Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis

Cédric Févotte
fevotte@telecom-paristech.fr
Nancy Bertin
nbertin@telecom-paristech.fr
Jean-Louis Durrieu
durrieu@telecom-paristech.fr
CNRS—TELECOM ParisTech, 75014 Paris, France

This letter presents theoretical, algorithmic, and experimental results about nonnegative matrix factorization (NMF) with the Itakura-Saito (IS) divergence. We describe how IS-NMF is underlaid by a well-defined statistical model of superimposed gaussian components and is equivalent to maximum likelihood estimation of variance parameters. This setting can accommodate regularization constraints on the factors through Bayesian priors. In particular, inverse-gamma and gamma Markov chain priors are considered in this work. Estimation can be carried out using a space-alternating generalized expectation-maximization (SAGE) algorithm; this leads to a novel type of NMF algorithm, whose convergence to a stationary point of the IS cost function is guaranteed.

We also discuss the links between the IS divergence and other cost functions used in NMF, in particular, the Euclidean distance and the generalized Kullback-Leibler (KL) divergence. As such, we describe how IS-NMF can also be performed using a gradient multiplicative algorithm (a standard algorithm structure in NMF) whose convergence is observed in practice, though not proven.

Finally, we report a furnished experimental comparative study of Euclidean-NMF, KL-NMF, and IS-NMF algorithms applied to the power spectrogram of a short piano sequence recorded in real conditions, with various initializations and model orders. Then we show how IS-NMF can successfully be employed for denoising and upmix (mono to stereo conversion) of an original piece of early jazz music. These experiments indicate that IS-NMF correctly captures the semantics of audio and is better suited to the representation of music signals than NMF with the usual Euclidean and KL costs.

1 Introduction

Nonnegative matrix factorization (NMF) is a popular dimension-reduction technique, employed for nonsubtractive, 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

$$\mathbf{V} \approx \mathbf{W}\mathbf{H}, \tag{1.1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $F\,K + K\,N \ll F\,N$, hence reducing the data dimension. Note that the factorization is in general only approximate, so that the terms *approximate nonnegative matrix factorization* and *nonnegative matrix approximation* also appear in the literature. NMF has been used for various problems in diverse fields. To cite a few, we mention the problems of learning parts of faces and semantic features of text (Lee & Seung, 1999), polyphonic music transcription (Smaragdis & Brown, 2003), object characterization by reflectance spectra analysis (Berry, Browne, Langville, Pauca, & Plemmons, 2007), portfolio diversification (Drakakis, Rickard, de Fréin, & Cichocki, 2008), and scotch whiskies clustering (Young, Fogel, & Hawkins, 2006).

In the literature, the factorization, equation 1.1, is usually sought after through the minimization problem

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D(\mathbf{V} \mid \mathbf{W}\mathbf{H}), \tag{1.2}$$

where $D(\mathbf{V} \mid \mathbf{W}\mathbf{H})$ is a cost function defined by

$$D(\mathbf{V} \mid \mathbf{W}\mathbf{H}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d([\mathbf{V}]_{fn} \mid [\mathbf{W}\mathbf{H}]_{fn}), \tag{1.3}$$

and where $d(x \mid y)$ is a scalar cost function. Popular choices are the Euclidean distance, which we here define as

$$d_{EUC}(x \mid y) = \frac{1}{2}(x - y)^2, \tag{1.4}$$

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

$$d_{KL}(x \mid y) = x \, \log \frac{x}{y} - x + y. \tag{1.5}$$

Both cost functions are positive and take value zero if and only if $x = y$.

Lee and Seung (2001) proposed gradient descent algorithms to solve the minimization problem, equation 1.2, under the latter two cost functions. When a suitable step size is used, the gradient descent update rules are turned into multiplicative rules, under which the cost function is shown to be nonincreasing. The simplicity of the update rules has undoubtedly contributed to the popularity of NMF, and most of the above-mentioned applications are based on Lee and Seung's algorithm for minimization of either the Euclidean distance or the KL divergence.

Nevertheless, some papers have considered NMF under other cost functions and other algorithmic structures. In particular Cichocki and coauthors have devised several types of NMF algorithms for cost functions such as Csiszár divergences (including Amari's α-divergence) and the β-divergence in Cichocki, Zdunek, and Amari (2006), with several other cost functions considered in Cichocki, Amari et al. (2006). Also, Dhillon and Sra (2005) have described multiplicative algorithms for the wide family of Bregman divergences. The choice of the NMF cost function should be driven by the type of data to analyze, 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 letter, we are specifically interested in NMF with the Itakura-Saito (IS) divergence, and we demonstrate its relevance to the decomposition of audio spectra. The expression of the IS divergence is given by

$$d_{IS}(x \mid y) = \frac{x}{y} - \log \frac{x}{y} - 1. \tag{1.6}$$

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 1970s. It was in particular praised for the good perceptual properties of the reconstructed signals it led to (Gray, Buzo, Gray, & Matsuyama, 1980).

As we shall see, this divergence has other 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, we describe how NMF in this case can be recast as ML estimation of $\mathbf{W}$ and $\mathbf{H}$ in superimposed signals under simple gaussian assumptions. Equivalently, we describe how IS-NMF can be interpreted as ML of $\mathbf{W}$ and $\mathbf{H}$ in multiplicative gamma noise.

The IS divergence belongs to the class of Bregman divergences and is a limit case of the β-divergence. Thus, the gradient descent multiplicative rules given in Dhillon and Sra (2005) and Cichocki, Zdunek et al. (2006), which coincide in the IS case, can be applied. If convergence of this algorithm is observed in practice, its proof is still an open problem. The statistical framework going along with IS-NMF allows deriving a new type of minimization method, derived from space-alternating expectation-maximization (SAGE), a variant of the standard expectation-maximization (EM) algorithm. This method leads to new update rules, which do not possess a multiplicative structure. The EM setting guarantees convergence of this algorithm to a stationary point of the cost function. Moreover, the statistical framework opens doors to Bayesian approaches for NMF, allowing elaborate priors on **W** and **H**, for which maximum a posteriori (MAP) estimation can again be performed using SAGE. Examples of such priors, yielding regularized estimates of **W** and **H**, are presented in this work.

IS-NMF underlies previous work in the area of automatic music transcription and single-channel audio source separation, but never explicitly so. In particular, our work builds on Benaroya, Gribonval, and Bimbot (2003), Benaroya, Blouet, Févotte, and Cohen (2006), Abdallah and Plumbley (2004), and Plumbley, Abdallah, Blumensath, and Davies (2006), and the connections between IS-NMF and these articles will be discussed.

This letter is organized as follows. Section 2 addresses general properties of IS-NMF. The relation between the IS divergence and other cost functions used in NMF is discussed in section 2.1, section 2.2 addresses scale invariance, and section 2.3 describes the statistical interpretations of IS-NMF. Section 3 presents two IS-NMF algorithms; an existing multiplicative algorithm is described in section 3.1, and section 3.2 introduces a new algorithm derived from SAGE. Section 4 reports an experimental comparative study of Euclidean-NMF, KL-NMF, or IS-NMF algorithms applied to the power spectrogram of a short piano sequence recorded in real conditions, with various initializations and model orders. These experiments show that IS-NMF correctly captures the semantics of the signal and is better suited to the representation of audio than NMF with the usual Euclidean and KL costs. Section 5 presents how IS-NMF can accommodate regularization constraints on **W** and **H** within a Bayesian framework and how SAGE can be adapted to MAP estimation. In particular, we give update rules for IS-NMF with gamma and inverse-gamma Markov chain priors on the rows of **H**. In section 6, we present audio restoration results of an original early recording of jazz music; we show how the proposed regularized IS-NMF algorithms can successfully be employed for denoising and upmix (mono to stereo conversion) of the original data. Finally, conclusions and perspectives of this work are given in section 7.

2 Properties of NMF with the Itakura-Saito Divergence

In this section we address the links between the IS divergence and other cost functions used for NMF. Then we discuss its scale invariance property and, finally, describe the statistical interpretations of IS-NMF.

2.1 Relation to Other Divergences Used in NMF.

2.1.1 β-Divergence. As observed by Cichocki, Amari et al. (2006) and Cichocki, Zdunek et al. (2006), the IS divergence is a limit case of the β-divergence introduced by Eguchi and Kano (2001) that we here define as

$$d_\beta(x \mid y) \overset{\text{def}}{=} \begin{cases} \dfrac{1}{\beta\,(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\,\log x/y + (y-x) & \beta = 1 \\ \dfrac{x}{y} - \log\dfrac{x}{y} - 1 & \beta = 0. \end{cases} \tag{2.1}$$

Eguchi and Kano (2001) assume $\beta > 1$, but the definition domain can very well be extended to $\beta \in \mathbb{R}$. The β-divergence is shown to be continuous in β by using the identity $\lim_{\beta\to 0}(x^\beta - y^\beta)/\beta = \log(x/y)$. It was considered in NMF by Cichocki, Zdunek et al. (2006) and also coincides up to a factor $1/\beta$ with the generalized divergence of Kompass (2007), which, in the context of NMF as well, was separately constructed so as to interpolate between the KL divergence ($\beta = 1$) and the Euclidean distance ($\beta = 2$). Note that the derivative of $d_\beta(x \mid y)$ with regard to y is also continuous in β and is simply written as

$$\nabla_y\, d_\beta(x \mid y) = y^{\beta-2}\,(y-x). \tag{2.2}$$

The derivative shows that $d_\beta(x|y)$, as a function of y, has a single minimum in $y = x$ and that it increases with $|y-x|$, justifying its relevance as a measure of fit. Figure 1 represents the Euclidean, KL, and IS costs for $x = 1$.

When equation 2.2 is used, the gradients of criterion $D_\beta(\mathbf{V} \mid \mathbf{WH})$ with regard to $\mathbf{W}$ and $\mathbf{H}$ are written as

$$\nabla_{\mathbf{H}}\, D_\beta(\mathbf{V} \mid \mathbf{WH}) = \mathbf{W}^T\left((\mathbf{WH})^{\cdot[\beta-2]} \,.\, (\mathbf{WH}-\mathbf{V})\right) \tag{2.3}$$

$$\nabla_{\mathbf{W}}\, D_\beta(\mathbf{V} \mid \mathbf{WH}) = \left((\mathbf{WH})^{\cdot[\beta-2]} \,.\, (\mathbf{WH}-\mathbf{V})\right)\mathbf{H}^T, \tag{2.4}$$

where . denotes Hadamard entrywise product and $\mathbf{A}^{\cdot[n]}$ denotes the matrix with entries $[\mathbf{A}]_{ij}^n$. The multiplicative gradient descent approach taken in Lee and Seung (2001) and Cichocki, Zdunek et al. (2006) is equivalent to updating each parameter by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion with regard 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 is with a nonnegative value. A fixed point $\theta^\star$ of the algorithm implies

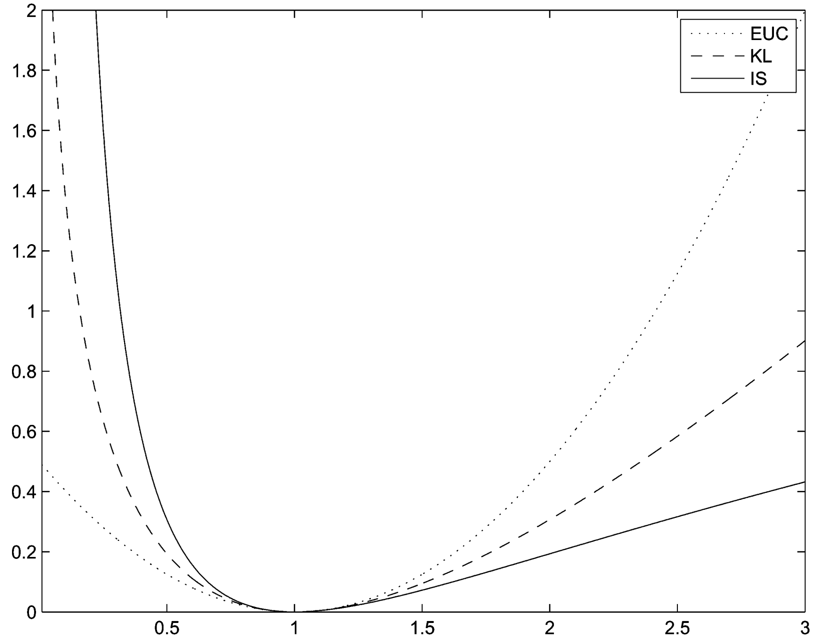

Figure 1: Euclidean, KL, and IS costs $d(x \mid y)$ as a function of y and for $x = 1$. The Euclidean and KL divergences are convex on $(0, \infty)$. The IS divergence is convex on $(0, 2x]$ and concave on $[2x, \infty)$.

either $\nabla f(\theta^\star) = 0$ or $\theta^\star = 0$. This leads to the following updates,

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T \left((\mathbf{WH})^{\cdot[\beta-2]}.\mathbf{V}\right)}{\mathbf{W}^T \, (\mathbf{WH})^{\cdot[\beta-1]}} \tag{2.5}$$

$$\mathbf{W} \leftarrow \mathbf{W}.\frac{\left((\mathbf{WH})^{\cdot[\beta-2]}.\mathbf{V}\right) \mathbf{H}^T}{(\mathbf{WH})^{\cdot[\beta-1]} \, \mathbf{H}^T}, \tag{2.6}$$

where $\frac{\mathbf{A}}{\mathbf{B}}$ denotes the matrix $\mathbf{A}.\mathbf{B}^{\cdot[-1]}$. Lee and Seung (1999) showed that $D_\beta(\mathbf{V} \mid \mathbf{W}\,\mathbf{H})$ is nonincreasing under the latter updates for $\beta = 2$ (Euclidean distance) and $\beta = 1$ (KL divergence). Kompass (2007) generalizes the proof to the case $1 \leq \beta \leq 2$. In practice, we observe that the criterion is still nonincreasing under updates 2.5 and 2.6 for $\beta < 1$ and $\beta > 2$ (and in particular for $\beta = 0$, corresponding to the IS divergence), but no proof is available. Indeed, the proof Kompass gives makes use of the convexity of $d_\beta(x|y)$ as a function of y, which is true only for $1 \leq \beta \leq 2$. In the rest of the letter, *EUC-NMF* will be used as shorthand for *Euclidean-NMF*.

2.1.2 Bregman Divergences. The IS divergence belongs to the class of Bregman divergences, defined as $d_\phi(x|y) = \phi(x) - \phi(y) - \nabla\phi(y)\,(x - y)$, where ϕ is a strictly convex function of $\mathbb{R}$ that has a continuous derivative $\nabla\phi$. The IS divergence is obtained with $\phi(y) = -\log(y)$. Using the same approach as in the previous paragraph, Dhillon and Sra (2005) derive the following update rules for minimization of $D_\phi(\mathbf{V} \mid \mathbf{WH})$:

$$\mathbf{H} \leftarrow \mathbf{H}.\, \frac{\mathbf{W}^T\,(\nabla^2\phi(\mathbf{WH}).\mathbf{V})}{\mathbf{W}^T\,(\nabla^2\phi(\mathbf{WH}).\mathbf{WH})} \tag{2.7}$$

$$\mathbf{W} \leftarrow \mathbf{W}.\, \frac{(\nabla^2\phi(\mathbf{WH}).\mathbf{V})\,\mathbf{H}^T}{(\nabla^2\phi(\mathbf{WH}).\mathbf{WH})\,\mathbf{H}^T}. \tag{2.8}$$

Again, the authors observed in practice continual descent of $D_\phi(\mathbf{V} \mid \mathbf{WH})$ under these rules, but a proof of convergence is yet to be found. Note that equations 2.5 and 2.6 coincide with equations 2.7 and 2.8 for the IS divergence.

2.2 Scale Invariance. The following property holds for any value of β:

$$d_\beta(\gamma\, x \mid \gamma\, y) = \gamma^\beta\, d_\beta(x \mid y). \tag{2.9}$$

It implies that the IS divergence is scale invariant (i.e., $d_{IS}(\gamma x \mid \gamma y) = d_{IS}(x \mid y)$) and is the only one of the β-divergence family to possess this property. *Scale invariance* means that same relative weight is given to small and large coefficients of $\mathbf{V}$ in cost function (see equation 1.3) in the sense that a bad fit of the factorization for a low-power coefficient $[\mathbf{V}]_{fn}$ will cost as much as a bad fit for a higher-power coefficient $[\mathbf{V}]_{f'n'}$. In contrast, factorizations obtained with $\beta > 0$ (such as 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 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. The results of the decomposition of a piano spectrogram presented in section 4 confirm these expectations by showing that IS-NMF extracts components corresponding to very low residual noise and hammer hits on the strings with great accuracy. These components are either ignored or severely degraded when using Euclidean or KL divergences.

2.3 Statistical Interpretations. We now turn to statistical interpretations of IS-NMF, which lead to the new EM-based algorithm described in section 3.

2.3.1 Notations. The entries of matrices $\mathbf{V}, \mathbf{W}$, and $\mathbf{H}$ are denoted v_{fn}, w_{fk}, and h_{kn}, respectively. Lowercase bold letters in general denote columns, such that $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, while lowercase 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.

2.3.2 Sum of Gaussian Components.

Theorem 1 *(IS-NMF as ML estimation in sum of gaussian components). Consider the generative model defined by,* $\forall n = 1, \ldots, N$,

$$\boldsymbol{x}_n = \overset{K}{\underset{k=1}{}} \boldsymbol{c}_{k,n}, \tag{2.10}$$

where $\boldsymbol{x}_n$ *and* $\boldsymbol{c}_{k,n}$ *belong to* $\mathbb{C}^{F \times 1}$ *and*

$$\boldsymbol{c}_{k,n} \sim \mathcal{N}_c(0, h_{kn}\, diag\,(\boldsymbol{w}_k)), \tag{2.11}$$

where $\mathcal{N}_c(\boldsymbol{\mu}, \quad)$ *denotes the proper multivariate complex gaussian distribution and where the components* $\boldsymbol{c}_{1,n}, \ldots, \boldsymbol{c}_{K,n}$ *are mutually independent and individually independently distributed. Define* $\mathbf{V}$ *as the matrix with entries* $v_{fn} = |x_{fn}|^2$*. Then, maximum likelihood estimation of* $\mathbf{W}$ *and* $\mathbf{H}$ *from* $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ *is equivalent to NMF of* $\mathbf{V}$ *into* $\mathbf{V} \approx \mathbf{WH}$*, where the Itakura-Saito divergence is used.*

Proof. Under the assumptions of theorem 1 and using the expression of $\mathcal{N}_c(\boldsymbol{\mu}, \quad)$ given in appendix A, the minus log-likelihood function $C_{ML,1}(\mathbf{W}, \mathbf{H}) \overset{\text{def}}{=} -\log p(\mathbf{X} \mid \mathbf{W}, \mathbf{H})$ simply factorizes as

$$C_{ML,1}(\mathbf{W}, \mathbf{H}) = -\sum_{n=1}^{N} \sum_{f=1}^{F} \log \mathcal{N}_c \left(x_{fn} \mid 0, \sum_k w_{fk}\, h_{kn} \right) \tag{2.12}$$

$$= NF \;\log \pi + \sum_{n=1}^{N} \sum_{f=1}^{F} \log \left(\sum_k w_{fk}\, h_{kn} \right) + \frac{|x_{fn}|^2}{(\sum_k w_{fk}\, h_{kn})} \tag{2.13}$$

$$\overset{c}{=} \sum_{n=1}^{N} \sum_{f=1}^{F} d_{IS} \left(|x_{fn}|^2 \mid \sum_k w_{fk}\, h_{kn} \right), \tag{2.14}$$

where $\overset{c}{=}$ denotes equality up to constant terms. The minimization of $C_{ML,1}(\mathbf{W}, \mathbf{H})$ with regard to $\mathbf{W}$ and $\mathbf{H}$ thus amounts to the NMF $\mathbf{V} \approx \mathbf{WH}$ with the IS divergence. Note that theorem 1 holds also for real-valued gaussian components. In that case, $C_{ML,1}(\mathbf{W}, \mathbf{H})$ equals $D_{IS}(\mathbf{V} \mid \mathbf{WH})$ up to a constant and a factor 1/2.

The generative model, equation 2.10, was introduced by Benaroya et al. (2003, 2006) for single-channel audio source separation. In that context, $\mathbf{x}_n = [x_{1n}, \ldots, x_{fn}, \ldots, x_{Fn}]^T$ is the short-time Fourier transform (STFT) of an audio signal x, where $n = 1, \ldots, N$ is a frame index and $f = 1, \ldots, F$ is a frequency index. The signal x is assumed to be the sum of two sources, $x = s_1 + s_2$, and the STFTs of the sources are modeled as $\mathbf{s}_{1,n} = \sum_{k=1}^{K_1} \mathbf{c}_{k,n}$ and $\mathbf{s}_{2,n} = \sum_{k=K_1+1}^{K_1+K_2} \mathbf{c}_{k,n}$, with $K_1 + K_2 = K$. This means that each source STFT is modeled as a sum of elementary components, each characterized by a power spectral density (PSD) $\mathbf{w}_k$ modulated in time by frame-dependent activation coefficients h_{kn}. The PSDs characterizing each source are learned on training data, before the mixture spectrogram $|\mathbf{X}|^{\cdot[2]}$ is decomposed onto the known dictionary $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{K_1}, \mathbf{w}_{K_1+1}, \ldots, \mathbf{w}_{K_1+K_2}]$. However, in these articles, the PSDs and the activation coefficients are estimated separately using somewhat ad hoc strategies (the PSDs are learned with vector quantization) and the equivalence between ML estimation and IS-NMF is not fully exploited.

Complex gaussian modeling of STFT frames of audio signals has been widely used in signal processing and has proven to be a satisfying model for many applications, in particular for audio denoising (see, e.g., Cohen & Gannot, 2007, for a review). But while denoising settings typically assume one observation frame $\mathbf{x}_n$ to be the sum of a source frame and a noise frame, IS-NMF in essence extends this modeling by assuming that one observation frame is the sum of several gaussian frames with different covariances.

The generative model, equation 2.10, may also be viewed as a generalization of well-known models of composite signals. For example, inference in superimposed components with gaussian structure can be tracked back to Feder and Weinstein (1988). In the latter article, however, the components are assumed stationary and solely modeled by their PSD $\mathbf{w}_k$, which in turn is parameterized by a set of parameters of interest $_k$, to be estimated. One extension brought in equation 2.10 is the addition of the amplitude parameters $\mathbf{H}$. This, however, has the inconvenience of making the total number of parameters $F\ K + K\ N$ dependent on N, with the consequence of losing the asymptotical optimality properties of ML estimation. But note that it is precisely the addition of the amplitude parameters in the model that allows $\mathbf{W}$ to be treated as a set of possibly identifiable parameters. Indeed, if h_{kn} is set to 1 for all k and n, the variance of $\mathbf{x}_n$ becomes $\sum_k \mathbf{w}_k$ for all n (i.e., is equal to the sum of the parameters). This would obviously make each PSD $\mathbf{w}_k$ not uniquely identifiable.

Interestingly, the equivalence between IS-NMF and ML inference in the sum of gaussian components provides means of reconstructing the components $\mathbf{c}_{k,n}$ with a sense of statistical optimality, which contrasts with NMF using other costs where methods of reconstructing components from the factorization $\mathbf{W}\mathbf{H}$ are somewhat ad hoc (see below). Indeed, given $\mathbf{W}$ and $\mathbf{H}$, minimum mean square error (MMSE) estimates can be obtained through Wiener filtering, such that

$$\hat{c}_{k,fn} = \frac{w_{fk}\, h_{kn}}{\sum_{l=1}^{K} w_{fl}\, h_{ln}} x_{fn}. \tag{2.15}$$

Because the Wiener gains sum up to 1 for a fixed entry (f, n), the decomposition is conservative:

$$\mathbf{x}_n = \sum_{k=1}^{K} \hat{\mathbf{c}}_{k,n}. \tag{2.16}$$

Note that a consequence of Wiener reconstruction is that the phase of all components $\hat{c}_{k,fn}$ is equal to the phase of x_{fn}.

Most works in audio have considered the NMF of magnitude spectra $|\mathbf{X}|$ instead of power spectra $|\mathbf{X}|^{\cdot[2]}$ (see, e.g., Smaragdis & Brown, 2003; Smaragdis, 2007; Virtanen, 2007; Bertin, Badeau, & Richard, 2007). In that case, it can be noted (see, e.g., Virtanen, Cemgil, & Godsill, 2008) that KL-NMF is related to the ML problem of estimating $\mathbf{W}$ and $\mathbf{H}$ in the model structure

$$|\mathbf{x}_n| = \sum_{k=1}^{K} |\mathbf{c}_{k,n}| \tag{2.17}$$

under Poissonian assumptions, that is, $|c_{k,fn}| \sim \mathcal{P}(w_{fk}h_{kn})$, where $\mathcal{P}(\lambda)$ is the Poisson distribution, defined in appendix A. Indeed, the sum of Poisson random variables being Poissonian itself (with the shape parameters summing up as well), one obtains $|x_{fn}| \sim \mathcal{P}(\sum_{k=1}^{K} w_{fk}h_{kn})$. Then it can easily be seen that the likelihood $-\log p(\mathbf{X} \mid \mathbf{W}, \mathbf{H})$ is equal up to a constant to $D_{KL}(\,|\mathbf{X}| \mid \mathbf{W}\,\mathbf{H})$. Here, $\mathbf{W}$ is homogeneous to a magnitude spectrum and not to a power spectrum. After factorization, component estimates are typically formed using the phase of the observations (Virtanen, 2007) such that

$$\hat{c}_{k,fn} = w_{fk}\, h_{kn} \arg(x_{fn}), \tag{2.18}$$

where $\arg(x)$ denotes the phase of complex scalar x. This approach is worth a few comments. First, the Poisson distribution is formerly defined only for integers, which impairs the statistical interpretation of KL-NMF on uncountable data such as audio spectra (but one could assume an appropriate data scaling and a very fine quantization to work around this).[1]

[1] Actually, KL-NMF has interesting parallels with inference in probabilistic latent variable models of histogram data; see Shashanka, Raj, and Smaragdis (2008a).

Second, this approach enforces nonnegativity in a somehow arbitrary way by taking the absolute value of data $\mathbf{X}$. In contrast, with gaussian modeling, nonnegativity arises naturally through the variance fitting problem equivalence. Similarly, the reconstruction method enforces the components to have same phase as observation coefficients, while this is a consequence of Wiener filtering only in the gaussian modeling framework. Last, the component reconstruction method is not statistically grounded and is not conservative: $\mathbf{x}_n \approx \sum_{k=1}^{K} \hat{\mathbf{c}}_{k,n}$. Note that Wiener reconstruction is used with KL-NMF of the magnitude spectrum $|\mathbf{X}|$ by Smaragdis (2007), where it is presented as spectral filtering, and its conservativity is pointed out.

2.3.3 Multiplicative Noise.

Theorem 2 *(IS-NMF as ML estimation in gamma multiplicative noise). Consider the generative model*

$$\boldsymbol{V} = (\boldsymbol{W}\boldsymbol{H})\,.\,\boldsymbol{E}, \tag{2.19}$$

where $\boldsymbol{E}$ is multiplicative independent and identically distributed (i.i.d.) gamma noise with mean 1. Then, maximum likelihood estimation of $\mathbf{W}$ and $\boldsymbol{H}$ is equivalent to NMF of $\mathbf{V}$ into $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$, where the Itakura-Saito divergence is used.

Proof. Let us note $\{e_{fn}\}$, the entries of $\mathbf{E}$. We have $v_{fn} = \hat{v}_{fn}\, e_{fn}$, with $p(e_{fn}) = \mathcal{G}(e_{fn} \mid \alpha, \beta)$, and where $\mathcal{G}(x \mid \alpha, \beta)$ is the gamma probability density function (PDF) defined in appendix A. Under the iid noise assumption, the minus log likelihood $C_{ML,2}(\mathbf{W}, \mathbf{H}) \overset{\text{def}}{=} -\log p(\mathbf{V} \mid \mathbf{W}, \mathbf{H})$ is

$$\begin{aligned}
C_{ML,2}(\mathbf{W}, \mathbf{H}) &= -\sum_{f,n} \log p(v_{fn} \mid \hat{v}_{fn}) && (2.20)\\
&= -\sum_{f,n} \log \mathcal{G}\left(v_{fn}/\hat{v}_{fn} \mid \alpha, \beta\right)/\hat{v}_{fn} && (2.21)\\
&\overset{c}{=} \beta \sum_{f,n} \frac{v_{fn}}{\hat{v}_{fn}} - \frac{\alpha}{\beta} \log \frac{v_{fn}}{\hat{v}_{fn}} - 1. && (2.22)
\end{aligned}$$

The ratio α/β is simply the mean of the gamma distribution. When it is equal to 1, we obtain that $C_{ML,2}(\ )$ is equal to $D_{IS}(\mathbf{V} \mid \hat{\mathbf{V}}) = D_{IS}(\mathbf{V} \mid \mathbf{WH})$ up to a positive factor and a constant.

The multiplicative noise equivalence explains the scale invariance of the IS divergence because the noise acts as a scale factor on $\hat{v}_{fn}$. In contrasts EUC-NMF is equivalent to the ML likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in additive iid gaussian noise. The influence of additive noise is greater on coefficients of $\hat{\mathbf{V}}$ with small amplitude (i.e., low SNR) than on the largest ones. As to KL-NMF, it corresponds to neither multiplicative nor additive noise but to ML estimation in Poisson noise.[2] To summarize, we have

$$\text{EUC-NMF: } p(v_{fn} \mid \hat{v}_{fn}) = \mathcal{N}(v_{fn}|\hat{v}_{fn}, \sigma^2), \tag{2.23}$$

$$\text{KL-NMF: } p(v_{fn} \mid \hat{v}_{fn}) = \mathcal{P}(v_{fn}|\hat{v}_{fn}), \tag{2.24}$$

$$\text{IS-NMF: } p(v_{fn} \mid \hat{v}_{fn}) = \frac{1}{\hat{v}_{fn}} \mathcal{G}\left(\frac{v_{fn}}{\hat{v}_{fn}}\middle|\alpha, \alpha\right), \tag{2.25}$$

and in all cases, $\mathrm{E}\{v_{fn} \mid \hat{v}_{fn}\} = \hat{v}_{fn}$.

Theorem 2 reports in essence how Abdallah and Plumbley (2004) derive a "statistically motivated error measure," which happens to be the IS divergence, in the very similar context of nonnegative sparse coding (see also developments in Plumbley et al., 2006). Pointing out the scale invariance of this measure, this work leads Virtanen (2007) to consider the IS divergence (but again without referring to it as such) for NMF in the context of single-channel source separation, but the algorithm is applied to the magnitude spectra instead of the power spectra, losing statistical coherence, and the sources are reconstructed through equation 2.18 instead of Wiener filtering.

3 Algorithms for NMF with the Itakura-Saito Divergence

In this section, we describe two algorithms for IS-NMF. The first one has a multiplicative structure and is only a special case of the derivations of section 2.1. The second one is a novel type, EM based, and is derived from the statistical presentation of IS-NMF given in theorem 1.

3.1 Multiplicative Gradient Descent Algorithm.

A multiplicative gradient descent IS-NMF algorithm is obtained by setting either $\beta = 0$ in equations 2.5 and 2.6 or $\phi(y) = -\log(y)$ in equations 2.7 and 2.8. The resulting update rules coincide and lead to algorithm 1:

```
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 $i = 1: n_{iter}$ 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
```

[2] KL-NMF is wrongly presented as ML estimation in additive Poisson noise in numerous publications.

These update rules were also obtained by Abdallah and Plumbley (2004), prior to Dhillon and Sra (2005) and Cichocki, Zdunek et al. (2006). In the following, we refer to this algorithm as IS-NMF/MU. This algorithm includes a normalization step at every iteration, which eliminates trivial scale indeterminacies, leaving the cost function unchanged. We impose $\|\mathbf{w}_k\|_2 = 1$ and scale h_k accordingly. Again, we emphasize that continual descent of the cost function is observed in practice with this algorithm but that a proof of convergence is yet to be found.

3.2 SAGE Algorithm. We now describe an EM-based algorithm for estimating the parameters $\ = \{\mathbf{W}, \mathbf{H}\}$, derived from the statistical formalism introduced in theorem 1. The additive structure of the generative model, equation 2.10, allows updating the parameters describing each component $\mathbf{C}_k \overset{\text{def}}{=} [\mathbf{c}_{k,1}, \ldots, \mathbf{c}_{k,N}]$ separately, using SAGE (Fessler & Hero, 1994). SAGE is an extension of EM for data models with particular structures, including data generated by superimposed components. It is known to converge faster in iterations than standard EM, though one iteration of SAGE is usually more computationally demanding than EM as it usually requires updating the sufficient statistics "more often." Let us consider a partition of the parameter space $\ = \bigcup_{k=1}^{K} \ _k$ with

$$\ _k = \{\mathbf{w}_k, h_k\}, \tag{3.1}$$

where we recall that $\mathbf{w}_k$ is the kth column of $\mathbf{W}$ and h_k is the kth row of $\mathbf{H}$. The SAGE algorithm involves choosing for each subset of parameters $\ _k$ a hidden-data space that is complete for this particular subset. Here, the hidden-data space for $\ _k$ is simply chosen to be $\mathbf{C}_k \overset{\text{def}}{=} [\mathbf{c}_{k,1}, \ldots, \mathbf{c}_{k,N}]$. An EM-like functional is then built for each subset $\ _k$ as the conditional expectation of the minus log likelihood of $\mathbf{C}_k$:

$$Q_k^{ML}(\ _k \mid \ ') \overset{\text{def}}{=} -\int_{\mathbf{C}_k} \log p(\mathbf{C}_k \mid \ _k)\, p(\mathbf{C}_k \mid \mathbf{X}, \ ')\, d\mathbf{C}_k. \tag{3.2}$$

One iteration i of the SAGE algorithm then consists of computing (E-step) and minimizing (M-step) $Q_k^{ML}(\ _k \mid \ ')$ for $k = 1, \ldots, K$. Note that $\ '$ always contains the most up-to-date parameter values, and not only the values at iteration $i - 1$ as in standard EM. This leads to the increase in computational burden, which is mild in our case.

The derivations of the SAGE algorithm for IS-NMF are detailed in appendix B. However, for a fixed k, the E-step merely consists of computing the posterior power $\mathbf{V}_k$ of component $\mathbf{C}_k$, defined by $[\mathbf{V}_k]_{fn} = v_{k,fn} = |\mu_{k,fn}^{post}|^2 + \lambda_{k,fn}^{post}$, where $\mu_{k,fn}^{post}$ and $\lambda_{k,fn}^{post}$ are the posterior mean and variance of $c_{k,fn}$, given by

$$\mu_{k,fn}^{post} = \frac{w_{fk}\, h_{kn}}{\sum_l w_{fl}\, h_{ln}}\, x_{fn}, \tag{3.3}$$

$$\lambda_{k,fn}^{post} = \frac{w_{fk}\, h_{kn}}{\sum_l w_{fl}\, h_{ln}} \sum_{l \neq k} w_{fl}\, h_{ln}. \tag{3.4}$$

The M-step is then shown to amount to the following one-component NMF problem,

$$\min_{\mathbf{w}_k,\, h_k \geq \mathbf{0}} D_{IS}(\mathbf{V}'_k \mid \mathbf{w}_k\, h_k), \tag{3.5}$$

where $\mathbf{V}'_k$ denotes $\mathbf{V}_k$ as computed from $'$. Interestingly, in the one-component case, the gradients simplify to

$$\nabla_{h_{kn}} Q_k^{ML}(\mathbf{w}_k, h_k \mid \;') = \frac{F}{h_{kn}} - \frac{1}{h_{kn}^2} \sum_{f=1}^{F} \frac{v'_{k,fn}}{w_{fk}}, \tag{3.6}$$

$$\nabla_{w_{fk}} Q_k^{ML}(\mathbf{w}_k, h_k \mid \;') = \frac{N}{w_{fk}} - \frac{1}{w_{fk}^2} \sum_{n=1}^{N} \frac{v'_{k,fn}}{h_{kn}}. \tag{3.7}$$

The gradients are easily zeroed, leading to the following updates,

$$h_{kn}^{(i+1)} = \frac{1}{F} \sum_f \frac{v'_{k,fn}}{w_{fk}^{(i)}}, \tag{3.8}$$

$$w_{fk}^{(i+1)} = \frac{1}{N} \sum_n \frac{v'_{k,fn}}{h_{kn}^{(i+1)}}, \tag{3.9}$$

which guarantees $Q_k^{ML}(\mathbf{w}_k^{(i+1)}, h_k^{(i+1)} \mid \;') \leq Q_k^{ML}(\mathbf{w}_k^{(i)}, h_k^{(i)} \mid \;')$. This can also be written in matrix form, as shown in algorithm 2, which summarizes the SAGE algorithm for IS-NMF:

```
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 $i = 1 \colon n_{iter}$ do
    for $k = 1 \colon K$ 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 $\mathbf{WH} - \mathbf{w}_k^{old}h_k^{old} + \mathbf{w}_k^{new}h_k^{new}$.
```

In the following, we refer to this algorithm as IS-NMF/EM.

IS-NMF/EM and IS-NMF/MU have the same complexity $\mathcal{O}(12\,F\,K\,N)$ per iteration, but can lead to different run times, as shown in the results below. Indeed, in our Matlab implementation, the operations in IS-NMF/MU can be efficiently vectorized using matrix entrywise multiplication, while IS-NMF/EM requires looping over the components, which is more time-consuming.

The convergence of IS-NMF/EM to a stationary point of $D_{IS}(\mathbf{V} \mid \mathbf{WH})$ is granted by property of SAGE. However, it can converge only to a point in the interior domain of the parameter space; $\mathbf{W}$ and $\mathbf{H}$ cannot take entries equal to zero. This is seen in equation 3.5: if w_{fk} or h_{kn} is zero, then the cost $d_{IS}(v'_{k,fn} \mid w_{fk}h_{kn})$ becomes infinite. This is not a feature shared by IS-NMF/MU, which does not a priori exclude zero coefficients in $\mathbf{W}$ and $\mathbf{H}$ (but excludes $\hat{v}_{fn} = 0$, which would lead to a division by zero). However, because zero coefficients are invariant under multiplicative updates (see section 2.1), if IS-NMF/MU attains a fixed-point solution with zero entries, then it cannot be determined if the limit point is a stationary point. Yet if the limit point does not take zero entries (i.e., it belongs to the interior of the parameter space), then it is a stationary point, which may or may not be a local minimum. This is stressed by Berry et al. (2007) for EUC-NMF but holds for IS-NMF/MU as well.

Note that SAGE has been used in the context of single-channel source separation by Ozerov, Philippe, Bimbot, and Gribonval (2007) for inference on a model somehow related to the IS-NMF model, equation 2.10. Indeed, these authors address voice and music separation using a generative model of the form $\mathbf{x}_n = \mathbf{c}_{V,n} + \mathbf{c}_{M,n}$ where the first component represents voice and the second one represents music. Then each component is given a gaussian mixture model (GMM). The GMM parameters for voice are learned from training data, while the music parameters are adapted to data. Though related, the GMM and NMF models are quite different in essence. The first one expresses the signal as a sum of two components that each can take different states. The second one expresses the signal as a sum of K components, each representative of one object. It cannot be claimed that one model is better than the other; rather, they address different characteristics. It is anticipated that the two models can be used jointly within the SAGE framework, for example, by modeling voice $\mathbf{c}_{V,n}$ with a GMM (i.e., a specific component with many states) and music $\mathbf{c}_{M,n}$ with an NMF model (i.e., a composite signal with many components).

4 Analysis of a Short Piano Excerpt

In this section, we report an experimental comparative study of the NMF algorithms applied to the spectrogram of a short monophonic piano sequence. In the first step, we compare the results of multiplicative Euclidean, KL, and IS NMF algorithms for several values of K before we more specifically compare the multiplicative and EM-based algorithms for IS-NMF in the second step.

4.1 Experimental Setup. A piano sequence played from the score given in Figure 2 on a Yamaha DisKlavier MX100A upright piano was recorded in a small-size room by a Schoeps omnidirectional microphone, placed about 15 cm (6 inches) above the opened body of the piano. The sequence is composed of four 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 $\nu_s = 22{,}050$ Hz, yielding $T = 339{,}501$ 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 in Figure 2.

IS-NMF/MU, IS-NMF/EM, and the multiplicative gradient descent NMF algorithms with Euclidean and KL costs were implemented in Matlab and run on data $\mathbf{V} = |\mathbf{X}|^{\cdot[2]}$. Note that in the following, the terms *EUC-NMF* and *KL-NMF* will implicitly refer to the multiplicative implementation of these NMF techniques. All algorithms were run for several values of the number of components, more specifically, for $K = 1, \ldots, 10$. For each value of K, 10 runs of each algorithm were produced from 10 random initializations of $\mathbf{W}$ and $\mathbf{H}$, chosen, in Matlab notation, as `W = abs(randn(F,K)) + ones(F,K)` and `H = abs(randn(K,N)) + ones(K,N)`. The algorithms were run for $n_{iter} = 5000$ iterations.

4.2 Pitch Estimation. In the following results, it will be observed that some of the basis elements (columns of $\mathbf{W}$) have a pitched structure, characteristic of individual musical notes. If pitch estimation is not the objective per se of the following study, it is informative to check if correct pitch values can be inferred from the factorization. As such, a fundamental frequency (or pitch) estimator is applied using the method described in Vincent, Bertin, and Badeau (2007). It consists of computing dot products of $\mathbf{w}_k$ with a set of J frequency combs and retaining the pitch number corresponding to the largest dot product. Each comb is a cosine function with period f_j, scaled and shifted to the amplitude interval [0 1], which takes its maximum value

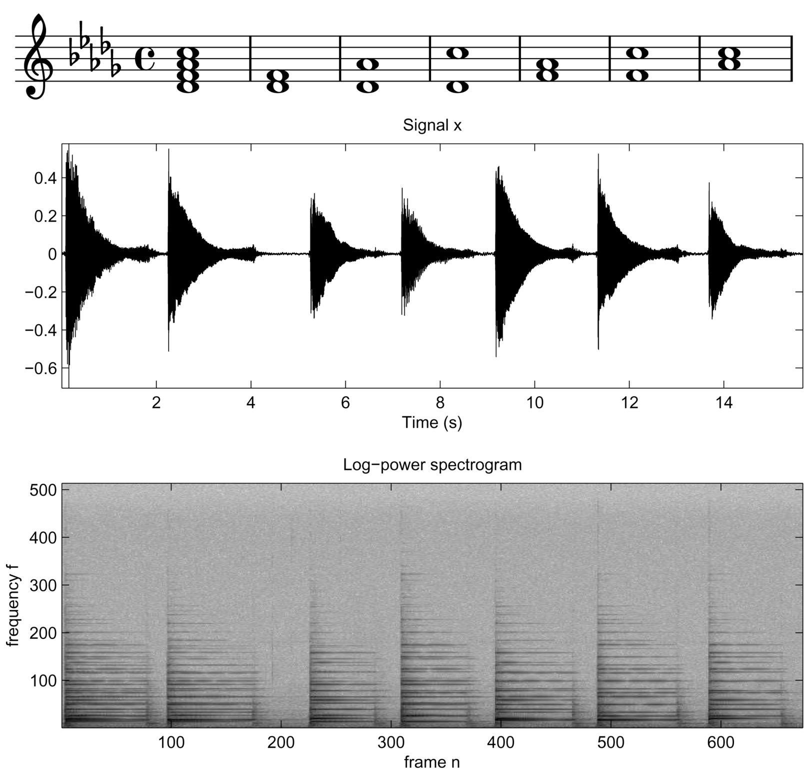

Figure 2: Three representations of data: (Top) Original score. (Middle) Time-domain recorded signal x. (Bottom) Log-power spectrogram $\log |\mathbf{X}|^{\cdot[2]}$. The four notes read $D_4^\flat$ (pitch 61), F_4 (pitch 65), $A_4^\flat$ (pitch 68), and C_5 (pitch 72). Together they form a $D^\flat$ major seventh chord. In the recorded interpretation, the third chord is slightly out of tempo.

1 at bins multiple of f_j. The set of fundamental frequency bins $f_j = \frac{\nu_j}{\nu_s} L$ is indexed on the MIDI logarithmic scale, such that

$$\nu_j = 440 \times 2^{\frac{p_j - 69}{12}}. \tag{4.1}$$

The piano note range usually goes from $p_{\min} = 21$, that is, note A_0 with fundamental frequency $f_{\min} = 27.5$ Hz, to $p_{\max} = 108$, that is, note C_8 with frequency $f_{\max} = 4186$ Hz. Two adjacent keys are separated by a semitone ($\Delta p = 1$). The MIDI pitch numbers of the notes pictured in Figure 2 are 61 ($D_4^\flat$), 65 (F_4), 68 ($A_4^\flat$), and 72 (C_5) and were chosen arbitrarily. In our implementation of the pitch estimator, the MIDI range was sampled from 20.6 to 108.4 with step 0.2. In the following, an arbitrary pitch value of 0 will be given to unpitched basis elements. The classification of pitched and

Table 1: Run Times in Seconds of 1000 Iterations of the NMF Algorithms Applied to the Piano Data.

K	1	2	3	4	5	10	$\mathcal{O}(.)$
EUC-NMF	17	18	20	24	27	37	$4\,F\,K\,N + 2\,K^2(F+N)$
KL-NMF	90	90	92	100	107	117	$8\,F\,K\,N$
IS-NMF/MU	127	127	129	135	138	149	$12\,F\,K\,N$
IS-NMF/EM	81	110	142	171	204	376	$12\,F\,K\,N$

Notes: This was implemented in Matlab on a 2.16 GHz Intel Core 2 Duo iMac with 2 GB RAM. The run times include the computation of the cost function at each iteration (for possible convergence monitoring). The last column shows the algorithm complexities per iteration, expressed in number of flops (addition, subtraction, multiplication, division). The complexity of EUC-NMF assumes $K < F, N$.

unpitched elements was done manually by looking at the basis elements and listening to the component reconstructions.

4.3 Results and Discussion

4.3.1 Convergence Behavior and Algorithm Complexities. Run times of 1000 iterations of each of the four algorithms are shown in Table 1, together with the algorithm complexities. Figure 3 shows for each algorithm and for every value of K the final cost values of the 10 runs, after the 5000 algorithm iterations. A first observation is that the minimum and maximum cost values differ: $K > 4$ in the Euclidean case, $K > 3$ in the KL case, and $K > 2$ in the IS case. This means either that the algorithms have failed to converge after 5000 iterations in some cases or the presence of local minima. Figure 4 displays for all four algorithms the evolution of the cost functions along the 5000 iterations for all 10 runs in the case $K = 6$.

4.3.2 Evolution of the Factorizations with Order K. In this paragraph, we examine in detail the underlying semantics of the factorizations obtained with all three cost functions. We address only the comparison of factorizations obtained from the three multiplicative algorithms. IS-NMF/EM and IS-NMF/MU will be more specifically compared in the next paragraph. Otherwise stated, the factorizations studied are those obtained from the run yielding the minimum cost value among the 10 runs. Figures 5 to 8 display the columns of $\mathbf{W}$ and corresponding rows of $\mathbf{H}$. The columns of $\mathbf{W}$ are represented against frequency bin f on the left (in $\log_{10}$ amplitude scale), and the rows of $\mathbf{H}$ are represented against frame index n on the right (in linear amplitude scale). Pitched components are displayed first (top to bottom, in ascending order of estimated pitch value), followed by the unpitched components. We reproduce only part of the results in this letter, but the factorizations obtained with all four algorithms for $K = 4, 5, 6$ are available online at http://www.tsi.enst.fr/~fevotte/Samples/is-nmf, together with

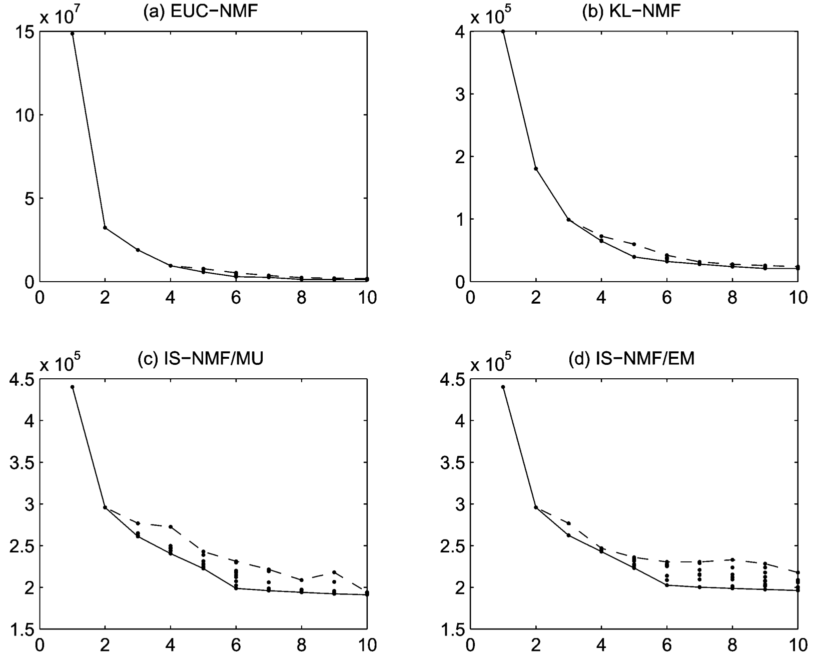

Figure 3: Cost values after 5000 iterations, obtained from 10 random initializations. (a) Euclidean distance. (b) KL divergence. (c) IS divergence (using IS-NMF/MU). (d) IS divergence (using IS-NMF/EM). On each plot, the solid line connects all minimum cost values, and the dashed line connects all maximum cost values.

sound reconstructions of the individual components. Component STFTs $\hat{\mathbf{C}}_k$ were computed by applying the Wiener filter, equation 2.15, to $\mathbf{X}$ using the factors $\mathbf{W}$ and $\mathbf{H}$ obtained with all three cost functions. Time-domain components c_k were then reconstructed by inverting the STFTs using an adequate overlap-add procedure with dual synthesis window. By conservativity of Wiener reconstruction and linearity of the inverse STFT, the time domain decomposition is also conservative, such that

$$x = \sum_{k=1}^{K} c_k. \tag{4.2}$$

Common sense suggests that choosing as many components as notes forms a sensible guess for the value of K so as to obtain a meaningful factorization of $|\mathbf{X}|^{\cdot[2]}$, where each component would be expected to represent one and only one note. The factorizations obtained with all three costs for $K = 4$ prove that this is not the case. Euclidean and KL-NMF rather

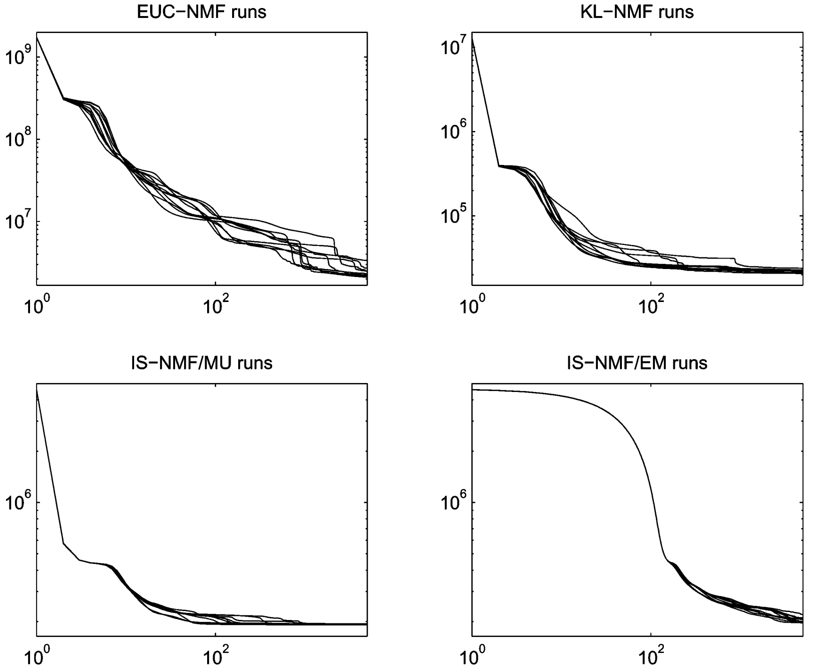

Figure 4: Evolution in log-log scale of the cost functions along the 5000 iterations of all 10 runs of the four algorithms in the specific case of $K = 6$.

successfully extract notes 65 and 68 into separate components (second and third), but notes 61 and 72 are melted into the first component, while a fourth component seems to capture transient events corresponding to the note attacks (the sound of the hammer hitting the string) and the sound produced by the release of the sustain pedal. The first two components obtained with IS-NMF have a similar interpretation to those given by EUC-NMF and KL-NMF. However, the two other components differ in nature: the third component comprises note 68 and transients, while the fourth component is akin to residual noise. It is interesting to notice how this last component, though of much lower energy than the other components (on the order of 1 compared to 10^4 for the others) bears equal importance in the decomposition. This is undoubtedly a consequence of the scale invariance property of the IS divergence discussed in section 2.2.

A fully separated factorization (at least as intended) is obtained for $K = 5$ with KL-NMF, as displayed in Figure 5. This results in four components, each made up of a single note, and a fifth component containing sound events corresponding to note attacks and pedal releases. However, these latter events are not well localized in time and suffer from an unnatural tremolo effect (oscillating variations in amplitudes), as can be heard from

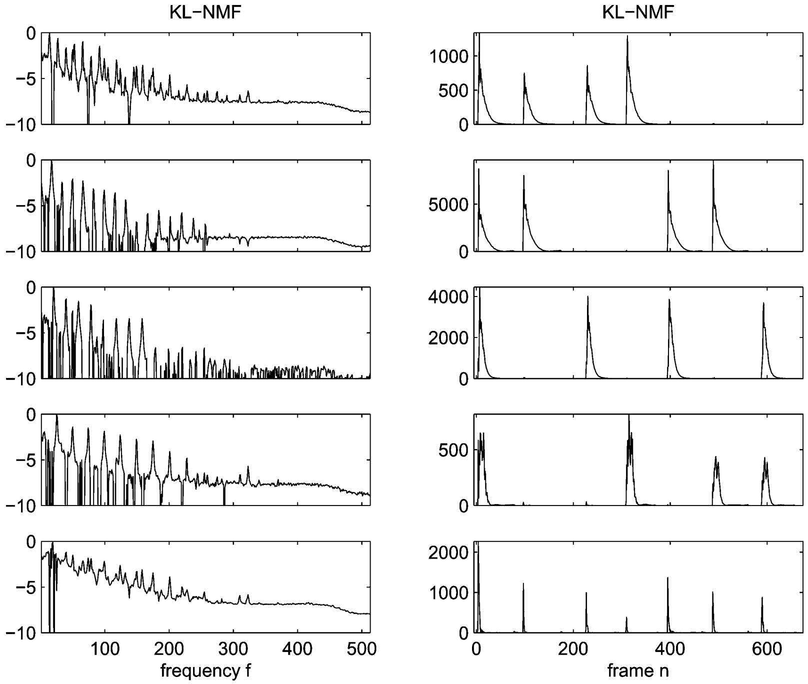

Figure 5: KL-NMF with $K = 5$. Pitch estimates: [61 65 68 72.2 0]. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Rows of $\mathbf{H}$.

the reconstructed sound files. Surprisingly, the decomposition obtained with EUC-NMF by setting $K = 5$ results in splitting the second component of the $K = 4$ decomposition into two components with estimated pitches 65 and 65.4 instead of actually demixing the third component, which comprises notes 61 and 72. As for IS-NMF, the first component now groups notes 61 and 68; the second and third components, respectively, capture notes 65 and 72; the fourth component is still akin to residual noise; and the fifth component perfectly renders the attacks and releases.

Full separation of the individual notes is finally obtained with Euclidean and IS costs for $K = 6$, as shown in Figures 6 and 7. KL-NMF produces an extra component (with pitch estimate 81) that is not clearly interpretable and is in particular not akin to residual noise as could have been hoped for. The decomposition obtained with the IS cost describes as follows. The four first components correspond to individual notes whose pitch estimate matches exactly the pitches of the notes played. The visual aspect of the PSDs is much better than the basis elements learned from EUC-NMF and KL-NMF. The fifth component captures the hammer hits and pedal releases with great accuracy, and the sixth component is akin to residual noise.

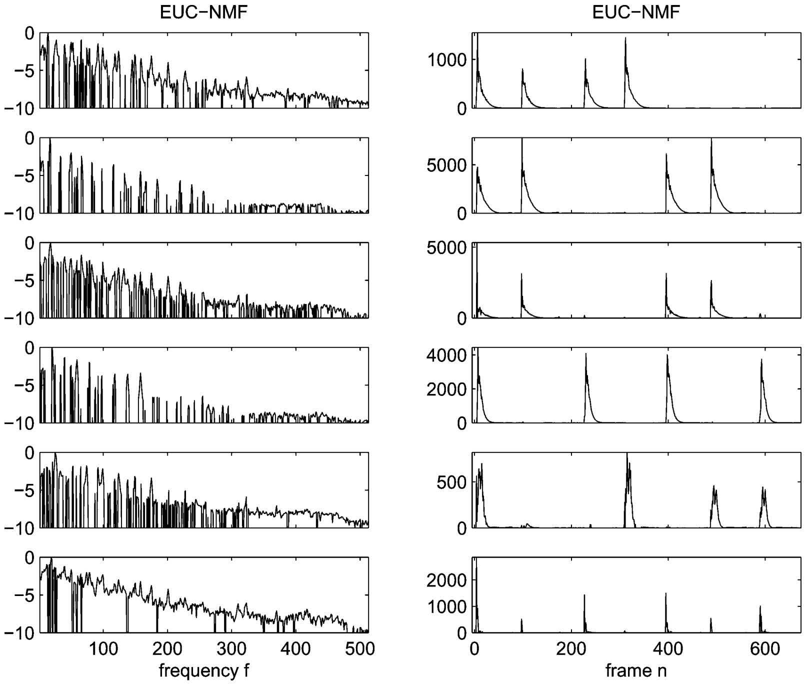

Figure 6: EUC-NMF with $K = 6$. Pitch estimates: [61 65 65.4 68 72 0]. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Rows of $\mathbf{H}$.

When the decomposition is carried beyond $K = 6$, EUC-NMF and KL-NMF split existing components into several subcomponents (such as components capturing sustained and decaying parts of one note) with pitch in the neighborhood of the note fundamental frequency. In contrast, IS-NMF/MU spends the extra components in fine-tuning the representation of the low-energy components—residual noise and transient events (as such, the hammer hits and pedal releases eventually get split in two distinct components). For $K = 10$, the pitch estimates read EUC-NMF: [61 64.8 64.8 65 65 65.8 68 68.4 72.2 0], KL-NMF: [61 61 65 65 66 68 72 80.2 0 0], IS-NMF/MU: [61 61 65 68 72 0 0 0 0 0]. If note 61 is indeed split into two components with IS-NMF/MU, one of the two components is actually inaudible.

The message of this experimental study is that the nature of the decomposition obtained with IS-NMF, and its progression as K increases, is in accord with an object-based representation of music, close to our own comprehension of sound. Entities with well-defined semantics emerge from the decomposition (individual notes, hammer hits, pedal releases, residual noise), while the decompositions obtained from the Euclidean and KL costs are less interpretable from this perspective. These conclusions do not

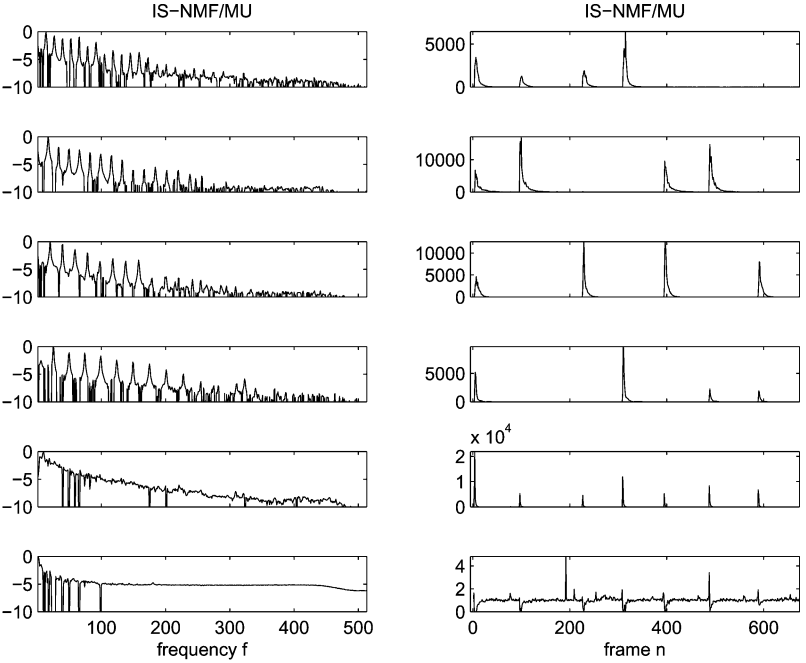

Figure 7: IS-NMF/MU with $K = 6$. Pitch estimates: [61 65 68 72 0 0]. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Rows of $\mathbf{H}$.

always hold when the factorization is not the one yielding the lowest-cost values from the 10 runs. As such, we also examined the factorizations with highest-cost values (with all three cost functions), and we found out that they did not reveal the same semantics, which was not always easily interpretable. The upside, however, is that the lowest IS cost values correspond to the most desirable factorizations, so that IS-NMF "makes sense."

4.3.3 Comparison of Multiplicative and EM-Based IS-NMF. Algorithms IS-NMF/MU and IS-NMF/EM are designed to address the same task of minimizing the cost $D_{IS}(\mathbf{V} \mid \mathbf{WH})$, so that the achieved factorization should be identical in nature, provided they complete this task. As such, the progression of the factorization provided by IS-NMF/EM is similar to the one observed for IS-NMF/MU, described in the previous paragraph. However, the resulting factorizations are not exactly equivalent, because IS-NMF/EM does not inherently allow zeros in the factors (see section 3.2). This feature can be desirable for $\mathbf{W}$, as the presence of sharp notches in the spectrum may not be physically realistic for audio, but can be considered a drawback as far as $\mathbf{H}$ is concerned. Indeed, the rows of $\mathbf{H}$ being akin to activation

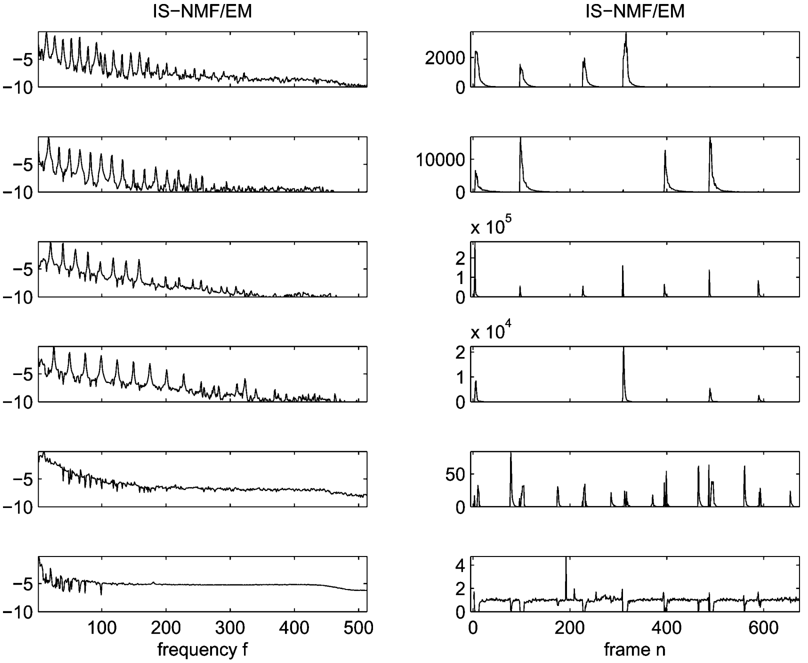

Figure 8: IS-NMF/EM with $K = 6$. Pitch estimates: [61 65 68 72 0 0]. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Rows of $\mathbf{H}$.

coefficients, when a sound object k is not present in frame n, then h_{kn} should be strictly zero. These remarks probably explain the factorization obtained from IS-NMF/EM with $K = 6$, displayed in Figure 8. The notches present in the PSDs learned with IS-NMF/MU, as seen in Figure 7, have disappeared from the PSDs on Figure 8, which exhibit better regularity. Unfortunately, IS-NMF/EM does not fully separate out the note attacks in the fifth component as IS-NMF/MU does. Indeed, some parts of the attacks appear in the second component, and the rest appear in the fifth component, which also contains the pedal releases. This is possibly explained by the a priori high sparsity of a transients component, which can be handled by IS-NMF/MU but not IS-NMF/EM (because it does not allow zero values in $\mathbf{H}$). Increasing the number of components K or the number of algorithm iterations n_{iter} does not solve this specific issue.

Regarding the compared convergence of the algorithms, IS-NMF/MU decreases the cost function much faster in the initial iterations and, with this data set, attains lower final cost values than IS-NMF/EM, as shown in Figure 3 or 4 for $K = 6$. Although the two algorithms have the same complexity, the run time per iteration of IS-NMF/MU is smaller than IS-NMF/EM for $K > 3$ (see Table 1).

5 Regularized IS-NMF

We now describe how the statistical setting of IS-NMF can be exploited to incorporate regularization constraints and prior information in the factors estimates.

5.1 Bayesian Setting. We consider a Bayesian setting where **W** and **H** are given (independent) prior distributions $p(\mathbf{W})$ and $p(\mathbf{H})$. We are looking for a joint MAP estimate of **W** and **H** through minimization of criterion

$$C_{MAP}(\mathbf{W},\mathbf{H}) \overset{\text{def}}{=} -\log p(\mathbf{W},\mathbf{H} \mid \mathbf{X}) \tag{5.1}$$

$$\overset{c}{=} D_{IS}(\mathbf{V} \mid \mathbf{WH}) - \log p(\mathbf{W}) - \log p(\mathbf{H}). \tag{5.2}$$

When independent priors of the form $p(\mathbf{W}) = \prod_k p(\mathbf{w}_k)$ and $p(\mathbf{H}) = \prod_k p(h_k)$ are used, then the SAGE algorithm presented in section 3.2 can be used again for MAP estimation. In that case, the functionals to be minimized for each component k are

$$Q_k^{MAP}(\ _k \mid \ ') \overset{\text{def}}{=} -\int_{\mathbf{C}_k} \log p(\ _k \mid \mathbf{C}_k)\, p(\mathbf{C}_k \mid \mathbf{X},\ ')\, d\mathbf{C}_k \tag{5.3}$$

$$\overset{c}{=} Q_k^{ML}(\mathbf{w}_k, h_k \mid \ ') - \log p(\mathbf{w}_k) - \log p(h_k). \tag{5.4}$$

Thus, the E-step still amounts to computing $Q_k^{ML}(\mathbf{w}_k, h_k \mid \ ')$, as done in section 3.2, 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.

Next, we more specifically consider Markov chain priors favoring smoothness over the rows of **H**. In the following results, no prior structure will be assumed for **W** (i.e., **W** is estimated through ML). However, we stress that the methodology presented for the rows of **H** can be transposed to the columns of **W**, that prior structures can be imposed on both **W** and **H**, and that these structures need not belong to the same class of models. Note also that since the components are treated separately, each can be given a different type of model (e.g., some components could be assigned a GMM, as discussed at the end of section 3.2).

We assume the following prior structure for h_k,

$$p(h_k) = \prod_{n=2}^{N} p\big(h_{kn} \mid h_{k(n-1)}\big)\; p(h_{k1}), \tag{5.5}$$

where $p(h_{kn} \mid h_{k(n-1)})$ 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. Possible PDF choices are, for $n = 2, \ldots, N$,

$$p\big(h_{kn} \mid h_{k(n-1)}\big) = \mathcal{IG}\big(h_{kn} \mid \alpha, (\alpha+1)\, h_{k(n-1)}\big) \tag{5.6}$$

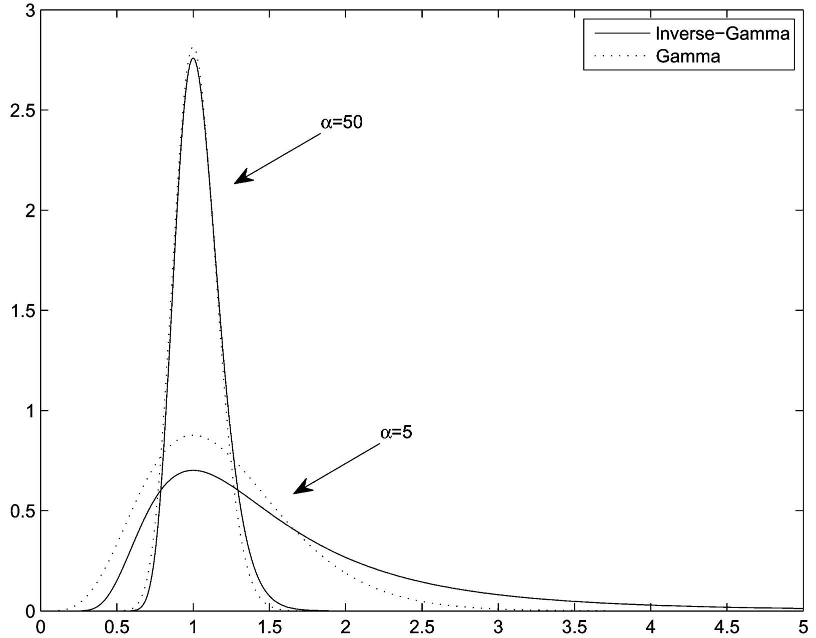

Figure 9: Prior PDFs $\mathcal{IG}(h_{kn} \mid \alpha - 1, \alpha\, h_{k(n-1)})$ (solid line) and $\mathcal{G}(h_{kn} \mid \alpha + 1, \alpha / h_{k(n-1)})$ (dashed line) for $h_{k(n-1)} = 1$ and for $\alpha = \{5, 50\}$.

and

$$p\big(h_{kn} \mid h_{k(n-1)}\big) = \mathcal{G}\big(h_{kn} \mid \alpha, (\alpha - 1)/h_{k(n-1)}\big), \tag{5.7}$$

where $\mathcal{G}(x|\alpha, \beta)$ is the previously introduced gamma PDF, with mode $(\alpha - 1)/\beta$ (for $\alpha \geq 1$) and $\mathcal{IG}(x|\alpha, \beta)$ is the inverse-gamma PDF (see appendix A), with mode $\beta/(\alpha + 1)$. Both priors are constructed so that their 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 thus accentuate the smoothness of h_k, while a low value of α will render the prior more diffuse and thus less constraining. The two priors become very similar for large values of α (see Figure 9). In the following, h_{k1} is assigned the scale-invariant Jeffreys noninformative prior $p(h_{k1}) \propto 1/h_{k1}$.

5.2 New Updates. Under prior structure 5.5, the derivative of $Q_k^{MAP}(\mathbf{w}_k, h_k \mid \ ')$ with regard to h_{kn} writes $\forall n = 2, \ldots, N - 1$,

$$\begin{aligned}\nabla_{h_{kn}} Q_k^{MAP}(\mathbf{w}_k, h_k \mid \ ') = \nabla_{h_{kn}} Q_k^{ML}(\mathbf{w}_k, h_k \mid \ ') - \nabla_{h_{kn}} \log p\big(h_{k(n+1)} \mid h_{kn}\big)\\ -\nabla_{h_{kn}} \log p\big(h_{kn} \mid h_{k(n-1)}\big).\end{aligned} \tag{5.8}$$

Table 2: Coefficients of the Order 2 Polynomial to Solve in Order to Update h_{kn} in Bayesian IS-NMF with a Markov Chain Prior.

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

Note: $\hat{h}_{kn}^{ML}$ denotes the ML update, given by equation 3.8.

This is shown to be equal to

$$\nabla_{h_{kn}} Q_k^{MAP}(\mathbf{w}_k, h_k \mid \quad ') = \frac{1}{h_{kn}^2}\big(p_2\, h_{kn}^2 + p_1\, h_{kn} + p_0\big), \tag{5.9}$$

where the values of p_0, p_1, and p_2 are specific to the type of prior employed (gamma or inverse-gamma chains), as given in Table 2. Updating h_{kn} then simply amounts to solving an order 2 polynomial. The polynomial has only one nonnegative root, given by

$$h_{kn} = \frac{\sqrt{p_1^2 - 4\,p_2\,p_0} - p_1}{2\,p_2}. \tag{5.10}$$

The coefficients h_{k1} and h_{kN} at the borders of the Markov chain require specific updates, but they also require solving polynomials of order 2 or 1, with coefficients given in Table 2 as well.

Note that the difference between the updates with the gamma and inverse-gamma chains prior mainly amounts to interchanging the positions of $h_{k(n-1)}$ and $h_{k(n+1)}$ in p_0 and p_2. Interestingly, using a backward gamma chain prior $p(h_k) = \prod_{n=1}^{N-1} p(h_{kn} \mid h_{k(n+1)})\ p(h_{kN})$ with shape parameter α is actually equivalent (in terms of MAP updates) to using a forward inverse-gamma chain prior as in equation 5.5 with shape parameter $\alpha - 2$. Respectively, using a backward inverse-gamma chain prior with shape parameter α is equivalent to using a forward-gamma chain prior with shape parameter $\alpha + 2$.

Note that Virtanen et al. (2008) recently considered gamma chains for regularization of KL-NMF. The modeling proposed in their work is, however, different from ours. Their gamma chain prior is constructed in a hierarchical setting, by introducing extra auxiliary variables, so as to ensure conjugacy of the priors with the Poisson observation model. Estimation of the factors is then carried out with the standard gradient descent multiplicative approach, and single-channel source separation results are presented from the factorization of the magnitude spectrogram $|\mathbf{X}|$ with component reconstruction 2.18. Regularized NMF algorithms for the Euclidean and KL costs with norm-2 constraints on $h_{kn} - h_{k(n-1)}$ have also been considered by Chen, Cichocki, and Rutkowski (2006) and Virtanen (2007). Finally, we also wish to mention that Shashanka, Raj, and Smaragdis (2008b) have recently derived a regularized version of KL-NMF with sparsity constraints in a Bayesian setting.

6 Learning the Semantics of Music with IS-NMF

The aim of the experimental study proposed in section 4 was to analyze the results of several NMF algorithms on a short, simple, and well-defined musical sequence with respect to the cost function, initialization, and model order. We now present the results of NMF on a long polyphonic recording. Our goal is to examine how much of the semantics NMF can learn from the signal, with a fixed number of components and a fixed random initialization. This is not easily assessed numerically in the most general context, but quantitative evaluations could be performed on specific tasks in simulation settings. Such tasks could include music transcription, as in Abdallah and Plumbley (2004), single-channel source separation, as in Benaroya et al. (2006, 2003), or content-based music retrieval based on NMF features.

Rather than choosing and addressing one of these specific tasks, we use NMF in an actual audio restoration scenario, where the purpose is to denoise and upmix original monophonic material (one channel) to stereo (two channels). This task is very close to single-channel source separation, with the difference that we are not aiming at perfectly separating each of the sources, but rather isolating subsets of coherent components that can be given different directions of arrival in the stereo remaster so as to render a sensation of spatial diversity. We will show in particular that the addition of smoothness constraints on the rows of $\mathbf{H}$ leads to more pleasing component reconstructions and brings out the pitched structure of some of the learned PSDs better.

6.1 Experimental Setup. We address the decomposition of a 108-second-long music excerpt from "My Heart (Will Always Lead Me Back to You)" recorded by Louis Armstrong and His Hot Five in the 1920s. The band features (to our best hearing) a trumpet, a clarinet, a trombone, a piano, and a double bass. The data are original unprocessed mono material containing substantial noise. The signal was downsampled to $\nu_s =$ 11,025 kHz, yielding $T =$ 1,191,735 samples. The STFT $\mathbf{X}$ of x was computed using a sinebell analysis window of length $L = 256$ (23 ms) with 50%

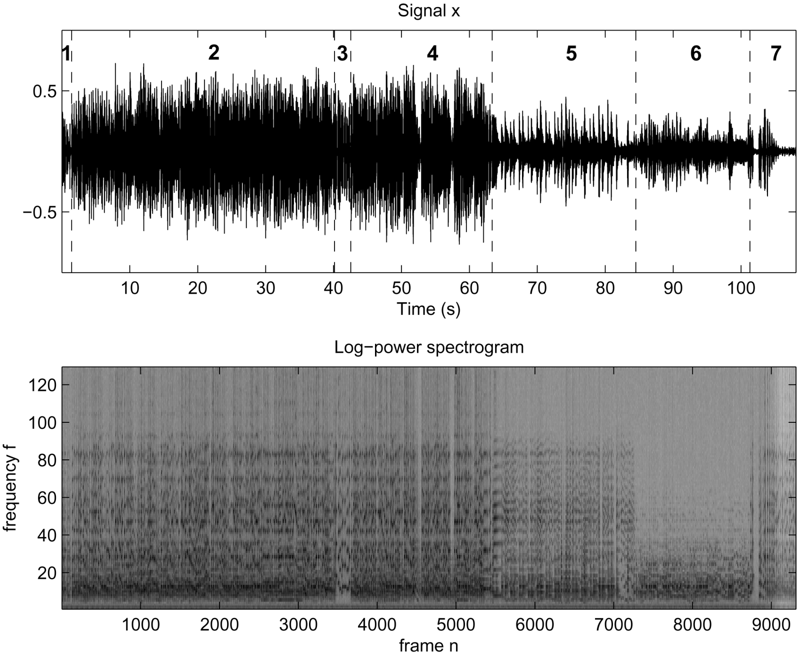

Figure 10: Original Louis Armstrong data. (Top) Time-domain recorded signal x. (Bottom) Log-power spectrogram. The vertical dashed lines on the top plot identify successive phases in the music piece, which we annotated manually: (2,4,7) all instruments, (1) clarinet only, (3) trumpet solo, (5) clarinet and piano, (6) piano solo.

overlap between two frames, leading to $N = 9312$ frames and $F = 129$ frequency bins. The time domain signal x and its log-power spectrogram are represented in Figure 10.

We applied EUC-NMF, KL-NMF, IS-NMF/MU, and IS-NMF/EM to $\mathbf{V} = |\mathbf{X}|^{\cdot[2]}$, as well as a regularized version of IS-NMF, as described in section 5. We used the inverse-gamma Markov chain prior (see equation 5.6) with α arbitrarily set to 10. We refer to this algorithm as IS-NMF/IG. Among many trials, this value of α provided a good trade-off between the smoothness of the component reconstructions and adequacy to data. Experiments with the gamma Markov chain prior, equation 5.6, did not lead to significant differences in the results and are not reported here.

The number of components K was arbitrarily set to 10. All five algorithms were run for $n_{iter} = 5000$ iterations and were initialized with the same random values. For comparison, we also applied KL-NMF to the magnitude spectrogram $|\mathbf{X}|$ with component reconstruction described by equation 2.18, as this can be considered state-of-the-art methodology for NMF-based single-channel audio source separation (Virtanen, 2007).

6.2 Results and Discussion. For conciseness, we here display only the decomposition obtained with IS-NMF/IG (see Figure 11) because it leads to the best results as far as our audio restoration task is concerned. (All decompositions and component reconstructions obtained from all NMF algorithms are available online at http://www.tsi.enst.fr/∼fevotte/Samples/is-nmf.) Figure 11 displays the estimated basis functions **W** in log-scale on the left and represents on the right the time-domain signal components reconstructed from Wiener filtering.

Figure 12 displays the evolution of the IS cost along the 5000 iterations with IS-NMF/MU, IS-NMF/EM, and IS-NMF/IG. In this case, IS-NMF/EM achieves a lower cost than IS-NMF/MU. The run times of 1000 iterations of the algorithms were, respectively: EUC-NMF, 1.9 min; KL-NMF, 6.8 min; IS-NMF/MU, 8.7 min; IS-NMF/EM, 23.2 min; and IS-NMF/IG, 32.2 min.

The comparison of the decompositions obtained with the three cost functions (Euclidean, KL, and IS), through visual inspection of **W** and listening to the components c_k, shows again that the IS divergence leads to the most interpretable results. In particular, some of the columns of matrix **W** produced by all three IS-NMF algorithms have a clear pitched structure, which indicates that some notes have been extracted. Furthermore, one of the components captures the hiss noise from the recording. Discarding this component from the reconstruction of x yields satisfying denoising (this is particularly noticeable during the piano solo, where the input SNR is low). Surprisingly, most of the rhythmic accompaniment (piano and double bass) is isolated in a single component (component 1 of IS-NMF/MU, component 2 of IS-NMF/EM and IS-NMF/IG), though its spectral content is clearly evolving in time. A similar effect happens with IS-NMF/IG and the trombone, which is mostly contained by component 7.

While we do not have a definite explanation for this, we believe that this is a consequence of Wiener reconstruction. Indeed, the Wiener component reconstruction is seen only as a set of K masking filters applied to x_{fn}, so that it does not constrain the spectrum of component k to be exactly $\mathbf{w}_k$ (up to amplitude h_{kn}), as the reconstruction method described by equation 2.18 does. So if one assumes that the IS-NMF model, equation 2.10, adequately captures some of the sound entities present in the mix (in our case, that would be the preponderant notes or chords and the noise), then the other entities are bound to be relegated in remaining components by conservativity of the decomposition $x = \sum_{k=1}^{K} c_k$.

As anticipated, the addition of frame-persistency constraints with IS-NMF/IG has an impact on the learned basis **W**. In particular, some of the components exhibit a more pronounced pitched structure. But more important, the regularization yields more pleasing sound reconstructions, which is particularly noticeable when listening to the accompaniment component obtained from IS-NMF/MU (component 1) or IS-NMF/EM (component 2) on the one side and from IS-NMF/IG (component 2) on the other side. Note also that in every case, the sound quality of Wiener

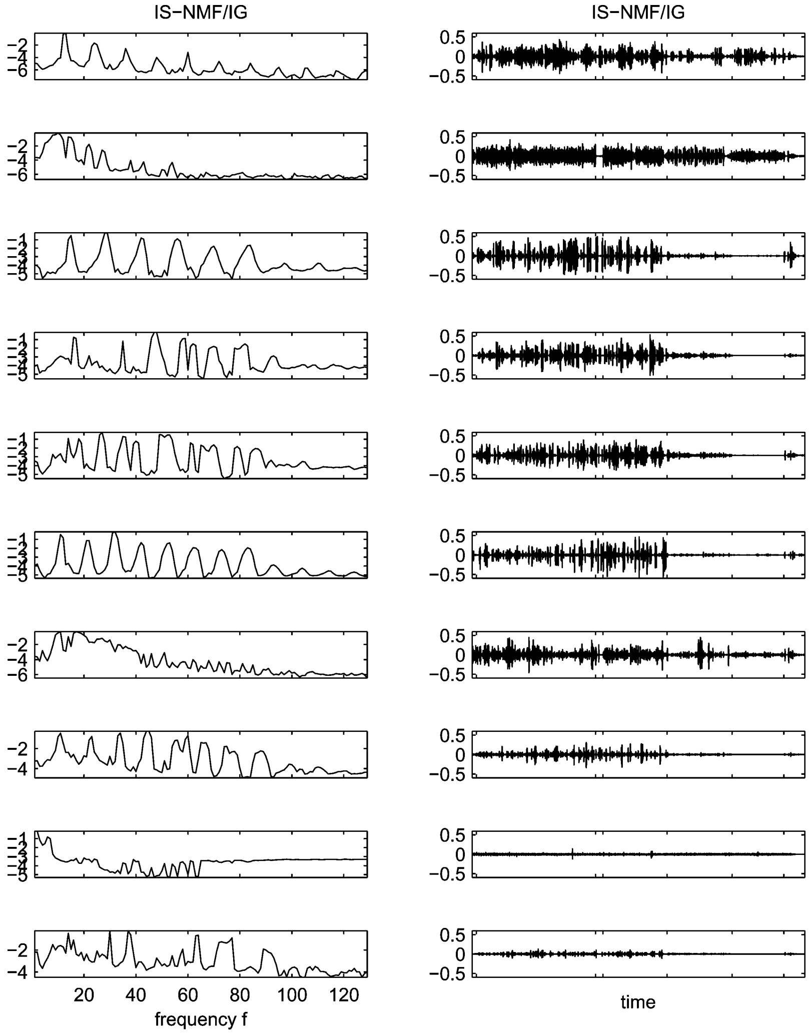

Figure 11: Decomposition of Louis Armstrong music data with IS-NMF/IG. (Left) Columns of **W** ($\log_{10}$ scale). (Right) Reconstructed components c_k. The x-axis ticks correspond to the temporal segmentation border lines displayed with signal x on Figure 10. Component 2 captures most of the acompaniment, component 7 most of the trombone, and component 9 most of the hiss noise. Summing up the other components leads to extracting the trumpet and clarinet, together with some piano notes.

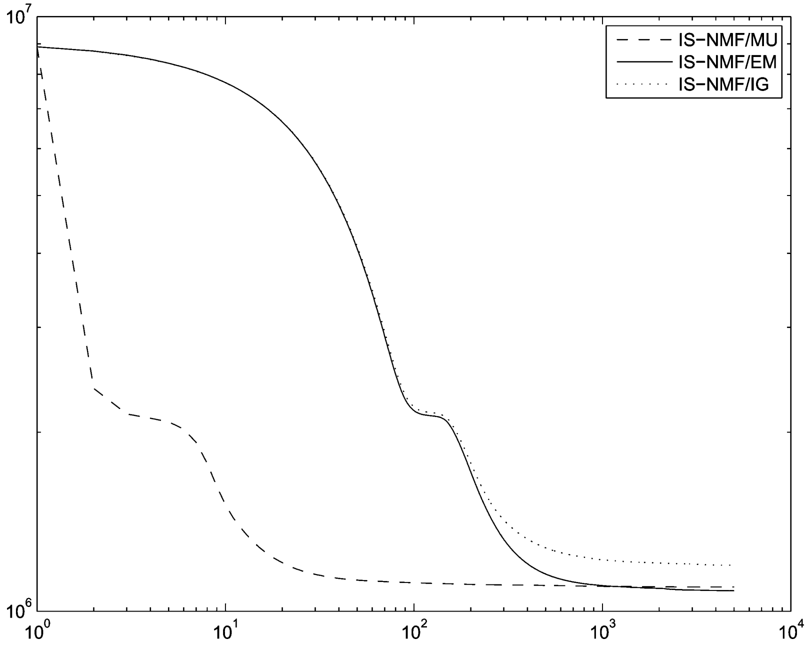

Figure 12: Evolution in log-log scale of the IS cost function along the 5000 iterations of IS-NMF/MU, IS-NMF/EM, and IS-NMF/IG, initialized with the same random values, with $K = 10$.

reconstructions is far better than state-of-the-art KL-NMF of $|\mathbf{X}|$ and ad hoc reconstruction described by equation 2.18.

To conclude this study, we provide online a restored version of the original recording, produced from the IS-NMF/IG decomposition. This is, to our best knowledge, the first use of NMF in an actual audio restoration scenario. The restoration includes denoising (by discarding component 9, which is regarded as noise) and upmixing. A stereo mix is produced by dispatching parts of each component to the left and right channels, hence simulating directions of arrival. As such, we manually created a mix where the components are arranged from 54 degrees left to 54 degrees right, such that the wind instruments (trumpet, clarinet, trombone) are placed left and the stringed instruments (piano, double bass) are placed right. While this stereo mix does render a sensation of spatialization, we emphasize that its quality could undoubtedly be improved with appropriate sound engineering skills.

The originality of our restoration approach lies in the joint noise removal and upmix (as opposed to a suboptimal sequential approach) and the genuine content-based remastering, as opposed to standard techniques based, for example, on phase delays or equalization.

7 Conclusions

We have presented modeling and algorithmic aspects of NMF with the Itakura-Saito divergence. On the modeling side, the following three features of IS-NMF have been demonstrated in this letter:

- IS-NMF is underlaid by a statistical model of superimposed gaussian components.
- This model is relevant to the representation of audio signals.
- This model can accommodate regularization constraints through Bayesian approaches.

On the algorithmic side, we have proposed a novel type of NMF algorithm, IS-NMF/EM, derived from SAGE, a variant of the EM algorithm. The convergence of this algorithm to a stationary point of the cost function $D_{IS}(\mathbf{V} \mid \mathbf{WH})$ is guaranteed by EM. This new algorithm was compared to an existing algorithm, IS-NMF/MU, whose convergence has not been proved, though it has been observed in practice. This letter also reports an experimental comparative study of the standard EUC-NMF and KL-NMF algorithms, together with the two described IS-NMF algorithms, applied to a given data set (a short piano sequence), with various random initializations and model orders. Such a furnished experimental study was, to our best knowledge, not yet available. This letter also reports a proof of concept of the use of IS-NMF for audio restoration, with an actual example. Finally, we believe we have shed light on the statistical implications of NMF with all of three cost functions.

We have shown how smoothness constraints on **W** and **H** can easily be handled in a Bayesian setting with IS-NMF. As such, we have shown how Markov chains' prior structures can improve both the auditory quality of the component reconstructions and the interpretability of the basis elements. The Bayesian setting opens doors to even more elaborate prior structures that can better fit the specificities of data. For music signals, we believe that two promising lines of research lay in (1) the use of switching state models for the rows of **H** that explicitly model the possibility for h_{kn} to be strictly zero with a certain prior probability (and time persistency could be favored by modeling the state sequence with a discrete Markov chain) and (2) the use of models that explicitly take into account the pitched structure of some of the columns of **W** and where the fundamental frequency could act as a model parameter. These models fit into the problem of object-based representation of sound, an active area of research in the music information retrieval and auditory scene analysis communities.

In section 4 we compared the factorization results of a short piano power spectrogram, obtained from three cost functions, given a common algorithmic structure: standard multiplicative updates. The experiments illustrate the slow convergence of this type of algorithm, which has has already been pointed out in other work (Cichocki, Amari et al., 2006; Berry et al., 2007;

Lin, 2007). If the proposed IS-NMF/EM does not improve on this issue, its strength is, however, to offer enough flexibility to accommodate Bayesian approaches. We believe we have made our point that the IS cost is well suited to the factorization of audio power spectrograms (i.e., independent of the type of algorithm used); future work will address the development of faster IS-NMF algorithms. Following developments for other cost functions, we intend to investigate projected gradient techniques (Lin, 2007), exponentiated gradient descent and generalizations (Cichocki, Amari et al., 2006), quasi-Newton second-order methods (Zdunek & Cichocki, 2007), and multilayered approaches (Cichocki & Zdunek , 2006).

Key issues that still need to be resolved in NMF concern identifiability and order selection. A related issue is the investigation into the presence of local minima in cost functions and ways to avoid them. In that matter, Markov chain Monte Carlo (MCMC) sampling techniques could be used as a diagnostic tool to better understand the topography of the criteria to minimize. While it is not clear whether these techniques can be applied to EUC-NMF or KL-NMF, they can readily be applied to IS-NMF, using its underlying gaussian composite structure the same way that IS-NMF/EM does. As to the avoidance of local minima, techniques inherited from simulated annealing could be applied with IS-NMF in either MCMC or EM inference.

Regarding order selection, usual criteria such as the Bayesian information criterion or Akaike's criterion (see, e.g., Stoica & Selén, 2004) cannot be directly applied to IS-NMF because the number of parameters ($F\ K + K\ N$) is not constant with regard to the number of observations N. This feature breaks the validity of the assumptions in which these criteria have been designed. As such, a final promising line of research concerns the design of methods characterizing $p(\mathbf{V} \mid \mathbf{W})$ instead of $p(\mathbf{V} \mid \mathbf{W}, \mathbf{H})$, treating $\mathbf{H}$ as a latent variable, as in independent component analysis (MacKay, 1996; Lewicki & Sejnowski, 2000). Besides allowing for model order selection, such approaches should lead to more reliable estimation of the basis $\mathbf{W}$.

Appendix A: Standard Distributions

Proper complex gaussian	$\mathcal{N}_c(\mathbf{x}\|\boldsymbol{\mu}, \quad) = \|\pi \quad\|^{-1} \exp -(\mathbf{x}-\boldsymbol{\mu})^H \quad ^{-1}(\mathbf{x}-\boldsymbol{\mu})$
Poisson	$\mathcal{P}(x\|\lambda) = \exp(-\lambda)\,\frac{\lambda^x}{x!}$
Gamma	$\mathcal{G}(u \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, u^{\alpha-1} \exp(-\beta\, u),\ u \geq 0$
Inverse-gamma	$\mathcal{IG}(u \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, u^{-(\alpha+1)} \exp(-\frac{\beta}{u}),\ u \geq 0$

The inverse-gamma distribution is the distribution of $1/X$ when X is gamma distributed.

Appendix B: Derivations of the SAGE Algorithm

In this appendix, we detail the derivations leading to algorithm 2. The functions involved in the definition of $Q_k^{ML}(\ _k \mid \ ')$, given by equation 3.2, can be derived as follows. For the hidden data minus log likelihood,

$$-\log p(\mathbf{C}_k \mid \quad _k) = -\sum_{n=1}^{N}\sum_{f=1}^{F}\log \mathcal{N}_c(c_{k,fn} \mid 0, h_{kn}\, w_{fk}) \tag{B.1}$$

$$\overset{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} \log\left(w_{fk}\, h_{kn}\right) + \frac{|c_{k,fn}|^2}{w_{fk}\, h_{kn}}. \tag{B.2}$$

Then the hidden-data posterior is obtained through Wiener filtering, yielding

$$p(\mathbf{C}_k \mid \mathbf{X}, \quad) = \prod_{n=1}^{N}\prod_{f=1}^{F} \mathcal{N}_c\left(c_{k,fn} \mid \mu_{k,fn}^{post}, \lambda_{k,fn}^{post}\right), \tag{B.3}$$

with $\mu_{k,fn}^{post}$ and $\lambda_{k,fn}^{post}$ given by equations 3.3 and 3.4. The E-step is performed by taking the expectation of equation B.2 with regard to the hidden-data posterior, leading to

$$Q_k^{ML}(\quad _k \mid \quad ') \overset{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} \log\left(w_{fk}\, h_{kn}\right) + \frac{\left|\mu_{k,fn}^{post\,\prime}\right|^2 + \lambda_{k,fn}^{post\,\prime}}{w_{fk}\, h_{kn}} \tag{B.4}$$

$$\overset{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} d_{IS}\left(\left|\mu_{k,fn}^{post\,\prime}\right|^2 + \lambda_{k,fn}^{post\,\prime} \mid w_{fk}\, h_{kn}\right). \tag{B.5}$$

The M-step thus amounts to minimizing $D_{IS}(\mathbf{V}_k' \mid \mathbf{w}_k\, h_k)$ with regard to $\mathbf{w}_k \geq 0$ and $h_k \geq 0$, as stated in section 3.2.

Acknowledgments

We acknowledge Roland Badeau, Olivier Cappé, Jean-François Cardoso, Maurice Charbit, and Alexey Ozerov for discussions related to this work. Many thanks to Gaël Richard for comments and suggestions about this manuscript and to Simon Godsill for helping us with the Louis Armstrong original data.

References

Abdallah, S. A., & Plumbley, M. D. (2004). Polyphonic transcription by nonnegative sparse coding of power spectra. In *5th International Symposium of Music Information Retrieval (ISMIR'04)* (pp. 318–325). Vienna: Austrian Computer Society.

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

Benaroya, L., Gribonval, R., & Bimbot, F. (2003). 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)* (pp. 613–616). Los Alamitos, CA: IEEE.

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

Bertin, N., Badeau, R., & Richard, G. (2007). Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*. Los Alamitos, CA: IEEE.

Chen, Z., Cichocki, A., & Rutkowski, T. M. (2006). Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer's disease. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*. Los Alamitos, CA: IEEE.

Cichocki, A., Amari, S.-I., Zdunek, R., Kompass, R., Hori, G., & He, Z. (2006). Extended SMART algorithms for non-negative matrix factorization. In *Proc. of the International Conference on Artificial Intelligence and Soft Computing (ICAISC'06)* (pp. 548–562). Calgary, Canada: ACTA Press.

Cichocki, A., & Zdunek, R. (2006). Multilayer nonnegative matrix factorization. *Electronics Letters, 42*(16), 947–948.

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

Cohen, I., & Gannot, S. (2007). Spectral enhancement methods. In M. M. Sondhi, J. Benesty, & Y. Huang (Eds.), *Springer handbook of speech processing*. New York: Springer.

Dhillon, I. S., & Sra, S. (2005). Generalized nonnegative matrix approximations with Bregman divergences. In M. I. Jordan, Y. Le Cun, & S. A. Solla (Eds.), *Advances in neural information processing systems, 19*. Cambridge, MA: MIT Press.

Drakakis, K., Rickard, S., de Fréin, R., & Cichocki, A. (2008). Analysis of financial data using non-negative matrix factorization. *International Mathematical Forum, 3*, 1853–1870.

Eguchi, S., & Kano, Y. (2001). *Robustifying maximum likelihood estimation*. (Research Memo 802). Tokyo: Institute of Statistical Mathematics. Available online at http://www.ism.ac.jp/~eguchi/pdf/Robustify_MLE.pdf.

Feder, M., & Weinstein, E. (1988). Parameter estimation of superimposed signals using the EM algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 36*, 477–489.

Fessler, J. A., & Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing, 42*, 2664–2677.

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

Itakura, F., & Saito, S. (1968). Analysis synthesis telephony based on the maximum likelihood method. In *Proc. 6th of the International Congress on Acoustics* (pp. C–17–C–20). Los Alamitos, CA: IEEE.

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

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

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural and information processing systems, 13* (pp. 556–562). Cambridge, MA: MIT Press.

Lewicki, M. S., & Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation, 12*, 337–365.

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

MacKay, D. (1996). *Maximum likelihood and covariant algorithms for independent component analysis*. Unpublished manuscript. Available online at http://www.inference.phy.cam.ac.uk/mackay/ica.pdf.

Ozerov, A., Philippe, P., Bimbot, F., & Gribonval, R. (2007). 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*, 1564–1578.

Plumbley, M. D., Abdallah, S. A., Blumensath, T., & Davies, M. E. (2006). Sparse representations of polyphonic music. *Signal Processing, 86*(3), 417–431.

Shashanka, M., Raj, B., & Smaragdis, P. (2008a). Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience, 2008*.

Shashanka, M., Raj, B., & Smaragdis, P. (2008b). Sparse overcomplete latent variable decomposition of counts data. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems, 20* (pp. 1313–1320). Cambridge, MA: MIT Press.

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

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

Stoica, P., & Selén, Y. (2004). Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine, 21*, 36–47.

Vincent, E., Bertin, N., & Badeau, R. (2007). Two nonnegative matrix factorization methods for polyphonic pitch transcription. In *Proc. of the Music Information Retrieval Evaluation eXchange (MIREX)*. Available online at http://www.music-ir.org.

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

Virtanen, T., Cemgil, A. T., & Godsill, S. (2008). Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)* (pp. 1825–1828). Los Alamitos, CA: IEEE.

Young, S. S., Fogel, P., & Hawkins, D. (2006). Clustering scotch whiskies using nonnegative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association, 14*, 11–13.

Zdunek, R., & Cichocki, A. (2007). Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing, 87*, 1904–1916.

Received April 29, 2008; accepted July 3, 2008.