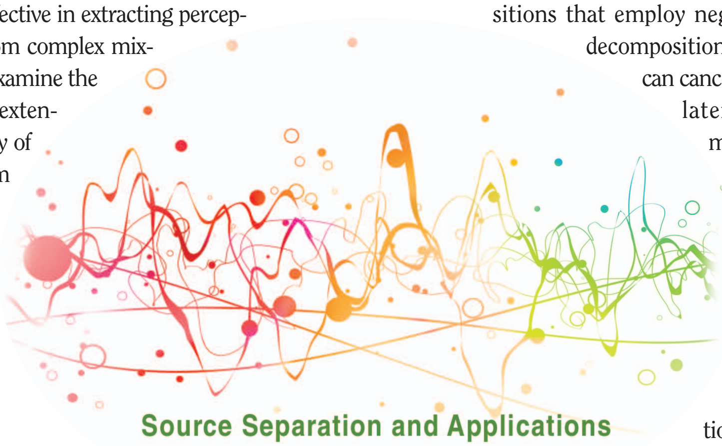
Paris Smaragdis, Cédric Févotte, Gautham J. Mysore, Nasser Mohammadiha, and Matthew Hoffman

Static and Dynamic Source Separation Using Nonnegative Factorizations

A unifed view

Source separation models that make use of nonnegativity in their parameters have been gaining increasing popularity in the last few years, spawning a significant number of publications on the topic. Although these techniques are conceptually similar to other matrix decompositions, they are surprisingly more effective in extracting perceptually meaningful sources from complex mixtures. In this article, we will examine the various methodologies and extensions that make up this family of approaches and present them under a unified framework. We will begin with a short description of the basic concepts and in the subsequent sections we will delve in more details and explore some of the latest extensions.



USING NONNEGATIVE FACTORIZATION MODELS FOR SEPARATION

The basic model we will use to get started is a bilinear factorization of a nonnegative input $\mathbf{V}$ into two nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$, i.e., $\mathbf{V} \approx \mathbf{WH}$, where both of the two factor matrices can be of lower rank than $\mathbf{V}$. This is known as the *nonnegative matrix factorization* (*NMF*) [1] model, and it is conceptually similar to other well-known matrix factorizations such as principal component analysis, independent component analysis, sparse linear models, or even vector quantization, which can all be expressed using the same equation [2]. What makes this model particularly interesting is the constraint that the matrices $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{H}$ are all nonnegative. This constraint ensures that the vectors making up the two factor matrices $\mathbf{W}$ and $\mathbf{H}$ can be interpreted as constructive building blocks of the input. Such an interpretation often does not apply to decompositions that employ negative-valued entries; in such decompositions, the elements of $\mathbf{W}$ and $\mathbf{H}$ can cancel each other out, obscuring the latent components' perceptual meaningfulness [1]. When NMF is applied to data that was generated by mixing a number of nonnegative sources, the components NMF discovers often correspond remarkably well to those sources, and the decomposition is able to separate out the contributions of each source to the data. Since NMF can operate even without any prior information about the nature of the sources in the data, it is particularly well suited to unsupervised or blind source separation problems. Some examples of interpretable components discovered by NMF are presented in Figure 1.

Sometimes it is more natural to represent complex sources using a linear combination of multiple latent components that collectively make up source dictionaries. In this case, we need one more level of hierarchy to group these components in terms of sources. Although in some cases this grouping could be obvious or analytically tractable, it is in principle not easy to compute. One can overcome this problem by using nonnegative factorization models in a supervised manner and explicitly providing cues

Digital Object Identifier 10.1109/MSP.2013.2297715
Date of publication: 7 April 2014

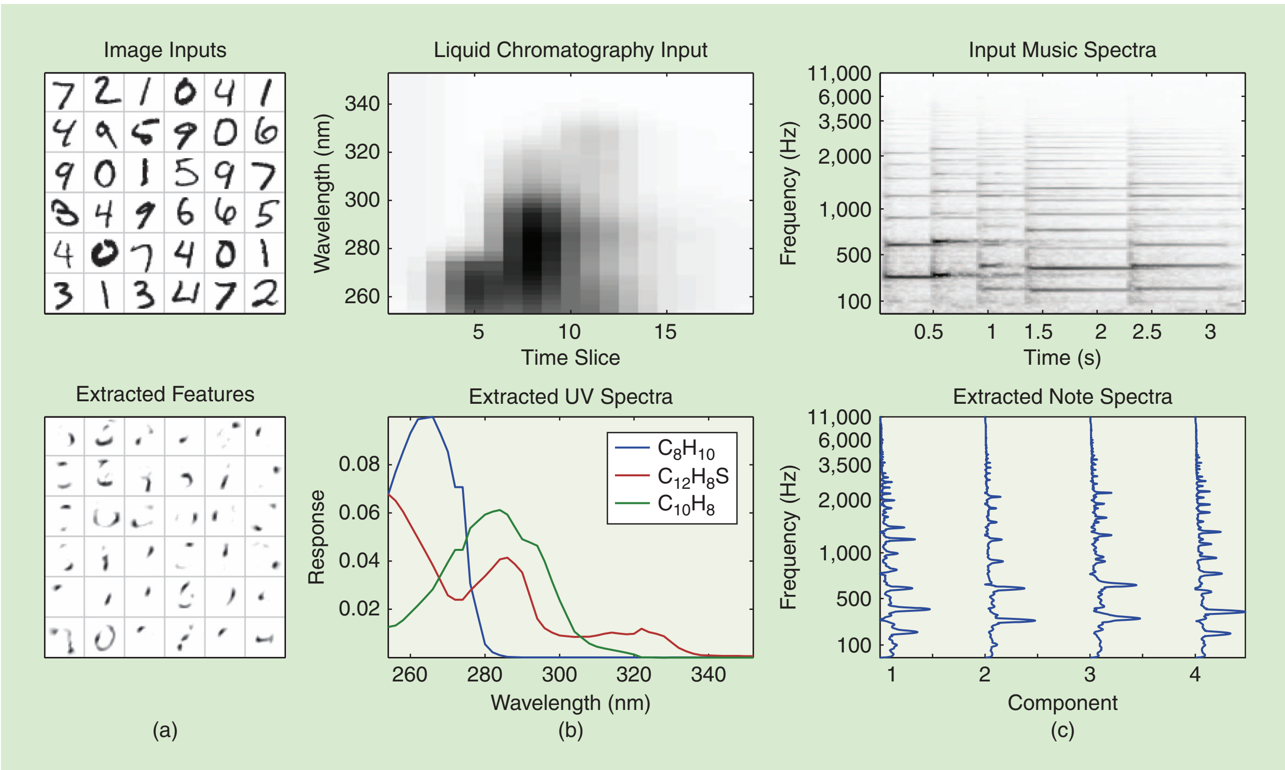
 1053-5888/14/$31.00©2014IEEE

to the nature of the sources. This involves learning a dictionary for each target source by using the above model on clean training data that presents that source in isolation, and then identifying where in a mixture the dictionary elements associated with each source lie. If our data is not nonnegative already, to employ a nonnegative factorization we need to transform our inputs to an additive (or approximately additive) nonnegative representation. For many kinds of time series, such a domain can be a time-frequency localized energy measure computed via a harmonic decomposition such as the Gabor transform, or a wavelet decomposition. Since most natural signals tend to be sparse in the magnitude or power, by using these transforms we can often guarantee with high probability that the transform of the sum of two sources will be equal or approximately equal to the sum of the transforms of the two sources separately, which can satisfy the additivity constraint. As we show later, depending on the exact NMF model and the representation used, the additivity assumption can be one that is either weak or strong.

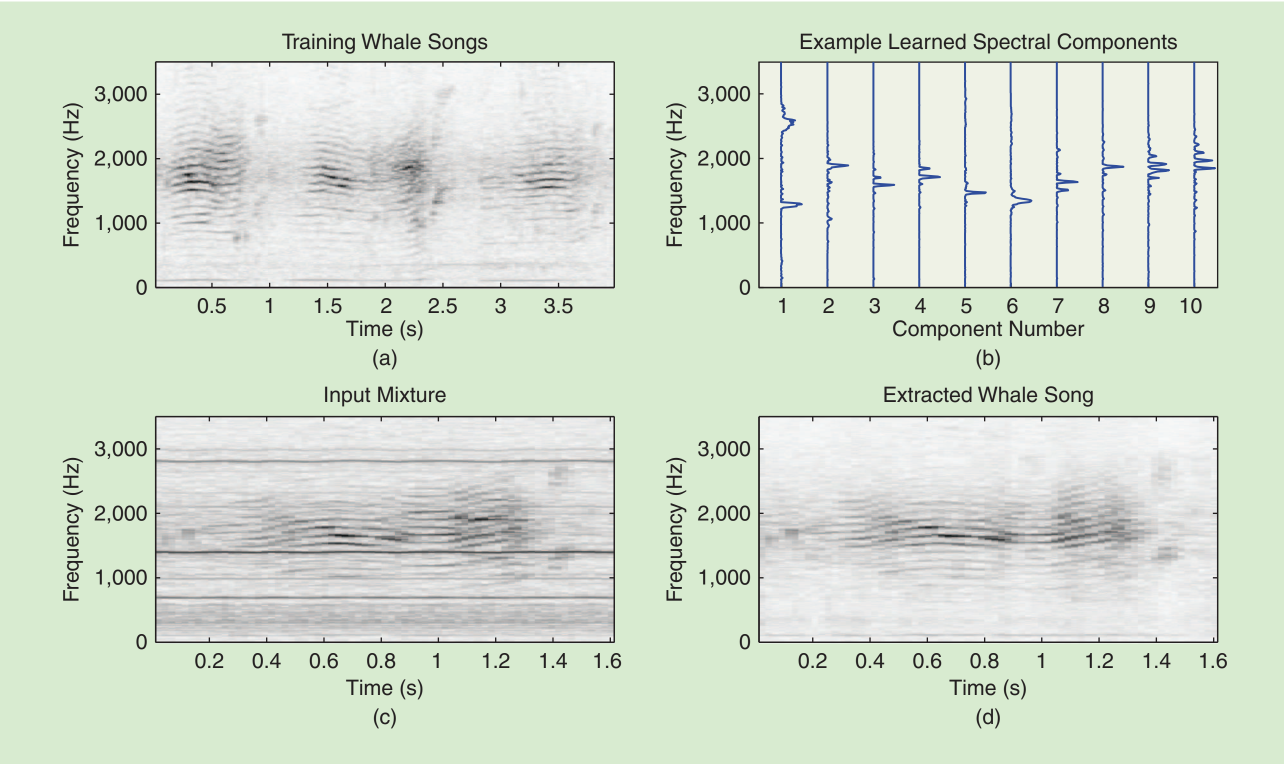
ALTHOUGH THESE TECHNIQUES ARE CONCEPTUALLY SIMILAR TO OTHER MATRIX DECOMPOSITIONS, THEY ARE SURPRISINGLY MORE EFFECTIVE IN EXTRACTING PERCEPTUALLY MEANINGFUL SOURCES FROM COMPLEX MIXTURES.

To demonstrate the separation process with a tangible example, let us look at a hydrophone mixture containing a whale song (target source) and sea clutter (background sources). We represent this mixture using a magnitude short-time Fourier transform (STFT), which is shown in Figure 2(c). To learn a target source dictionary we use a clean recording of whale songs [Figure 2(a)]. This is done by analyzing the matrix containing the STFT representation using any of the models that we detail in the remainder of this article. A learned dictionary is shown in Figure 2(b), and as one can see its elements represent salient spectral features that comprise the whale song recording. We can repeat this process for the sea clutter source to get components that describe it too. In practice, a few seconds of training data is usually enough to learn an adequate model of a source, although this can vary depending on the domain and source characteristics we are dealing with. The number of components per dictionary determines how accurately we want to model the sources, with more components giving us more expressive power but at the cost of making a dictionary so rich that it could describe other sources as well.

Given the approximate additivity assumption and a representative set training data, we can now hypothesize that the mixture recording will be explained by a linear combination of the elements in the source dictionaries, i.e., that $X \approx [W_1, W_2]H$ will



[FIG1] Extracted NMF components from various domains. (a) The analysis of handwritten digital data results in parts of penstrokes, (b) the analysis of chemometric data results in the spectral profiles of the three constituent components (oxylene, napthalene, dibenzothiophene), and (c) the analysis of music spectrograms results in spectra of musical notes.



[FIG2] Extracting a target source from a hydrophone ocean mixture using a nonnegative dictionary. The training data in (a) are isolated whale songs used to learn the dictionary shown in (b). Not shown are the equivalent plots for sea clutter sounds. These dictionaries are then used to extract their respective sources from a mixture that includes them, shown in (c). The extracted whale song is shown in (d).

approximately hold, where $\mathbf{X}$ contains the magnitude STFT of the mixture and $\mathbf{W}_1$ and $\mathbf{W}_2$ are the learned left factors from the training data of the two sounds. We thus only need to compute the matrix $\mathbf{H}$. Given the ability to compute the full NMF model, the estimation of the $\mathbf{H}$ matrix can be easily obtained by fixing $[\mathbf{W}_1, \mathbf{W}_2]$ and only updating the estimate for $\mathbf{H}$. Once this is computed we can reconstruct the mixture using only the dictionary of one source at a time, which will produce in a time-frequency representation of the two sources separately, which can then be inverted back to the time domain. The only assumption that needs to hold at this point is that the two source dictionaries are sufficiently different from each other so that they do not model the same elements in the mixture. Although there is no easy way of quantifying the required degree of dissimilarity in real-world examples, this is a process that works even in cases where the sources are very similar (e.g., two speakers of the same gender), and by incorporating the ideas in the remainder of this article we can even separate sources that share identical dictionaries by making use of their temporal statistics. In this particular case, the dictionaries that characterize the two sources have minimal similarities and produce a very clean separation. The result of extracting the whale song from the hydrophone mixture is shown in Figure 2. The details of this process and its generalization in the case where we might not have dictionaries for all the sources is described in [3].

This basic approach of supervised separation has spawned much subsequent research using varying approaches and methodologies, often seemingly incompatible with each other. In the following sections we will take a closer look at the details of various formulations of nonnegative factorization models, and will show a unified progression of techniques that spans from the simple static models (such as the ones shown above) to more complex dynamic approaches that incorporate more temporal information and can produce higher-quality results. We will predominantly focus on the statistical interpretation (and variation) within NMF algorithms and then we will show how these can be extended to two kinds of useful temporal models: continuous state and discrete state models, which in turn can take advantage of temporal information to improve the performance of source separation tasks.

STATIC MODELS

A PROBABILISTIC VIEW OF NMF

Traditionally NMF is applied by solving the optimization problem defined by

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}\,|\,\mathbf{WH}) \quad \text{s.t} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{1}$$

where $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{H}$ are nonnegative matrices of size $F \times T$, $F \times K$, and $K \times T$, respectively. The notation $\mathbf{M} \geq 0$ denotes element-wise nonnegativity of $\mathbf{M}$ (and not semidefinite positivity) and $D(\mathbf{V}\,|\,\mathbf{WH})$ is a separable measure of fit such that

$$D(\mathbf{V}\,|\,\mathbf{WH}) = \sum_{t=1}^{T} D(\mathbf{v}_t\,|\,\mathbf{Wh}_t). \tag{2}$$

$D(\mathbf{x}\,|\,\mathbf{y})$ is a divergence between vectors $\mathbf{x}$ and $\mathbf{y}$, i.e., a nonnegative function of $\mathbf{y} \in \mathbb{R}_+^F$ given $\mathbf{x} \in \mathbb{R}_+^F$, with a single minimum (zero) for $\mathbf{x} = \mathbf{y}$. For convenience we will use the same notation $D(\cdot|\cdot)$ to denote the divergence between vectors or matrices, with the convention that in the matrix case the divergences between columns simply add up as in (2). Common divergences used in NMF include the squared Euclidean distance (see [46]), variants of the Kullback–Leibler (KL) divergence [1], and the Itakura–Saito (IS) divergence [4]. More general families of divergences considered for NMF include alpha-beta [5] and Bregman divergences [6]. A comprehensive review of divergences and algorithms used for NMF can be found in [7].

In many cases divergences are likelihoods in disguise (and are as such sometimes referred to as *pseudolikelihoods*) in the sense that they underlie a probabilistic generative model of the data. The correspondence is such that there exists a probability density function (pdf) $p(\mathbf{V}\,|\,\mathbf{W},\mathbf{H})$ that satisfies

$$-\log p(\mathbf{V}\,|\,\mathbf{WH}) = a\,D(\mathbf{V}\,|\,\mathbf{WH}) + b, \tag{3}$$

where a and b are constants with respect to $\mathbf{WH}$. Some examples of correspondences are given in Table 1. Note that this correspondence does not automatically imply a coherent generative model for nonnegative real-valued data; e.g., although the generalized KL divergence is a valid measure of fit on the whole positive orthant, the corresponding Poisson likelihood is only a true likelihood on the nonnegative integers, and in the large-variance setting the additive Gaussian model could generate negative data. However, these theoretical issues can usually be resolved; see, e.g., [8].

In this article we focus on two probabilistic NMF models that have been widely used in source separation: probabilistic latent component analysis (PLCA), which is closely related to NMF with the KL divergence [9], and the Gaussian composite model (GCM), which is closely related to NMF with the IS divergence [4]. A common feature of these models, shared by the models in Table 1 as well, is that the conditional expectation of $\mathbf{V}$ is $\mathbf{WH}$ (i.e., $\mathbb{E}[\mathbf{V}\,|\,\mathbf{WH}] = \mathbf{WH}$), and that the data points are conditionally independent given $\mathbf{WH}$ [i.e., $p(\mathbf{V}\,|\,\mathbf{WH}) = \prod_t p(\mathbf{v}_t\,|\,\mathbf{Wh}_t)$]. These simple factorization models are "static" in the sense that data points (columns of $\mathbf{V}$) could be exchanged without any effect on the estimates other than a permutation of $\mathbf{H}$. Dynamic, nonexchangeable models will be introduced later in the article using temporal priors on $\mathbf{H}$.

WHEN NMF IS APPLIED TO DATA THAT WAS GENERATED BY MIXING A NUMBER OF NONNEGATIVE SOURCES, THE DISCOVERED COMPONENTS OFTEN CORRESPOND REMARKABLY WELL TO THOSE SOURCES.

PROBABILISTIC LATENT COMPONENT ANALYSIS

PLCA is an extension of probabilistic latent semantic indexing (PLSI) for signal processing applications [9]. PLSI is a method for text analysis based on word counts from documents [10]. In PLCA, the input matrix $\mathbf{V}$ is a magnitude spectrogram $v_{ft} = |x_{ft}|$, where x_{ft} is the complex-valued STFT of some time-domain data. PLCA interprets the entries of each column $\mathbf{v}_t$ of $\mathbf{V}$ as a sort of histogram of independent identically distributed (i.i.d.) frequency "quanta" $f \in \{1,\dots,F\}$ in each time frame t. The data distribution in PLCA is therefore

$$\mathbf{v}_t \sim \mathrm{Mult}(\mathbf{v}_t\,|\,\|\mathbf{v}_t\|_1, \hat{\mathbf{v}}_t), \tag{4}$$

where $\|\mathbf{v}\|_1 = \sum_f |v_f|$ is the ℓ_1 norm, $\hat{\mathbf{v}}_t = \mathbf{Wh}_t$, and $\mathrm{Mult}(N, \mathbf{p})$ denotes the multinomial distribution. In PLCA it is imposed that $\|\mathbf{w}_k\|_1 = \|\mathbf{h}_t\|_1 = 1$, which in turn implies that $\|\hat{\mathbf{v}}_t\|_1 = 1$. A draw from $\mathrm{Mult}(N, \mathbf{p})$ returns an integer-valued vector of dimension F whose entries sum to N. The fth entry of this vector corresponds to the number of times event f was sampled in N independent draws from the discrete distribution defined by $\mathbf{p}$. Although usual inputs in source separation problems are not integer valued, the negative log-likelihood of the data and parameters in PLCA provides a valid divergence for nonnegative real-valued data. Specifically, under (4) and introducing the normalized data $\bar{\mathbf{v}}_t = \mathbf{v}_t/\|\mathbf{v}_t\|_1$, the negative log-likelihood is given by

$$-\log p(\mathbf{V}\,|\,\hat{\mathbf{V}}) = \sum_t \|\mathbf{v}_t\|_1 D_{\mathrm{KL}}(\bar{\mathbf{v}}_t\,|\,\hat{\mathbf{v}}_t) + \mathrm{cst}, \tag{5}$$

where "cst" denotes terms constant with respect to $\hat{\mathbf{V}}$ and $D_{\mathrm{KL}}(\mathbf{x}\,|\,\mathbf{y}) = \sum_f x_f \log(x_f/y_f)$ is the KL divergence between discrete distributions. As such, PLCA essentially minimizes a weighted KL divergence between the normalized input and its factorized approximation, where every data point is given a weight equal to its sum.

IS-NMF AND THE GAUSSIAN COMPOSITE MODEL

Underlying IS-NMF is a multiplicative noise model of the form $v_{fn} = \hat{v}_{fn} \,.\, \epsilon_{fn}$, where ϵ_{fn} has a Gamma distribution with

[TABLE 1] COMMON DIVERGENCES AND THEIR CORRESPONDING PROBABILISTIC GENERATIVE MODELS. WE DEFINE $\hat{\mathbf{v}}_t = \mathbf{Wh}_t$, WHOSE COEFFICIENTS ARE DENOTED $\hat{v}_{ft}$. ALL THREE MODELS VERIFY $\mathbf{E}[\mathbf{v}_t\,|\,\hat{\mathbf{v}}_t] = \hat{\mathbf{v}}_t$.

DIVERGENCE $D(\mathbf{v}_t\,|\,\hat{\mathbf{v}}_t)$	LATENT GENERATIVE MODEL $p(\mathbf{v}_t\,|\,\hat{\mathbf{v}}_t)$
SQUARED EUCLIDEAN DISTANCE $\frac{1}{2\sigma^2}\sum_f (v_{ft} - \hat{v}_{ft})^2$	ADDITIVE GAUSSIAN $\prod_f \mathrm{N}(v_{ft}\,|\,\hat{v}_{ft}, \sigma^2)$
GENERALIZED KL DIVERGENCE $\sum_f (v_{ft} \log \frac{v_{ft}}{\hat{v}_{ft}} - v_{ft} + \hat{v}_{ft})$	POISSON $\prod_f \mathrm{P}(v_{ft}\,|\,\hat{v}_{ft})$
IS DIVERGENCE $\sum_f \left(\frac{v_{ft}}{\hat{v}_{ft}} - \log\frac{v_{ft}}{\hat{v}_{ft}} - 1\right)$	MULTIPLICATIVE GAMMA $\prod_f G(v_{ft}\,|\,\alpha, \alpha/\hat{v}_{ft})$

expectation one. The resulting data distribution is given in Table 1 and the negative log-likelihood is such that

$$-\log p(\mathbf{V} \mid \hat{\mathbf{V}}) = \alpha D_{\mathrm{IS}}(\mathbf{V} \mid \hat{\mathbf{V}}) + \mathrm{cst}, \tag{6}$$

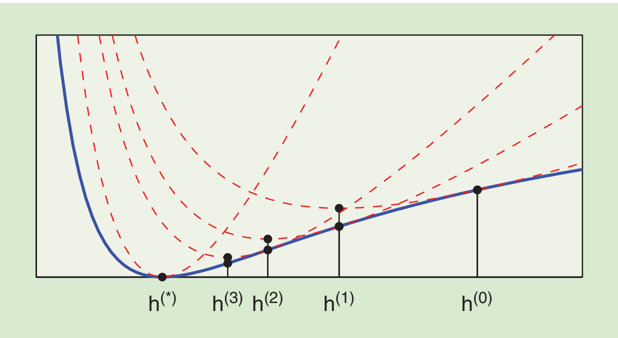
where $D_{\mathrm{IS}}(\cdot \mid \cdot)$ is the IS divergence defined in Table 1.

When $\alpha = 1$, i.e., when the multiplicative noise has an exponential distribution, the multiplicative noise model can be related to a generative model of real- or complex-valued data coined *Gaussian composite model* (*GCM*) [4]. The model is in particular a valid probabilistic model of STFTs. Let x_{ft} be the complex-valued STFT of some time-domain signal. The GCM is defined by $x_{ft} = \sum_k c_{fkt}$ and $c_{fkt} \sim N_c(0, w_{fk} h_{kt})$, where $N_c(0, \lambda)$ refers to the circular complex Gaussian distribution with zero mean. A random variable has distribution $N_c(0, \lambda)$ if its real and imaginary parts are independent centered Gaussian variables with variance $\lambda/2$. In other words, the GCM models the STFT as a sum of uncorrelated centered Gaussian components structured through their variance. The variance of the kth component is characterized by the spectral pattern $\mathbf{w}_k$, amplitude-modulated in time by the coefficients $\{h_{kt}\}_t$. The centered assumption reflects an equivalent assumption in the time domain, which holds for many signals (in particular audio signals). The latent components c_{fkt} can trivially be marginalized from the generative model, yielding $x_{ft} \sim N_c\left(0, \sum_k w_{fk} h_{kt}\right)$. It follows that the power spectrogram $v_{ft} = |x_{ft}|^2$ of x_{ft} is exponentially distributed with mean $\hat{v}_{ft} = \sum_k w_{fk} h_{kt}$, and can thus be written as a special case of the multiplicative Gamma model given in Table 1 with $\alpha = 1$. Under this model, minimum mean squares estimate (MMSE) of the components can be obtained by Wiener filtering and given by $\hat{c}_{fkt} = [(w_{fk} h_{kt})/\hat{v}_{ft}] x_{ft}$.

A MORE FLEXIBLE APPROACH FOR MODELING TEMPORAL STATISTICS IS TO IMPOSE CONSTRAINTS ON THE MODEL ACTIVATIONS.

WHICH MODEL TO USE?

An important feature of the GCM is that the phase of the original complex-valued data is preserved in the generative model (though it is modeled in an uninformative way, owing to the circular assumption) rather than discarded, as in PLCA. Additionally, the additivity assumption holds strongly in the original STFT domain. The IS divergence turns out to be a scale-invariant measure, i.e., $d_{\mathrm{IS}}(\lambda x \mid \lambda y) = d_{\mathrm{IS}}(x \mid y)$, where x, y, and λ are positive scalars. This makes it well suited to audio spectrograms and their widely varying ranges of magnitudes; a more detailed discussion is in [4]. In contrast, PLCA will rely more heavily on data vectors with large norms, as can been seen from the divergence expression in (5). Whether this is a desirable property or not depends on the data and specific task. A downside of the IS divergence with respect to the weighted KL divergence of PLCA is its lack of convexity with respect to its second argument, which leads more often to local solutions in practice, as explained in the next section. PLCA and IS-NMF were benchmarked in [11] for speech separation and audio interpolation tasks. However, a consensus did not clearly emerge from the experiments as to which method is best, and the conclusions were often data or task dependent.



[FIG3] An illustration of the MM principle on a unidimensional problem. Given a current estimate of $\mathbf{W}$, the blue curve acts as the objective function $C(\mathbf{H}) = D(\mathbf{V} \mid \mathbf{WH})$ to be minimized with respect to $\mathbf{H}$. The MM approach relies on the iterative minimization of tight upper bounds (dashed red curves). The algorithm is initialized at $\mathbf{H}^{(0)}$, at which the first upper bound is minimized during the first iteration to yield $\mathbf{H}^{(1)}$, and so on until convergence.

ESTIMATION

We now discuss estimation in PLCA and IS-NMF, i.e., the optimization of the objective functions (5) and (6) with respect to $\mathbf{W}$ and $\mathbf{H}$. Like virtually all NMF algorithms, PLCA and IS-NMF rely on a block-coordinate descent structure that alternates between updating $\mathbf{W}$ holding $\mathbf{H}$ fixed and updating $\mathbf{H}$ holding $\mathbf{W}$ fixed. It is easy to see that the updates of $\mathbf{W}$ and $\mathbf{H}$ are essentially the same by transposition $(\mathbf{V} \approx \mathbf{WH} \Leftrightarrow \mathbf{V}^T \approx \mathbf{H}^T \mathbf{W}^T)$. Each update can be carried out by majorization-minimization (MM) [12]. MM consists in upper bounding the objective function with an auxiliary function that is tight at the current estimate and that can be minimized in closed form. The principle of MM is illustrated in Figure 3. Details of the algorithms can be found in [9] for PLCA and in [13] for IS-NMF. The resulting updates are given in Table 2. Their multiplicative structure automatically ensures the nonnegativity of the updates given positive initialization.

It should be pointed out that in every NMF problem the objective function $D(\mathbf{V} \mid \mathbf{WH})$ is not jointly convex with respect to $\mathbf{W}$ and $\mathbf{H}$. When the divergence $D(\mathbf{x} \mid \mathbf{y})$ is convex with respect to its second argument $\mathbf{y}$, like in PLCA, the problem is at least convex with respect to $\mathbf{H}$ given $\mathbf{W}$ and vice versa. However it is never convex with respect to both. This means that the block-coordinate approach may converge to local solutions that will depend on initialization. Some recent work (e.g., [14] and [15]) has explored alternate estimation algorithms that avoid formulating NMF as a nonconvex optimization and thereby sidestep the local-optima problem. The guarantees associated with these algorithms are dependent on separability and/or sparsity assumptions that may be more appropriate for extremely high-dimensional data like document word counts than for moderately high-dimensional data like audio spectra. However, as shown in [16], separability is not necessary for uniqueness in NMF, and such a

[TABLE 2] PLCA AND IS-NMF FOR THE GCM SUMMARIZED. IN THE UPDATE RULES, $\tilde{w}_{fk}$ AND $\tilde{h}_{kt}$ DENOTE CURRENT PARAMETER VALUES. $\hat{v}_{ft}$ DENOTES THE CURRENT DATA APPROXIMATION, I.E., $\sum_k w_{fk}\tilde{h}_{kt}$ IN THE UPDATE OF $\mathbf{H}$ AND $\sum_k \tilde{w}_{fk} h_{kt}$ IN THE UPDATE OF $\mathbf{W}$.

	PLCA	IS-NMF FOR THE GCM
NONNEGATIVE DATA	$\mathbf{V} = \lvert \mathbf{X} \rvert$	$\mathbf{V} = \lvert \mathbf{X} \rvert^2$
OBJECTIVE FUNCTION	$D(\mathbf{V} \mid \mathbf{WH}) = \sum_t \lVert \mathbf{v}_t \rVert_1 D_{KL}(\bar{\mathbf{v}}_t \mid \hat{\mathbf{v}}_t)$	$D(\mathbf{V} \mid \mathbf{WH}) = D_{IS}(\mathbf{V} \mid \mathbf{WH})$
CONSTRAINTS	$\lVert \mathbf{w}_k \rVert_1 = \lVert \mathbf{h}_t \rVert_1 = 1$	—
LATENT GENERATIVE MODEL	$p(\mathbf{v}_t \mid \hat{\mathbf{v}}_t) = \mathrm{Mult}(\mathbf{v}_t \mid \lVert \mathbf{v}_t \rVert_1, \hat{\mathbf{v}}_t)$	$p(\mathbf{x}_t \mid \hat{\mathbf{v}}_t) = \prod_f \mathbf{N}_c(x_{ft} \mid 0, \hat{v}_{ft})$
UPDATES	$h_{kt} = \dfrac{\tilde{h}_{kt} \sum_f w_{fk}(v_{ft}/\hat{v}_{ft})}{\sum_k \tilde{h}_{kt} \sum_f w_{fk}(v_{ft}/\hat{v}_{ft})}$ $w_{fk} = \dfrac{\tilde{w}_{fk} \sum_t h_{kt}(v_{ft}/\hat{v}_{ft})}{\sum_f \tilde{w}_{fk} \sum_t h_{kt}(v_{ft}/\hat{v}_{ft})}$	$h_{kt} = \tilde{h}_{kt} \dfrac{\sum_f w_{fk}(v_{ft}/\hat{v}_{ft}^2)}{\sum_f w_{fk}(1/\hat{v}_{ft})}$ $w_{fk} = \tilde{w}_{fk} \dfrac{\sum_n h_{kt}(v_{ft}/\hat{v}_{ft}^2)}{\sum_n h_{kt}(1/\hat{v}_{ft})}$

constraint can be too restrictive when using convex formulations. Regardless, for our purposes, the block-coordinate approach is practical and effective on a wide range of problems, despite its lack of theoretical guarantees.

So far we have presented a basic version of NMF in which the data is approximated as $\mathbf{V} \approx \mathbf{WH}$ without any structural priors (aside from nonnegativity) on either $\mathbf{W}$ or $\mathbf{H}$. However, in many cases one is expecting the latent factors to have a certain structure, such as smoothness or sparsity. As such, a large part of the NMF literature has concentrated on penalized variants of NMF, in which penalty functions of either $\mathbf{W}$ or $\mathbf{H}$ are added to the divergence $D(\mathbf{V} \mid \mathbf{WH})$. In our probabilistic setting, this can be viewed as setting prior distributions for the latent factors. In particular, the next section will review temporal priors $p(\mathbf{H})$ that have been used in the literature. In most cases, penalized NMF can be handled with MM, by simply adding the penalty term, or a local majorization of the latter, to the auxiliary function obtained in the static case.

DYNAMIC MODELS

Temporal continuity is one of the most important features of time-series data. Our aim here is to present some of the basic as well as advanced ideas to make use of this information by modeling time dependencies in NMF. These dependencies between consecutive columns of $\mathbf{V}$ can be imposed either on the basis matrix $\mathbf{W}$ or on the activations $\mathbf{H}$. The former case is known as the convolutive NMF [17]–[19]. In these approaches, the repeating patterns within data are represented with multidimensional bases which are not vectors anymore, but functions that can span an arbitrary number of dimensions (e.g., both frequency and time in examples like the previous one). These models can be seen as a deterministic way to model temporal dependencies. Although they are useful in extracting temporal components, they most often result in very structured representations that do not generalize well enough to be successfully employed for source separation. A more flexible approach for modeling temporal statistics is to impose constraints on the model activations. Such methods are very much in line with traditional dynamic models that have been studied extensively in signal processing, and in this section we will turn our attention to these.

Most models considered in the literature are special cases of the general dynamic model given by

$$\mathbf{h}_t \sim p(\mathbf{h}_t \mid \mathbf{h}_{t-1}, \theta), \tag{7}$$

$$\mathbf{v}_t \sim p(\mathbf{v}_t \mid \mathbf{W}\mathbf{h}_t). \tag{8}$$

We assume that (8) defines a probabilistic NMF observation model such that $\mathbb{E}[\mathbf{V} \mid \mathbf{WH}] = \mathbf{WH}$. As such, it may refer to any of the static models discussed in the previous section. Equation (7) introduces temporal dynamics by assuming a Markov structure for the activation coefficients. θ denotes the prior parameters. The aim of this section is to describe the general concepts of dynamic NMF and provide references for specific instantiations related to given probabilistic NMF models (PLCA, IS-NMF, generalized KL-NMF, etc.). Two broad classes of models are discussed next, continuous and discrete models.

CONTINUOUS MODELS

SMOOTH NMF

A straightforward approach to use temporal continuity is to apply some constraints that reduce fluctuations in each individual row of $\mathbf{H}$. This corresponds to assuming that different rows of $\mathbf{H}$ are independent.

In these approaches, the general equation (7) can be written as

$$\mathbf{h}_t \sim \prod_{k=1}^{K} p(h_{kt} \mid h_{k(t-1)}, \theta). \tag{9}$$

A natural choice for $p(h_{kt} \mid h_{k(t-1)}, \theta)$ is a pdf that either takes its mode at $h_{k(t-1)}$ or is such that $\mathbb{E}[h_{kt} \mid h_{k(t-1)}, \theta] = h_{k(t-1)}$. Various papers have dealt with smooth NMF and they typically differ by the choice of observation models and priors (or in nonprobabilistic settings, penalty term) that is used [4], [20]–[27]. Gaussian priors (or equivalently, squared differences) of the form $p(h_{kt} \mid h_{k(t-1)}) = \mathcal{N}(h_{kt} \mid h_{k(t-1)}, \sigma^2)$ are used in [20], [21], and [26]. Nonnegativity-preserving Gamma or inverse-Gamma Markov chains are considered in [4], [23], [25], and [27]–[30] and Markov random fields in [31].

NONNEGATIVE STATE-SPACE MODELS

Smooth NMF does not capture the full extent of frame-to-frame dependencies in its input. In practice we will observe various temporal correlations between adjacent time frames that will be more nuanced than the continuity that smooth NMF implies. In other words, there is correlation both within (smoothness) and between (transitions) the time frames of the coefficients of $\mathbf{H}$. For real-valued time series, this type of structure can be handled with the classical linear dynamical system, using dynamics of the form $\mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \epsilon_t$, where ϵ_t is a centered Gaussian innovation. This model is not natural in the NMF setting because it may not maintain nonnegativity in the activations. However it is possible to design alternative dynamic models that maintain nonnegativity while preserving

$$\mathbb{E}[\mathbf{h}_t \,|\, \mathbf{A}\mathbf{h}_{t-1}] = \mathbf{A}\mathbf{h}_{t-1}. \tag{10}$$

The statistical models considered in the section "Static Models" are good candidates by exchanging $\mathbf{v}_t$ for $\mathbf{h}_t$ and $\hat{\mathbf{v}}_t$ for $\mathbf{h}_{t-1}$. Following that idea, a nonnegative dynamical system (NDS) with multiplicative Gamma innovations was proposed in [32], in conjunction with multiplicative Gamma noise for the observation (IS-NMF model). Note that in the case of the Gaussian linear dynamical system, integration of the activation coefficients from the joint likelihood $p(\mathbf{V}, \mathbf{H} \,|\, \mathbf{W})$ is feasible using the Kalman filter. Such computations are unfortunately intractable with NDS, and a MAP approach based on an MM algorithm is pursued in [32].

Dynamic filtering of the activation coefficients in the PLCA model has also been considered [33], [34], where the proposed algorithms use Kalman-like prediction strategies.

The technique in [34] considers a more general multistep predictor such that $\mathbf{h}_t \approx \sum_j \mathbf{A}_j \mathbf{h}_{t-j}$, and describes an approach for both the smoothing (which relies on both past and future data) and causal filtering (which relies only on the past data) problems.

DISCRETE MODELS

Time-series data often has hidden structure in which each time frame corresponds to a discrete hidden state q_t. Moreover, there is typically a relationship between the hidden states at different time frames, in the form of temporal dynamics. For example, each time frame of a speech signal corresponds to a subunit of speech such as a phoneme, which can be modeled as a distinct state. The subunits evolve over time as governed by temporal dynamics. Hidden Markov models (HMMs) [35] have been used extensively to model such data. They model temporal dynamics with a transition matrix defined by the distribution $p(q_t \,|\, q_{t-1})$. There has been a recent thread of literature [36]–[40] that combines these ideas with NMF to model nonnegative data with such structure.

The notion of a state is incorporated in the NMF framework by associating distinct dictionary elements with each state. This is done by allowing each state to determine a different support of the activations, which we express with the distribution $p(\mathbf{h}_t \,|\, q_t)$. This is to say that given a state, the model allows only certain dictionary elements to be active. Some techniques [36], [39] define the support of each state to be a single dictionary element, while other techniques [37], [38], [40], called nonnegative HMMs (N-HMMs), allow the support of each state to be a number of dictionary elements. Since only a subset of the dictionary elements are active at each time frame (as determined by the state at that time frame), we can interpret these models as imposing block sparsity on the dictionary elements [41].

[TEMPORAL CONTINUITY IS ONE OF THE MOST IMPORTANT FEATURES OF TIME-SERIES DATA.]

As in (7), there is a dependency between $\mathbf{h}_t$ and $\mathbf{h}_{t-1}$. However, unlike the continuous models, this dependency is only through the hidden states, which are in turn related through the temporal dynamics. Therefore $\mathbf{h}_t$ is conditionally independent of $\mathbf{h}_{t-1}$ given q_t or q_{t-1}. In the case of discrete models, we can therefore replace (7) with

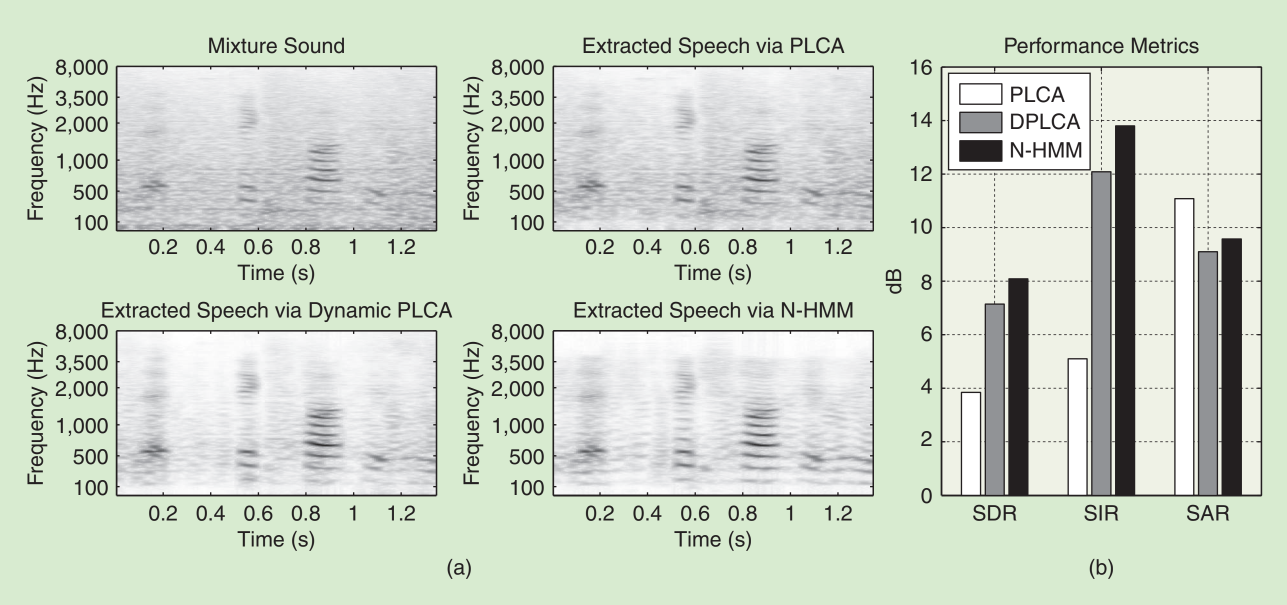
$$q_t \sim p(q_t \,|\, q_{t-1}), \tag{11}$$
$$\mathbf{h}_t \sim p(\mathbf{h}_t \,|\, q_t). \tag{12}$$

Since these models incorporate an HMM structure into an NMF framework, one can make use of the vast theory of Markov chains to extend these models in various ways. For example, one can incorporate high-level knowledge of a particular class of signals into the model, use higher-order Markov chains, or use various natural language processing techniques. Language models were recently incorporated in this framework [42] as typically done in the speech recognition literature [35]. Similarly, one can incorporate other types of temporal structure like music theory rules when dealing with music signals.

The above techniques discuss how to model a single source using an HMM structure. However, to perform source separation, we need to model mixtures. This is typically done by combining the individual source models into a factorial HMM [28], [36]–[38], [40], which allows each source to be governed by a distinct pattern of temporal dynamics. One issue with this strategy is that the computational complexity of inference is exponential in the number of sources. This can be circumvented using approximate inference techniques such as variational inference [43], which makes the complexity linear in the number of sources.

THE USE OF DYNAMIC MODELS IN SOURCE SEPARATION

To demonstrate the utility of dynamic models in context, we will once again use a real-world source separation example. This time it will be an acoustic mixture of speech mixed with background noise from a factory (using the TIMIT [47] and NOISEX-92 [48] databases). The mixture is shown using a magnitude STFT representation in Figure 4. This particular case is interesting because of the statistics of speech. We note that human speech



[FIG4] An example of dynamic models for source separation. (a) The four spectrograms show the mixture and the extracted speech for three different approaches. (b) A quantitative evaluation of the separation performance of each approach.

tends to have a smooth acoustic trajectory, which means that there is a strong temporal correlation between adjacent time frames. On the other hand, we also know that speech has a strong discrete hidden structure that is associated with the sequence of spoken phonemes. These properties make this example a good candidate for demonstrating the differences between the methods discussed so far and their effects on source separation.

We performed source separation using the three main approaches that we covered in this article. These include a static PLCA model [44], a dynamic PLCA model [34], and an N-HMM [37]. In all three cases, we trained a model for speech and a model for background noise from training data. The dictionary size for the noise was fixed to 30 elements, whereas the speech model had 60 dictionary elements for PLCA and dynamic PLCA, and 40 states with ten dictionary elements each for the N-HMM. For the dynamic models, we learned the temporal statistics as well. To separate a mixture of test data of the sources, we fixed the learned $\mathbf{W}$ matrices for both the speech and noise models and estimated their respective activations $\mathbf{H}$ using the context of each model. In Figure 4, we show the reconstruction of speech using each model. We also show a set of objective metrics that evaluate the quality of separation in each case. These include the source-to-distortion ratio (SDR), the source-to-interference ratio (SIR), and the source-to-artifacts ratio (SAR) as defined in [45]. These results are averaged over 20 different speakers to reduce biasing and initialization effects.

WE HOPE THAT BY PRESENTING THIS STREAMLINED FORMULATION WE CAN HELP READERS TO EXPERIMENT WITH THE OTHER MANY POSSIBILITIES IN FORMULATING DYNAMIC SOURCE SEPARATION ALGORITHMS.

For the static PLCA model, we see that there is a detectable amount of visible suppression of the background noise, which amounts to a modest SIR of about 5 dB. The dynamic PLCA model on the other hand, by taking advantage of the temporal statistics of speech, does a much better job resulting in more than double the SIR. Note however that in the process of adhering to the expected statistics, it introduces artifacts, which result in a lower SAR as compared to the static model. The N-HMM results in an even higher SIR and a better SAR than the dynamic PLCA model. This is because the specific signal we are modeling has a temporal structure that is well described by a discrete dynamic model as we transition from phoneme to phoneme. By constraining our model to only use a small dictionary at each discrete state, we obtain a cleaner estimate of the source. An example of that can be seen when comparing the separation results in Figure 4, where unwanted artifacts between the harmonics of speech in the dynamic PLCA example are not present in the N-HMM example since the dictionary elements within a state cannot produce such complex spectra.

WHICH MODEL TO USE?

Now, in addition to pondering on which divergence function is the most appropriate to employ, we also have a decision to make on which model is best fo,r a source separation approach. As always, the answer depends on the nature of the sources in the mixture. In general, the static model has found success in a variety of areas but does not take advantage of temporal correlations. In domains where we do not expect a high degree of correlations across time (e.g., short, burstlike sources) this model works well, but in cases where we expect a strong sense of continuity (e.g., a smooth source like a whale song), then a continuous dynamic model would work better. Furthermore, if we know that a source exhibits a behavior of switching through different states, each with its own unique character (e.g., speech), then a model like the N-HMM is more appropriate since it will eliminate the concurrent use of elements that belong at different states and produce a more plausible reconstruction. Of course, by using the generalized formulation we present in this article, there is nothing that limits us from employing different models concurrently. It is entirely plausible to design a source separation system where one source is modeled by a static model and other by a dynamic one, or even have both being described by different kinds of dynamic models. Doing so usually requires a relatively straightforward application of the estimation process that we outlined earlier.

CLOSING THOUGHTS

In this article we presented a unifying look at source separation approaches that employ nonnegative factorizations, and showed how they can be easily extended to temporal models that are either continuous or discrete. Using this methodology one can come up with many more alternative formulations, e.g., factorial HMMs, switching models, etc. and incorporate even more complex priors to better model sources in mixtures. We hope that by presenting this streamlined formulation we can help readers to experiment with the many other possibilities in formulating dynamic source separation algorithms and to help highlight relationships between a family of approaches that can initially seem divergent despite their common roots.

AUTHORS

Paris Smaragdis (paris@illinois.edu) is an assistant professor in the Computer Science Department and the Electrical and Computer Science Department at the University of Illinois at Urbana-Champaign, as well as a senior research scientist at Adobe Research. Prior to that he was a research scientist at Mitsubishi Electric Research Labs, during which time he was selected by *MIT Technology Review* as one of the top 35 young innovators of 2006. His research interests lie in the intersection of machine learning and signal processing. He is a Senior Member of the IEEE.

Cédric Févotte (cfevotte@unice.fr) received his Ph.D. degree from École Centrale de Nantes (2000–2003) and was with the University of Cambridge as a postdoctoral researcher (2003–2006). Previously, he was affiliated with Mist-Technologies (now Audionamix) as a research engineer (2006) and Télécom ParisTech (2007–2013). His research interests generally concern statistical signal processing and machine learning for inverse problems and source separation. He is a CNRS research scientist at the Laboratoire Lagrange (CNRS, Observatoire de la Côte d'Azur and the Université de Nice Sophia Antipolis), France. He is a member of the IEEE Machine Learning for Signal Processing Technical Committee.

Gautham J. Mysore (gmysore@adobe.com) received the M.A. and Ph.D. degrees from the Center for Computer Research in Music and Acoustics at Stanford University and the M.S. degree in electrical engineering, also from Stanford. He is a senior research scientist at Adobe Research in San Francisco, California, and was previously a visiting researcher at the Gatsby Computational Neuroscience Unit at the University College London. His research interests include various aspects of machine learning and signal processing for audio. He is currently a member of the IEEE Technical Committee on Audio and Acoustic Signal Processing.

Nasser Mohammadiha (nmoh@kth.se) received the M.S. degree in electronics engineering from Sharif University of Technology, Tehran, Iran, in 2006. He worked on digital hardware and software design until 2008. He received his Ph.D. degree from the Communication Theory Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden, in 2013. He received the Best Student Paper Award at the 2013 European Signal Processing Conference. Since November 2013, he has been a postdoctoral researcher at the University of Oldenburg, Germany. His research interests include machine learning for signal processing, statistical signal modeling, speech processing, and image processing. He is a Member of the IEEE.

Matthew Hoffman (mathoffm@adobe.com) is a research scientist in the Creative Technologies Laboratory at Adobe. He was previously a postdoctoral researcher working with Prof. Andrew Gelman in the Statistics Department at Columbia University. He completed his Ph.D. degree at Princeton University in computer science working in the Sound Lab with Prof. Perry Cook and Prof. David Blei. His research interests include developing efficient Bayesian (and pseudo-Bayesian) inference algorithms; hierarchical probabilistic modeling of audio, text, and marketing data; and audio feature extraction, music information retrieval, and the application of music information retrieval and modeling techniques to musical synthesis.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] A. Singh and G. Gordon, "A unified view of matrix factorization models," *Mach. Learn. Knowledge Discovery Databases*, vol. 5212, pp. 358–373, 2008.

[3] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Independent Component Analysis and Signal Separation* (Lecture Notes in Computer Science, vol. 4666). Berlin, Heidelberg, Germany: Springer, 2007, pp. 414–421.

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

[5] A. Cichocki, S. Cruces, and S. Amari, "Generalized Alpha-Beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, pp. 134–170, 2011.

[6] I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," *Adv. Neural Inform. Process. Syst. (NIPS)*, vol. 19, pp. 283–290, 2005.

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

[8] M. D. Hoffman, "Poisson-uniform nonnegative matrix factorization," in *Proc. IEEE Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2012, pp. 5361–5364.

[9] P. Smaragdis, B. Raj, and M. V. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *NIPS Workshop on Advances in Models for Acoustic Processing*, 2006.

[10] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Int. Conf. Research Development Information Retrieval (SIGIR)*, 1999, pp. 50–57.

[11] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *Proc. IEEE Int. Workshop Machine Learning Signal Processing (MLSP)*, Santander, Spain, Sep. 2012, pp. 1–6.

[12] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, pp. 30–37, 2004. [Online]. Available: http://amstat.tandfonline.com/doi/abs/10.1198/0003130042836#tabModule

[13] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computat.*, vol. 23, no. 9, pp. 2421–2456, Sept. 2011.

[14] S. Arora, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *Proc. Int. Conf. Machine Learning*, 2013, pp. 280–288.

[15] A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade, "Learning latent Bayesian networks and topic models under expansion constraints," submitted for publication. [Online]. Available: http://arxiv.org/pdf/1209.5350.pdf

[16] K. Huang, N. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Processing*, vol. 62, no. 1, pp. 211–224, 2014.

[17] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[18] P. D. O'Grady and B. A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputation*, vol. 72, nos. 1–3. pp. 88–101, 2008.

[19] W. Wang, A. Cichocki, and J. A. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance," *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2858–2864, July 2009.

[20] Z. Chen, A. Cichocki, and T. M. Rutkowski, "Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer's disease," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[21] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech Lang. Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[22] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proc. Int. Conf. Spoken Language Processing (Interspeech)*, 2008, pp. 411–414.

[23] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 1825–1828.

[24] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proc. IEEE Workshop Applications Signal Processing Audio Acoustics (WASPAA)*, Oct. 2011, pp. 45–48.

[25] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, May 2011, pp. 1980–983.

[26] S. Essid and Févotte, "Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 415–425, Feb. 2013.

[27] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using NMF," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[28] M. Nakano, J. Le Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama, "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model," in *Proc. IEEE Workshop Applications Signal Processing Audio Acoustics (WASPAA)*, New Paltz, NY, Oct. 2011, pp. 325–328.

[29] K. Yoshii and M. Goto, "Infinite composite autoregressive models for music signal analysis," in *Proc. 13th Int. Society Music Information Retrieval Conf. (ISMIR)*, Oct. 2012, pp. 79–84.

[30] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Mar. 2012, pp. 4561–4564.

[31] M. Kim and P. Smaragdis, "Single channel source separation using smooth nonnegative matrix factorization with Markov random fields," in *IEEE Workshop Machine Learning Signal Processing (MLSP2013)*, Southampton, U.K., Sept. 2013, pp. 1–6.

[32] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 3158–3162.

[33] J. Nam, G. Mysore, and P. Smaragdis, "Sound recognition in mixtures," in *Latent Variable Analysis and Signal Separation* (Lecture Notes in Computer Science, vol. 7191). Berlin, Heidelberg, Germany: Springer, 2012, pp. 405–413.

[34] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, May 2013, pp. 873–877.

[35] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[36] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop Applications Signal Processing Audio Acoustics (WASPAA)*, Oct. 2009, pp. 121–124.

[37] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Int. Conf. Latent Variable Analysis and Signal Separation*, 2010, pp. 140–148.

[38] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, May 2011, pp. 17–20.

[39] M. Nakano, J. Roux, H. Kameoka, Y. Kitano, N. Ono, and S. Sagayama, "Nonnegative matrix factorization with markov-chained bases for modeling time-varying patterns in music spectrograms," in *Latent Variable Analysis and Signal Separation* (Lecture Notes in Computer Science, vol. 6365). Berlin, Heidelberg, Germany: Springer, 2010, pp. 149–156.

[40] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 5, pp. 998–1011, May 2013.

[41] G. J. Mysore, "A block sparsity approach to multiple dictionary learning for audio modeling," in *Proc. Int. Conf. Machine Learning (ICML) Workshop Sparsity, Dictionaries, Projections Machine Learning Signal Processing*, June 2012.

[42] G. Mysore and P. Smaragdis, "A non-negative approach to language informed speech separation," in *Latent Variable Analysis and Signal Separation* (Lecture Notes in Computer Science, vol. 7191). Berlin, Heidelberg, Germany: Springer, 2012, pp. 356–363.

[43] G. J. Mysore and M. Sahani, "Variational inference in non-negative factorial hidden Markov models for efficient audio source separation," in *Proc. Int. Conf. Machine Learning (ICML)*, June 2012, pp. 1887–1894.

[44] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Independent Component Analysis and Signal Separation* (Lecture Notes in Computer Science, vol. 4666). Berlin, Heidelberg: Springer, 2007, pp. 414–421.

[45] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[46] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie, "Nonnegative matrix and tensor factorizations," *IEEE Signal Processing Mag.*, vol. 31, no. 3, pp. 54–65, May 2014.

[47] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallet. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI Tech. Rep. N, 93, 27403.

[48] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[SP]