ITAKURA-SAITO NONNEGATIVE MATRIX FACTORIZATION WITH GROUP SPARSITY

Augustin Lefèvre[⋆†] *Francis Bach*[⋆] *Cédric Févotte*[†]

[⋆] INRIA / ENS - Sierra team [†] CNRS LTCI / Telecom ParisTech

ABSTRACT

We propose an unsupervised inference procedure for audio source separation. Components in nonnegative matrix factorization (NMF) are grouped automatically in audio sources via a penalized maximum likelihood approach. The penalty term we introduce favors sparsity at the group level, and is motivated by the assumption that the local amplitude of the sources are independent. Our algorithm extends multiplicative updates for NMF ; moreover we propose a test statistic to tune hyperparameters in our model, and illustrate its adequacy on synthetic data. Results on real audio tracks show that our sparsity prior allows to identify audio sources without knowledge on their spectral properties.

Index Terms— Blind source separation, audio signal processing, unsupervised learning, nonnegative matrix factorization, sparsity priors

1. INTRODUCTION

In this paper, we propose a contribution to the problem of unsupervised source separation of audio signals, more specifically single channel audio signals. Nonnegative matrix factorization (NMF) of time-frequency representations such as the power spectrogram has become a popular tool in the signal processing community. Given such a time-frequency representation $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, NMF consists in finding a factorization of the form $\mathbf{V} \simeq \mathbf{WH}$ where $\mathbf{W} \in \mathbb{R}_+^{F\times K}$, $\mathbf{H} \in \mathbb{R}_+^{K\times N}$, and $K \ll F, N$. The factorization is obtained by minimizing a loss function of the form $D(\mathbf{V}, \mathbf{WH})$. For simple signals, individual components of NMF were found to retrieve meaningful signals such as notes or events [1, 2]. However, when applied to more complex signals, such as music instruments, it is more reasonable to suppose that each sound source corresponds to a subset of components. Grouping is usually done either by the user, or based on heuristics, but as the number of components grows large, this task becomes even more time-consuming than the parameter inference task (it involves considering all permutations of K components). In this paper, we argue that grouping may be incorporated in the inference of the dictionary $\mathbf{W}$ as part of a structured statistical model. We make the hypothesis that the instantaneous local amplitudes (i.e., the "volume") of the sources are independent and derive a marginal distribution for $\mathbf{H}$. This results in a maximum likelihood problem penalized with a sparsity-inducing term. Sparsity-inducing functions have been a subject of intensive research. According to the loss function used, either sparsity-inducing norms [3, 4] or divergences [1, 5] are preferred. The penalty term we introduce is designed to deal with a specific choice of loss function, the Itakura-Saito divergence. This paper is organized as follows : in Section 2 we propose a penalized maximum-likelihood estimation method, that favors group-sparsity in NMF. We provide in Section 3 an efficient descent algorithm, building on a majorization-minimization procedure. In Section 4.2 we propose a statistic to select hyperparameters. In Section 5, we validate our algorithm and parameter selection procedure on synthetic data and discuss the influence of remaining free parameters. Finally, we emphasize the benefits of our approach in an unsupervised audio source separation task.

Notation. Matrices are bold upper-case (e.g., $\mathbf{X} \in \mathbb{R}^{F\times N}$), column vectors are bold lower-case (e.g., $\mathbf{x} \in \mathbb{R}^F$), and scalars are plain lower case (e.g., $x \in \mathbb{R}$). $\mathbf{x}_{\cdot n}$ denotes the n-th column of matrix $\mathbf{X}$, $\mathbf{x}_{f\cdot}$ the k-th line, while x_{fn} is the (f, n) coefficient. Moreover, if g is a set of integers, then $\mathbf{h}_g$ is a vector in $\mathbb{R}^{|g|}$ of elements of $\mathbf{h}$ indexed by g. In algorithms we write elementwise matrix multiplication $\mathbf{A} \odot \mathbf{B}$, division $\frac{\mathbf{A}}{\mathbf{B}}$, matrix power $\mathbf{A}^{\cdot k}$, and coefficientwise modulus $|\mathbf{A}|$. For any vector or matrix $\mathbf{X}$, $\mathbf{X} \geq 0$ means that all entries are nonnegative. Sums are for $k \in \{1 \dots K\}, f \in \{1 \dots F\}, n \in \{1 \dots N\}$, unless otherwise stated. Finally, we use the convention $\tilde{\mathbf{V}} = \mathbf{WH}$ throughout the paper.

2. STATISTICAL FRAMEWORK AND OPTIMIZATION PROBLEM

2.1. Overview of the generative model

Given a short time Fourier transform $\mathbf{X} \in \mathbb{C}^{F\times N}$ of an audio track, we make the assumption that $\mathbf{X}$ is a linear instantaneous mixture of i.i.d. Gaussian signals :

$$x_{fn} = \sum_k x_{fn}^{(k)} \quad \text{where} \quad x_{fn}^{(k)} \sim \mathcal{N}(0, w_{fk}h_{kn}) \,. \qquad (1)$$

As a consequence, we have $\mathbb{E}(\mathbf{V}) = \mathbf{WH}$ where $\mathbf{V} = |\mathbf{X}|^2$ is the observed power spectrogram. Furthermore, $\mathbf{V}$ has the

This work is supported by project ANR-09-JCJC-0073-01 TANGERINE and SIERRA-ERC-239993 .

following distribution :

$$p(\mathbf{V}|\tilde{\mathbf{V}}) = \prod_{f,n} \frac{1}{\tilde{v}_{fn}} \exp\left(-\frac{v_{fn}}{\tilde{v}_{fn}}\right). \quad (2)$$

As shown in [2], maximum-likelihood estimation of $(\mathbf{W}, \mathbf{H})$ is equivalent to minimizing the Itakura-Saito divergence between $\mathbf{V}$ and $\mathbf{WH}$. The Itakura-Saito loss is defined on strictly positive scalars by : $d_{IS}(x, y) = \frac{x}{y} - \log\frac{x}{y} - 1$. It may be generalized to vectors or matrices by summing over all available components, e.g., $D_{IS}(\mathbf{X}, \mathbf{Y}) = \sum_{f,n} d_{IS}(x_{fn}, y_{fn})$. In our model, the power spectral density of each component k is $\mathbf{w}_{\cdot k}\mathbf{h}_{k\cdot} \in \mathbb{R}_+^{F\times N}$. If we assume a source is characterized by a subset of components g, then the Wiener filter is an unbiased estimator of $\mathbf{X}^{(g)} = \sum_{k\in g} \mathbf{X}^{(k)}$:

$$\mathbb{E}\left(\mathbf{X}^{(g)}|\mathbf{X}, \mathbf{W}, \mathbf{H}\right) = \frac{\mathbf{W}_{\cdot g}\mathbf{H}_{g\cdot}}{\mathbf{WH}} \odot \mathbf{X}\,. \quad (3)$$

2.2. Maximum Likelihood with a sparsity penalty

We wish to partition the K components into G non-overlapping groups. In the following a source will be uniquely identified by the subset g to which it corresponds. In the framework of statistical inference, many priors have been proposed to identify components, either on $\mathbf{W}$ or $\mathbf{H}$ or both (see e.g., [1, 5]). We focus here on a simple grouping principle : if a source is inactive at a given frame n of the spectrogram, then all the corresponding gains $\mathbf{h}_{gn}$ should be set to zero. Instead of modelling individual gains we model the local amplitudes $\alpha_n^{(g)}$ of the sources, which we define as follows : assume $\|\mathbf{w}_{\cdot k}\|_1 = 1$ without loss of generality, and define $\alpha_n^{(g)} = \|\mathbf{h}_{gn}\|_1$. We make the hypothesis that $\alpha_n^{(g)}$ are mutually independent variables drawn from an inverse Gamma distribution with shape parameter b and scale a. Furthermore we suppose that the conditional distribution of the gains h_{kn} factorizes in groups, i.e., $p(\mathbf{h}_{\cdot n}|(\alpha_n^{(g)})_{g\in\mathcal{G}}) = \prod_g \prod_{k\in g} p(h_{kn}|\alpha_n^{(g)})$; and that h_{kn} are exponentially distributed conditionally on $\alpha_n^{(g)}$, with mean $\alpha_n^{(g)}$. The marginal distribution of $\mathbf{h}_{\cdot n}$ is then given by :

$$p(\mathbf{h}_{\cdot n}) = \prod_g \frac{\Gamma(|g| + b)}{\Gamma(b)} \frac{a^b}{(a + \|\mathbf{h}_{gn}\|_1)^{b+|g|}}\,. \quad (4)$$

Combining this prior and the likelihood term, we propose a penalized maximum likelihood inference in the following form :

$$\begin{array}{cc} \min & D_{IS}(\mathbf{V}, \mathbf{WH}) + \lambda\Psi(\mathbf{H}), \\ \mathbf{W} \geq 0, \mathbf{H} \geq 0 & \\ \forall k,\ \|\mathbf{w}_{\cdot k}\|_1 = 1 & \end{array} \quad (5)$$

with $\Psi(\mathbf{H}) = \sum_{g,n} \psi(\|\mathbf{h}_{gn}\|_1)$ and $\psi(x) = \log(a + x)$. We refer to Eq. (5) as the GIS-NMF problem (group Itakura-Saito NMF), and call $\mathcal{L}(\mathbf{W}, \mathbf{H})$ the objective function. Eq. (5) generalizes IS-NMF in the sense that when $\lambda = 0$ we recover the standard IS-NMF problem. In GIS-NMF a tradeoff is made between the fit to data as measured by the loss term, and the grouping criterion defined by Ψ. Although we impose a particular choice of ψ, note that for optimization purposes we only require that ψ be a differentiable, concave, increasing function.

3. MAJORIZATION-MINIMIZATION ALGORITHMS

In this Section we derive an efficient algorithm for solving Eq. (5), inspired by multiplicative updates and majorization minimization techniques [6]. We optimize alternately in $\mathbf{W}$ and $\mathbf{H}$. Descent at each step yields a descent algorithm.

3.1. Updates in H

The objective function is separable in the columns of $\mathbf{H}$, so we need only consider the following subproblem :

$$\min_{\mathbf{h}\geq 0} \quad D_{IS}(\mathbf{v}, \mathbf{Wh}) + \lambda\Psi(\mathbf{h})\,. \quad (6)$$

where $\mathbf{h} \in \mathbb{R}_+^K$. Let $\underline{\mathbf{h}} \geq 0$ be the current estimate for $\mathbf{h}$. The authors of [6] derived an auxiliary function for $D_{IS}(\mathbf{v}, \mathbf{Wh})$ in the following form :

$$g(\mathbf{h}, \underline{\mathbf{h}}) = \sum_k \left(-p_k \frac{\underline{h}_k}{h_k} + q_k\right)(h_k - \underline{h}_k)\,. \quad (7)$$

These derivations may easily accommodate additional concave or convex terms. Applying the tangent inequality to $x \mapsto \log(a + x)$, we derive an auxiliary function for the objective function in (6) :

$$g(\mathbf{h}, \underline{\mathbf{h}}) = \sum_{g,k\in g}\left(-p_k \frac{\underline{h}_k}{h_k} + q_k + \lambda\psi'(\|\underline{\mathbf{h}}_g\|_1)\right)(h_k - \underline{h}_k)\,, \quad (8)$$

We may either (a) minimize the right hand side with respect to h_k, or (b) set each term $-p_k\frac{\underline{h}_k}{h_k} + q_k$ to 0, yielding the multiplicative updates found in [2]. We obtain new multiplicative updates :

$$\forall g, \forall k \in g,\ h_k = \underline{h}_k\left(\frac{\sum_f w_{fk}\frac{v_f}{(\mathbf{Wh})_f^2}}{\sum_f w_{fk}\frac{1}{(\mathbf{Wh})_f} + \lambda\psi'(\|\mathbf{h}_g\|_1)}\right)^\delta, \quad (9)$$

where $\delta = 0.5$ (a) or $\delta = 1$ (b). The additional term in the denominator in Equation (9) favors low values of $\mathbf{h}_{kn}$: since $\psi'(x)$ decreases with x (ψ is concave), low values of $\|\mathbf{h}_g\|_1$ are more penalized than high values. Moreover the quantity $\|\mathbf{h}_g\|_1$ is the same for all k in group g. Thus, if at a given frame n the volume of source g is small with respect to that of source g', the updates in (5) tend to mute source g. We thus get the same grouping effect than the traditional penalization by the ℓ_2-norms $\|\mathbf{h}_g\|_2$ [3], but with the added benefit of natural multiplicative updates.

3.2. Updates in W

To optimize with respect to $\mathbf{W}$, we notice that the minimizers of Eq. (5) are also minimizers of :

$$\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} D_{IS}(\mathbf{V}, \mathbf{WH}) + \lambda \Phi(\mathbf{W}, \mathbf{H}), \quad (10)$$

where $\Phi(\mathbf{W}, \mathbf{H}) = \sum_g \sum_n \psi(\sum_{k \in g} h_{kn} \|\mathbf{w}_{\cdot k}\|_1)$. Thus updates for $\mathbf{W}$ may be derived in the same way as for $\mathbf{H}$. Since the objective function in (10) is unchanged under the transformation $\mathbf{W} \leftarrow \mathbf{W}\Lambda^{-1}$, $\mathbf{H} \leftarrow \Lambda\mathbf{H}$, where Λ is a diagonal matrix, we may rescale matrices $\mathbf{W}$ and $\mathbf{H}$ at each step to return to the feasible set of (5). Thus, we derived a descent algorithm to solve Eq. (5), that is summed up in Algorithm 1.

4. PROCEDURE FOR DICTIONARY LEARNING AND INFERENCE

Algorithm 1 Algorithm for GIS-NMF

Input $\mathbf{V}$, $(\mathbf{W}, \mathbf{H})$, $\mathcal{G}$, (λ, a), δ,t
For t iterations
$\hat{\mathbf{V}} \leftarrow \mathbf{WH}$
For $n = 1 \dots N, g \in \mathcal{G}, k \in g$
$p_{kn} \leftarrow \psi'(\|\mathbf{h}_{gn}\|_1)$
End

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^\top (\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{W}^\top (\hat{\mathbf{V}}^{\cdot -1}) + \lambda \mathbf{P}} \right)^{\cdot \delta}, \quad \hat{\mathbf{V}} \leftarrow \mathbf{WH},$$

For $f = 1 \dots F, g \in \mathcal{G}, k \in g$
$r_{fk} = \sum_n h_{kn} \psi'(\|\mathbf{h}_{gn}\|_1)$
End

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\frac{\mathbf{H}^\top (\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{H}^\top (\hat{\mathbf{V}}^{\cdot -1}) + \lambda \mathbf{R}} \right)^{\cdot \delta},$$

$\mathbf{\Lambda} = \text{diag}(\|\mathbf{w}_{\cdot 1}\|_1, \dots, \|\mathbf{w}_{\cdot K}\|_1)$
$\mathbf{W} \leftarrow \mathbf{W}\,\Lambda^{-1} \quad \mathbf{H} \leftarrow \Lambda\,\mathbf{H}, \quad \hat{\mathbf{V}} \leftarrow \mathbf{WH}.$
End

In Section 4.1, we summarize our procedure in a simple and fast algorithm with multiplicative updates. In practice the choice of regularization parameters a and λ is crucial. In Section 4.2, we propose a test statistic inspired from the Kolmogorov-Smirnov test to perform this selection automatically.

4.1. Multiplicative updates algorithm

Algorithm 1 sums up our discussion in Section 3. It encompasses both multiplicative updates with or without a square root exponent by adding a parameter $\delta = 0.5$ or 1. Our algorithm is of complexity $O(FKN)$ in time and memory, like many multiplicative updates algorithms. We run the algorithm with several different initializations and keep the result that yields the lowest cost value, in order to avoid local minima. The objective function decreases at each step, and convergence of the parameters is observed in practice. We stop the algorithm after a fixed number of iterations.

4.2. Selection of hyperparameters with a Kolmogorov-Smirnov statistics

Define standardized observations $\varepsilon_{fn} = v_{fn}/\tilde{v}_{fn}$. Then if the observed data follow the model in Eq.(2), the empirical distribution function of $\mathbf{E}$ converges towards that of an exponential random variable. We propose to select the parameters of our model $(a,\ \lambda)$ that yield the minimum Kolmogorov-Smirnov (KS) statistic (see [7] for more details).

5. RESULTS

5.1. Validation on synthetic data

We designed an optimization procedure to enforce structured sparsity on the columns of $\mathbf{H}$. In order to validate our algorithm, we picked $\mathbf{W}^{(\star)} \in \mathbb{R}_+^{100 \times 20}$ at random and $\mathbf{H}^{(\star)}$ with two groups of 10 components each and disjoint supports. 10 synthetic data sets of various sizes were generated according to model (2). Define the support recovery error as the proportion of frames where the active sources are incorrectly identified. Figure 1 displays, for various data set sizes N, how the test statistic and the support recovery error vary with λ. For fixed N, the KS statistic reaches a minimum in the interval $[10^0, 10^2]$. As N grows large, the support recovery error decreases towards zero, and the minimizer of the KS statistic (which does not require to know the ground truth) matches the one of the recovery error.

5.2. Results in single channel source separation

track	source	GIS-NMF	base	random	ideal
love	bass	**8.88**	-67.53	-8.55	8.86
0 %	guitar	**13.60**	3.77	-2.19	13.94 [1]
love	bass	**4.33**	-4.60	-8.74	4.56
33 %	guitar	**9.77**	-7.40	-2.02	9.90
love	bass	**1.47**	-5.29	-9.08	3.12
66 %	guitar	**7.72**	-8.11	-1.94	8.68
love	bass	-5.13	**-4.16**	-9.02	2.54
100 %	guitar	**-0.21**	-2.68	-2.02	8.09

Table 1. Source to distortion ratios (SDR) for the track "We are in love" [2] . $x\%$ is the overlap between sources.

We experiment our algorithm on two audio tracks found on the Internet Archive (`www.archive.org`) : the individual sources $\mathbf{x}^{(g)}$, $g = 1 \dots 2$. were available, from which we took 20-30 seconds excerpts [2]. For each track, we propose the

1. "ideal NMF" serves for comparison, but is not an upper bound for the performance of our algorithm, see text.

2. Complete results on all mixtures, including .wav files, are available online (`www.di.ens.fr/~lefevrea/demos.html`)

following mixture :

$$x_n = \begin{cases} x_n^{(1)} & \text{if } n \leq \frac{1-p}{2} T \\ x_n^{(2)} & \text{if } n \geq \frac{1+p}{2} T \\ x_n^{(1)} + x_n^{(2)} & \text{otherwise} \end{cases} . \quad (11)$$

where T is the total length of the track : thus if $p = 0.33$, we make sure that sources overlap over no more than 33% of the track. The goal is to analyze how important sparsity is to estimate the mixtures correctly by varying p. Table 1 compares our algorithm (GIS-NMF) with a baseline strategy, an ideal strategy and the result of a random binary mask. We take SDR (source to distortion ratio) as a performance indicator (see [8]). The baseline we took consists in estimating Itakura-Saito NMF and then group components so as to minimize $\Psi(\mathbf{H})$, so that $\Psi(\mathbf{H})$ plays the role of a heuristic criterion to group components. Ideal NMF consists in running NMF and choose groups that yield optimal SDR (by selecting from all possible of $K!$ permutations) : the aim of our procedure is to obtain the same performance. However, note that it is not an oracle performance (not the same objective function). Finally, we computed the average SDR of 10 random binary masks. In Table 1 we display our results on one audio track. In GIS-NMF, parameters (a, λ) were chosen to minimize the test statistic, then we tuned the number of components per group as to maximize SDR. In most cases, we perform better than a random binary mask, unlike the baseline. For overlap p up to 66%, we obtain SDR values close to that of the ideal i.e., we find the best assignment for source separation. Thus group-sparsity in the columns of $\mathbf{H}$ plays a key role in identifying sources. Our algorithm meets his limits when there is too much overlap, then we fail to identify the sources correctly, and more knowledge about the sources is needed.

6. CONCLUSION

We introduced an optimization procedure to find groups in NMF with the Itakura-Saito divergence. Instead of finding groups after running NMF, we incorporate grouping in the optimization. Our algorithm keeps the attractive features of multiplicative updates algorithm (low complexity, descent property), and allows to perform blind source separation on complex signals, with no assumption on the frequency profiles of the sources. We are working on adding temporal smoothness priors to improve separation quality.

7. REFERENCES

[1] P. Smaragdis, B. Raj, and M.V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *In Proc. Int. Conf. on ICA and Signal Separation. London, UK*, September 2007.

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

[3] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *Adv. NIPS*, 2010.

[4] R. Jenatton, F. Bach, and J.-Y. Audibert, "Structured variable selection with sparsity-inducing norms," Tech. Rep. 0904.3523, arXiv, 2009.

[5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 15, no. 3, March 2007.

[6] Y. Cao, P. Eggermont, and S. Terebey, "Cross Burg entropy maximization and its application to ringing suppression in image reconstruction," *IEEE Trans. on Image Proc.*, vol. 8, no. 2, pp. 286–292, Feb 1999.

[7] E. Lehmann and J.P. Romano, *Testing Statistical Hypotheses (Springer Texts in Statistics)*, Springer, 3rd edition, April 2005.

[8] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, no. 4, 2006.

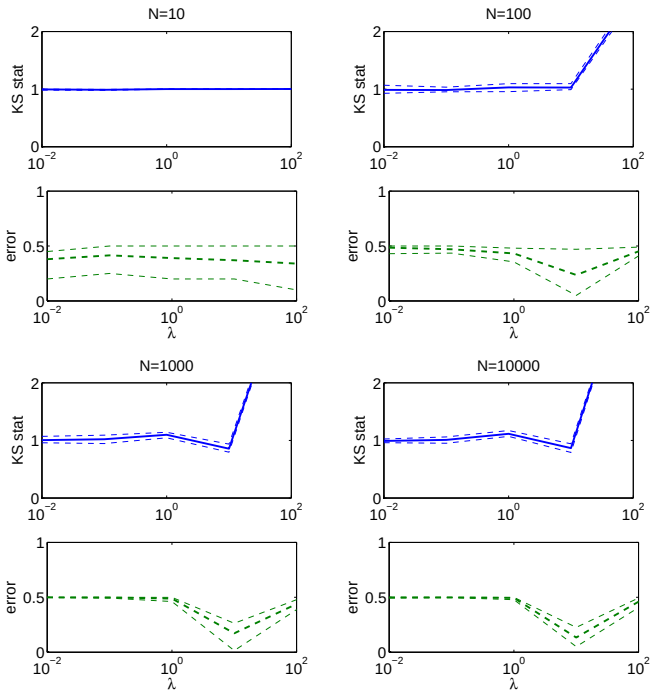

Fig. 1. Relationship between support recovery error and KS statistic as the size N of the data set increases. x-axis : regularization parameter λ. y-axis : KS statistic (solid line) and the support recovery error (dashed line). Thin dashed lines are error bars