, where $\mathbf{w}_k$ is column k of $\hat{\mathbf{W}}$ and h_k is row k of $\mathbf{H}$. It also intrinsically implies a linear instantaneous mixture but requires a post-processing binding step in order to group the K elementary patterns into J sources, based on clustering of the ratios $\{q_{1k}/q_{2k}\}_k$ (in the stereo case). To ease comparison, our model can be rewritten as

$$\hat{v}_{i,fn} = \sum_k q_{ij_k} \, w_{fk} \, h_{kn} \tag{69}$$

where we recall that $q_{ij_k} = q_{ij}$ if and only if $k \in \mathcal{K}_j$. Hence our model has the merit of imposing that the K mixing proportions $\{q_{ik,f}\}_k$ can only take J possible values out of K, which implies that the clustering of the components is taken care of within the decomposition as opposed to after the decomposition. PARAFAC-NTF is then achieved by minimizing $D(\mathbf{V}|\hat{\mathbf{V}})$ using multiplicative updates given in (Welling and Weber, 2001; Shashua and Hazan, 2005).

4.2 Convolutive mixing

4.2.1 Generative model

The instantaneous model has allowed us to introduce the concepts of multichannel NMF based on simple mixing assumptions. However in practice, most audio recordings involve convolution, i.e,

$$x_i(t) = \sum_j \sum_{\tau=0}^{L} a_{ij}(\tau) s_j(t-\tau) \tag{70}$$

When the convolution length L is "significantly" shorter than the STFT analysis window size the time-domain convolutive mixing can be approximated by linear instantaneous mixing in each frequency band f, i.e

$$x_{i,fn} = \sum_j a_{ij,f} s_{j,fn}, \tag{71}$$

which is to be contrasted with the linear instantaneous mixture of equation (53), i.e, a_{ij} has been replaced by $a_{ij,f}$. From there, everything described in previous section holds, from source assumptions to inference, given that q_{ij} is everywhere replaced by $q_{ij,f} = |a_{ij,f}|^2$. In the following we note $\mathbf{q}_{ij,f} = [q_{ij,1}, \ldots, q_{ij,F}]^T$ and $\mathbf{Q}_f$ the $I \times J$ matrix with coefficients $\{q_{ij,f}\}_{ij}$. In the convolutive mixture, the variance model for channel i now writes

$$\hat{\mathbf{V}}_i = \sum_j \mathrm{diag}\,(\mathbf{q}_{ij})\, \mathbf{W}_j \mathbf{H}_j \tag{72}$$

and we aim at minimizing criterion (59).

4.2.2 Indeterminacies

Structure (72) suffers from obvious scale and permutation indeterminacies.[2] Regarding scale, let $\boldsymbol{\theta}_1 = \{\{\mathbf{Q}_f\}_f, \{\mathbf{W}_j, \mathbf{H}_j\}_j\}$ be a minimizer of (59) and let $\{\mathbf{D}_f\}_f$ and $\{\boldsymbol{\Lambda}_j\}_j$ be sets of nonnegative diagonal matrices. Then, the set $\boldsymbol{\theta}_2 = \{\{\mathbf{Q}_f\, \mathbf{D}_f^{-1}\}_f, \{\mathrm{diag}\,([d_{jj,f}]_f)\ \mathbf{W}_j\, \boldsymbol{\Lambda}_j^{-1}\}_j, \{\boldsymbol{\Lambda}_j\, \mathbf{H}_j\}_j\}$ leads to $v_{i,fn}(\boldsymbol{\theta}_1) = v_{i,fn}(\boldsymbol{\theta}_2)$, hence same criterion value. Similarly, permuted diagonal matrices would also leave the criterion unchanged. In practice, we remove these scale ambiguities by imposing $\sum_i q_{ij,f} = 1$ (and scaling the rows of $\mathbf{W}_j$ accordingly) and then by imposing $\sum_f w_{fk} = 1$ (and scaling the rows of $\mathbf{H}_j$ accordingly). Note again that the indeterminacy on the phase of $a_{ij,f}$ is total.

4.2.3 Multiplicative updates

Like in the single-channel case we adopt a multiplicative gradient descent approach for minimization of criterion $C(\mathbf{W}, \mathbf{H}, \mathbf{Q})$, which in particular allows to ensure the nonnegativity of all parameters. The following derivatives may be obtained

$$\nabla_{q_{ij,f}} C(\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{k \in \mathcal{K}_j} w_{fk} h_{kn}\, d'_{IS}(v_{i,fn} | \hat{v}_{i,fn}) \tag{73}$$

$$\nabla_{w_{jfk}} C(\boldsymbol{\theta}) = \sum_{i=1}^{I} \sum_{n=1}^{N} q_{ij,f}\, h_{j,kn}\, d'_{IS}(v_{i,fn} | \hat{v}_{i,fn}) \tag{74}$$

$$\nabla_{h_{jkn}} C(\boldsymbol{\theta}) = \sum_{i=1}^{I} \sum_{f=1}^{F} q_{ij,f}\, w_{j,fk}\, d'_{IS}(v_{i,fn} | \hat{v}_{i,fn}) \tag{75}$$

$$\tag{76}$$

where $d'_{IS}(x|y) = 1/y - x/y^2$. Separating the positive and negative summands of each equations and rearranging the expressions in a matrix form leads to Algorithm 5, whose convergence was

[2] There might also be other less obvious indeterminacies, such as those inherent to NMF (see, e.g., Laurberg et al. (2008)), but this study is here left aside.

Algorithm 5 Multichannel IS-NMF

Input : nonnegative tensor $\mathbf{V}$
Output : nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$, nonnegative matrix or tensor $\mathbf{Q}$
Initialize $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{Q}$ with nonnegative values
for $l = 1 : n_{iter}$ **do**
$\mathbf{q}_{ij} \leftarrow \mathbf{q}_{ij}.\frac{\left[\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i.(\mathbf{W}_j\mathbf{H}_j)\right]\mathbf{1}_{N\times 1}}{\left[\hat{\mathbf{V}}_i^{-1}.(\mathbf{W}_j\mathbf{H}_j)\right]\mathbf{1}_{N\times 1}}$
$\mathbf{W}_j \leftarrow \mathbf{W}_j.\frac{\sum_{i=1}^{I}\mathrm{diag}(\mathbf{q}_{ij})(\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i)\mathbf{H}_j^T}{\sum_{i=1}^{I}\mathrm{diag}(\mathbf{q}_{ij})\hat{\mathbf{V}}_i^{-1}\mathbf{H}_j^T}$
$\mathbf{H}_j \leftarrow \mathbf{H}_j.\frac{\sum_{i=1}^{I}(\mathrm{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T(\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i)}{\sum_{i=1}^{I}(\mathrm{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T\hat{\mathbf{V}}_i^{-1}}$
Normalize $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$
end for

observed in practice. Updates for the linear instantaneous mixing are obtained as a special case, by setting $q_{ij,f} = q_{ij}$ everywhere and using the update

$$q_{ij} \leftarrow q_{ij}.\frac{\mathrm{sum}\left[\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i.(\mathbf{W}_j\mathbf{H}_j)\right]}{\mathrm{sum}\left[\hat{\mathbf{V}}_i^{-1}.(\mathbf{W}_j\mathbf{H}_j)\right]} \tag{77}$$

where sum[$\mathbf{M}$] is the sum of all coefficients in $\mathbf{M}$. In the specific linear instantaneous case, multiplicative updates based on Khatri-Rao and contracted tensor products can be exhibited for the whole matrices $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ (instead of individual updates for $q_{ij}, \mathbf{W}_j, \mathbf{H}_j$), but are not given here for conciseness. They are similar in form to (Welling and Weber, 2001) and (Shashua and Hazan, 2005) and lead to a faster MATLAB implementation.

4.3 Results

We present decomposition results of a real stereo music excerpt taken as the first 24 s of the song *So Much Trouble in the World* by Bob Marley. The excerpt is composed two guitars, drums & percussions, bass, voice and synthesizer sounds. One guitar plays the off-beat rhythmic component characteristic of reggae and we will refer to it as lead guitar, while we will refer to the other as second guitar. The audio signal is downsampled at 22.05 kHz and a 50 % overlap sinebell STFT was computed with window size 92 ms - a rather long window size is here preferable for model (71) to hold. We decomposed the signal with the convolutive NTF method described in previous section, with $J = 4$ sources and only 2 components per source, i.e, $K = 8$. It is clear that such a small number of components is insufficient to capture every elementary object of this rich and diverse audio scene, but we wish to inspect what can our convolutive IS-NTF model learn with such a restricted number of components. We run the algorithm for 1000 iterations (approximatively 30 min with our MATLAB implementation on a Mac 2.6 GHz with 2 Go RAM) and as before we select the solution with minimum cost value from 10 random initializations. The learnt dictionary $\mathbf{W}$, coefficients $\mathbf{H}$ and reconstructed stereo components are displayed on figure 8. The sound samples can be listened to online at `http://www.tsi.enst.fr/~fevotte/Samples/machine-audition/`. The MATLAB code of the algorithm is available from the author webpage as well.

Very interestingly the decomposition captures high-level structures from the data in the sense that each of the components captures pieces of a subset of the instrumental sources. The decomposition describes coarsely as follows. The first component contains mainly all of the bass, the bass drum, one part of the second guitar, one sort of the synthesizer sounds. The second and fourth component contains mainly parts of the voice. The fifth component contains mainly the lead guitar and the other part of the second guitar. Finally, the remaining components each capture one of the drum sources. Hence, binding components 1 and 5 mainly provide the bass, guitars and synthesizer part, components 2 and 4 mainly provide the voice and the four remaining components mainly provide the drums. Note the "instrumental sources" resulting from the manual binding of the components belong to different "directional sources" as intended in model (71), showing the limits of the point-source model. The separation between all the instruments can be further improved

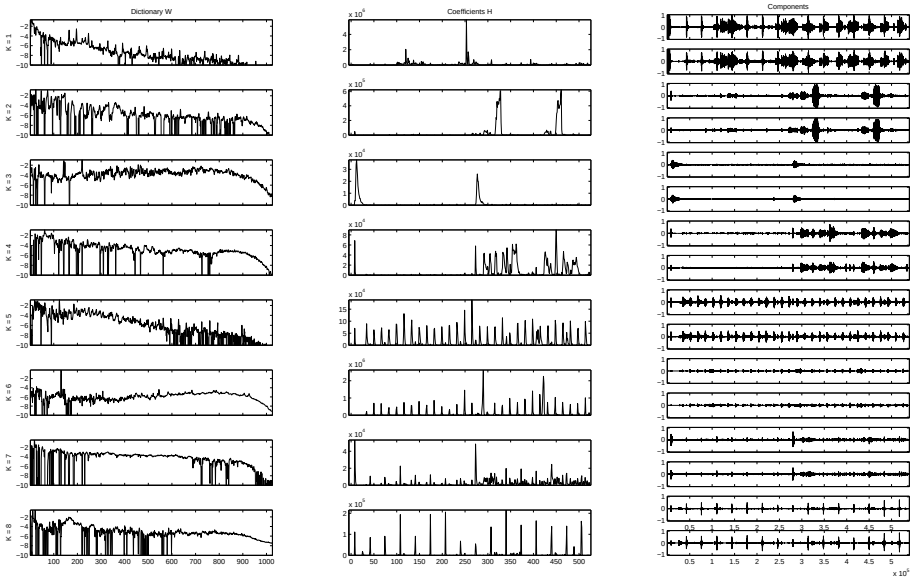

Figure 8: Convolutive IS-NTF with $K = 8$. Left : columns of $\mathbf{W}$ ($\log_{10}$ scale). Middle : rows of $\mathbf{H}$. Right : Component reconstructions with equation (63).

by allotting more components to each directional source j. We present online separation results obtained with $J = 4$ and 6 component per source, i.e, a total number of $K = 24$ components, that were manually bound into 6 instrumental sources containing respectively the lead guitar, the second guitar, the voice, the bass, the drums and the synthesizer.

5 Conclusions

In this chapter we have attempted to show the relevance of using nonnegative factorization of the power spectrogram with the Itakura-Saito divergence, for musical audio decomposition. On the modelling side, we wish to bring out the following three features of IS-NMF : 1) IS-NMF is underlain by a sound statistical model of superimposed Gaussian components, 2) this model is relevant to the representation of audio signals, 3) this model can accommodate regularization constraints and model selection scenarios through Bayesian approaches. We compared the decompositions of a well structured piano sequence obtained using IS-NMF of the power spectrogram and using the more common approach of KL-NMF of the magnitude spectrogram. The organization of the decomposition with the former better matches our own comprehension of sound.

In the second part of the paper we described extensions of NMF-based sound decomposition to the multichannel case. NMF-based models usually convey the idea that the dictionary elements should represent low-level elementary objects such as notes. However, experiments on a real stereo musical excerpt showed that, given a small number of components, the model is able to retrieve both low-level objects and higher-level structures encompassing rich and diverse sources.

On the algorithmic part, one of the limitations of the decomposition techniques we described is their sensitivity to local minima, and we found out in practice that the decompositions obtained from the local solutions were semantically not as satisfactory as the "more optimal" ones - this observation also corroborates the relevance of the IS divergence for audio, in the sense that lower cost solutions are better indeed. In order to reduce this problem we systematically ran the algorithms several times from different random solutions, which prove rather satisfactory. However, more advanced MCMC inference strategies have also been mentioned, though we do not give any results at this stage.

The computation times involved by the multiplicative algorithms presented in this paper is fair; though still far from real-time (in the order of 1 min for 1 s of data, to decompose stereo signals sampled at 22 kHz, using a MATLAB implementation on a standard machine) the methods may be used in an off-line setting, typically for sound edition.

Finally, as also discussed in the chapter, the plain IS-NMF model can be sophisticated in many ways, for example to model voice or include explicit pitched structure. The general inference methodology presented in this chapter holds in every case and combinations of various component models are possible, either in the single or multichannel cases.

A Standard distributions

Proper complex Gaussian $\mathcal{N}_c\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right) = |\pi\,\boldsymbol{\Sigma}|^{-1}\,\exp -(\mathbf{x}-\boldsymbol{\mu})^H\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})$

Poisson $\mathcal{P}(x|\lambda) = \exp(-\lambda)\,\frac{\lambda^x}{x!}$

Gamma $\mathcal{G}(u|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,u^{\alpha-1}\,\exp(-\beta\,u),\; u \geq 0$

inverse-Gamma $\mathcal{IG}(u|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,u^{-(\alpha+1)}\,\exp(-\frac{\beta}{u}),\; u \geq 0$

(The inverse-Gamma distribution is the distribution of $1/X$ when X is Gamma distributed.)

half-normal $\mathcal{HN}(x|\phi) = \left(\frac{\pi\phi}{2}\right)^{-\frac{1}{2}}\,\exp\left(-\frac{1}{2}\frac{x^2}{\phi}\right)$

(The half-normal distribution is the distribution of $|X|$ when X is normally distributed with mean 0.)

Acknowledgements

I would like to acknowledge Nancy Bertin, Raphaël Blouet, Taylan Cemgil, Maurice Charbit and Jean-Louis Durrieu for collaborations related to the content of this chaper, with very special thanks to Alexey Ozerov and Vincent Y. F. Tan.

References

S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by nonnegative sparse coding of power spectra. In *Proc. 5th International Symposium Music Information Retrieval (ISMIR'04)*, pages 318–325, Barcelona, Spain, Oct. 2004.

L. Benaroya, R. Gribonval, and F. Bimbot. Non negative sparse representation for Wiener based source separation with a single sensor. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, pages 613–616, Hong Kong, 2003.

L. Benaroya, R. Blouet, C Févotte, and I. Cohen. Single sensor source separation using multiple-window STFT representation. In *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sep. 2006a.

L. Benaroya, R. Gribonval, and F. Bimbot. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):191–199, Jan. 2006b.

C. M. Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems (NIPS)*, pages 382–388, 1999.

R. Blouet, G. Rapaport, I. Cohen, and C. Févotte. Evaluation of several strategies for single sensor speech/music separation. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Las Vegas, USA, Apr. 2008.

R. Bro. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, Oct. 1997.

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.

S. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1995, Dec. 1995.

A. Cichocki, S.-I. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He. Extended SMART algorithms for non-negative matrix factorization. In *Proc. International Conference on Artificial Intelligence and Soft Computing (ICAISC'06)*, pages 548–562, Zakopane, Poland, June 2006.

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

A. Doucet, S. Sénécal, and T. Matsui. Space alternating data augmentation: Application to finite mixture of gaussians and speaker recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, pages IV–713 – IV–716, Philadelphia, 2005.

J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David. Main instrument separation from stereophonic audio signals using a source/filter model. In *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*, pages 15–19, Glasgow, Scotland, Aug. 2009.

J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010. Submitted.

C. Févotte and A. T. Cemgil. Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*, pages 1913–1917, Glasgow, Scotland, Aug. 2009.

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), Mar. 2009.

D. FitzGerald, M. Cranitch, and E. Coyle. Non-negative tensor factorisation for sound source separation. In *Proc. of the Irish Signals and Systems Conference*, Dublin, Ireland, Sep. 2005.

R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):367–376, Aug. 1980.

F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proc 6^{th} International Congress on Acoustics*, pages C–17 – C–20, Tokyo, Japan, Aug. 1968.

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

K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15(2):349–396, 2003.

H. Laurberg, M. Græbøll 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.

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.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

D. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. `http://www.inference.phy.cam.ac.uk/mackay/ica.pdf`, 1996. Unpublished.

D. J. C. Mackay. Probable networks and plausible predictions – a review of practical Bayesian models for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret. Separation of non-negative mixture of non-negative sources using a Bayesian approach and mcmc sampling. *IEEE Trans. on Signal Processing*, 54(11):4133–4145, Nov. 2006.

A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 2010. In press.

A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1564–1578, Jul. 2007.

A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden markov model for polyphonic audio representation and source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, Mohonk, NY, USA, Oct. 2009.

P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37(1):23–25, May 1997.

R. M. Parry and I. Essa. Phase-aware non-negative spectrogram factorization. In *In Proc. International Conference on Independent Component Analysis and Signal Separation (ICA'07)*, pages 536–543, London, UK, Sep. 2007.

R. M. Parry and I. A. Essa. Estimating the spatial position of spectral components in audio. In *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, pages 666–673, Charleston SC, USA, Mar. 2006.

M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *In Proc. 8th Internation conference on Independent Component Analysis and Signal Separation (ICA'09)*, Paraty, Brazil, Mar. 2009.

M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008(Article ID 947438):8 pages, 2008. doi:10.1155/2008/947438.

A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proc. 22nd International Conference on Machine learning*, pages 792 – 799, Bonn, Germany, 2005. ACM.

P. Smaragdis. Convolutive speech bases and their application to speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1–12, Jan. 2007.

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. Stoica and Y. Selén. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, Jul. 2004.

V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization. In *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS'09)*, Saint-Malo, France, Apr. 2009.

M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211 – 244, Sep. 2001.

E. Vincent, N. Bertin, and R. Badeau. Two nonnegative matrix factorization methods for polyphonic pitch transcription. In *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.

T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, Mar. 2007.

T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 1825–1828, Las Vegas, Nevada, USA, Apr. 2008.

W. Wang, A. Cichocki, and J. A. Chambers. A multiplicative algorithm for convolutive non-negative matrix factorization based on squared euclidean distance. *IEEE Trans. on Signal Processing*, 57(7):2858–2864, July 2009. doi: 10.1109/TSP.2009.2016881.

M. Welling and M. Weber. Positive tensor factorization. *Pattern Recognition Letters*, 22(12): 1255–1261, 2001.