MAJORIZATION-MINIMIZATION ALGORITHM FOR SMOOTH ITAKURA-SAITO NONNEGATIVE MATRIX FACTORIZATION

Cédric Févotte

CNRS LTCI; Télécom ParisTech
Paris, France

ABSTRACT

Nonnegative matrix factorization (NMF) with the Itakura-Saito divergence has proven efficient for audio source separation and music transcription, where the signal power spectrogram is factored into a "dictionary" matrix times an "activation" matrix. Given the nature of audio signals it is expected that the activation coefficients exhibit smoothness along time frames. This may be enforced by penalizing the NMF objective function with an extra term reflecting smoothness of the activation coefficients. We propose a novel regularization term that solves some deficiencies of our previous work and leads to an efficient implementation using a majorization-minimization procedure.

Index Terms— Nonnegative matrix factorization (NMF), Itakura-Saito divergence, regularization by smoothness, audio signal representation, single-channel source separation.

1. INTRODUCTION

Nonnegative matrix factorization (NMF) is a linear regression technique, employed for non-subtractive, part-based representation of nonnegative data [1]. Given a data matrix $\mathbf{V}$ of dimensions $F \times N$ with nonnegative entries, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $F\,K + K\,N << F\,N$, hence reducing the data dimension. Much research about NMF has been driven by applications in audio, namely automatic music transcription and source separation, where the data $\mathbf{V}$ is taken as the magnitude or power spectrogram of the audio signal, see, e.g., [2, 3]. In this setting the factorization amounts to decomposing the spectrogram data into a sum of rank-1 spectrograms, each of which being the expression of an elementary spectral pattern (columns of $\mathbf{W}$) amplitude-modulated in time (rows of $\mathbf{H}$).

This work is supported by project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization). Many thanks to J. Idier, F. Bach and A. Lefèvre for discussions related to this work.

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

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

where $d(x|y)$ is a scalar cost function, typically a positive function with a single minimum 0 for $x = y$. The minimization, with respect to (w.r.t) $\mathbf{W}$ and $\mathbf{H}$, is subject to nonnegativity constraints on the coefficients of both factors. In [4] it was shown that factorizing the *power* spectrogram using the Itakura-Saito (IS) divergence, defined by

$$d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1, \tag{3}$$

is relevant to audio for its two following properties. Firstly, the IS divergence is scale-invariant, i.e., $d_{IS}(\lambda\,x|\lambda\,y) = d_{IS}(x|y)$, a property which is not shared by the more common Euclidean distance and generalized Kullback-Leibler divergence. The scale-invariance is relevant to the decomposition of audio spectra, which typically have a large dynamic range and also comprise low-power transient components such as note attacks together with higher power components such as tonal parts of sustained notes. Secondly, the IS divergence leads to desirable statistical interpretations of the NMF problem. Indeed, IS-NMF of the power spectrogram can be recast as maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in a variance model of superimposed Gaussian components, i.e., a generative model of the short-time Fourier transform relevant to audio decomposition applications. See details in [4].

Given the nature of audio signals it is expected that the rows of $\mathbf{H}$ exhibit smoothness along time frames. The statistical composite model inherent to IS-NMF was exploited in [4] to design an EM algorithm for maximum a posteriori estimation of $\mathbf{H}$ under Gamma and inverse-Gamma Markov chain priors. As a matter of fact the regularization term resulting from this prior leads to an ill-posed optimization problem, as discussed later. The aim of this paper is to propose an alternative to the latter EM approach, based on a majorization-minimization (MM) procedure [5] which in turn leads to a more efficient implementation.[1]

[1] The EM algorithm is a special case of MM algorithm, but the MM algo-

Note that other works have considered NMF with smoothness constraints on $\mathbf{H}$, e.g., [6, 3, 7, 8], but using either the Euclidean distance or KL divergence as the measure of fit to data, while we are here specifically interested with the IS divergence for its relevance in the audio setting.

The paper is organized as follows. Section 2 describes our previous work and propose a novel scale-invariant measure of smoothness. A MM algorithm for IS-NMF under this new smoothness constraint is then described in Section 3. Section 4 reports results on the decomposition of a jazz excerpt and Section 5 concludes. This paper comes with a companion webpage [9] which offers MATLAB code of the presented algorithm and audio demo samples.

2. SMOOTHNESS CONSTRAINTS FOR IS-NMF

In the following, the entries of matrices $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$ are denoted v_{fn}, w_{fk} and h_{kn} respectively. Lower case bold letters denote columns, such that $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]$.

2.1. Previous work

In [4] were proposed smoothness constraints in the form of nonnegative Markov chains, such that

$$p(\mathbf{H}) = \prod_{k=1}^{K} \prod_{n=2}^{N} p(h_{kn}|h_{k(n-1)})\, p(h_{k1}), \tag{4}$$

where the Markov kernel $p(h_{kn}|h_{k(n-1)})$ is a probability density function (pdf) defined on the nonnegative orthant, with mode at $h_{k(n-1)}$. Possible choices of kernels are Gamma or inverse-Gamma pdfs, both considered in [4]. In the second case, the kernel writes

$$p(h_{kn}|h_{k(n-1)}) = \mathcal{IG}(h_{kn}|\alpha, (\alpha+1)\, h_{k(n-1)}) \tag{5}$$

where $\mathcal{IG}(u|\alpha,\beta) = \beta^{\alpha} u^{-(\alpha+1)} \exp(-\beta/u)\, \Gamma(\alpha)$, $u \geq 0$. The MAP criterion function under this prior leads to the following penalized functional

$$C_1(\mathbf{W},\mathbf{H}) = D_{IS}(\mathbf{V}|\mathbf{WH}) + (\alpha+1)P_1(\mathbf{H}), \tag{6}$$

$$P_1(\mathbf{H}) = \sum_{k=1}^{K} \sum_{n=2}^{N} d_{IS}(h_{k(n-1)}|h_{kn}) + \frac{\log h_{k(n-1)}}{\alpha+1} + cst, \tag{7}$$

and where cst denotes constant terms w.r.t $\mathbf{H}$. The latter expression assumes a constant prior for h_{k1}. Considering a Gamma kernel instead of inverse-Gamma would lead to a similar expression but where $d_{IS}(h_{k(n-1)}|h_{kn})$ is replaced with $d_{IS}(h_{kn}|h_{k(n-1)})$. As it turns out the optimization of $C_1(\mathbf{W},\mathbf{H})$ under mere nonnegativity constraints of $\mathbf{W}$ and $\mathbf{H}$ is ill-posed. Indeed, for any $\mathbf{W}$, $\mathbf{H}$ and nonnegative diagonal matrix $\mathbf{\Delta}$ with coefficients δ_k we have

$$C_1(\mathbf{W}\mathbf{\Delta}^{-1}, \mathbf{\Delta}\mathbf{H}) = C_1(\mathbf{W},\mathbf{H}) + (N-1)\sum_k \log \delta_k \tag{8}$$

so that one obtains the degenerate solution $\hat{\mathbf{W}} \to \infty$, $\hat{\mathbf{H}} \to 0$ in the limit when $\mathbf{\Delta} \to 0$. In our previous work [4], this degenerate solution was avoided by mistakenly renormalizing $\mathbf{W}$ and $\mathbf{H}$ at every iteration, as commonly done in unpenalized NMF to eliminate scale indeterminacies. But this as a mater of fact changes the objective function value as evidenced by Eq. (8) and should thus be avoided. A standard and feasible solution to prevent from the latter degenerate solution would consist in penalizing the norm of the columns of $\mathbf{W}$. We resort to a simpler solution, described in the next section.

2.2. A scale-invariant measure of smoothness

The degenerate solution $\hat{\mathbf{W}} = \infty$, $\hat{\mathbf{H}} = 0$ is caused by the term in $\log h_{k(n-1)}$ in Eq. (7). We propose to simply discard this term, leading to the following ad-hoc penalty criterion

$$P_2(\mathbf{H}) = \sum_{k=1}^{K} \sum_{n=2}^{N} d_{IS}(h_{k(n-1)}|h_{kn}) \tag{9}$$

which penalizes large deviations of h_{kn} from $h_{k(n-1)}$, as measured by the IS divergence. The statistical interpretation of the penalty term is lost, but leads in turn to a well posed optimization problem. We define the following objective function

$$C_2(\mathbf{W},\mathbf{H}) = D_{IS}(\mathbf{V}|\mathbf{WH}) + \lambda P_2(\mathbf{H}) \tag{10}$$

which is scale-invariant, i.e., $C_2(\mathbf{W}\mathbf{\Delta}^{-1}, \mathbf{\Delta}\mathbf{H}) = C_2(\mathbf{W},\mathbf{H})$. A MM algorithm for minimizing criterion (10) is presented in the next section.

3. MM ALGORITHM FOR SMOOTH IS-NMF

We propose an iterative algorithm that updates $\mathbf{W}$ given $\mathbf{H}$ and $\mathbf{H}$ given $\mathbf{W}$. In the next subsection we first recall how to handle the unpenalized IS-NMF problem using auxiliary functions. Then we show in Section 3.2 how to include the smoothness penalty on $\mathbf{H}$ in the optimization.

3.1. Unpenalized case

In the unpenalized case ($\lambda = 0$), the updates of $\mathbf{W}$ and $\mathbf{H}$ are essentially the same, by symmetry of the factorization ($\mathbf{V} \approx \mathbf{WH}$ is equivalent to $\mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$ and the roles of $\mathbf{W}$ and $\mathbf{H}$ are simply exchanged), and because we are not making any assumption on the relative values of F and N. Because the objective function (2) separates into the columns of $\mathbf{H}$ (or the rows $\mathbf{W}$), we are essentially left with solving the problem

$$\min_{\mathbf{h}} C(\mathbf{h}) = D_{IS}(\mathbf{v}|\mathbf{Wh}) \text{ subject to } \mathbf{h} \geq 0 \tag{11}$$

rithm that we propose in this paper does not exploit any hidden data and is derived in a deterministic setting.

where $\mathbf{v} \in \mathbb{R}_+^F$, $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{h} \in \mathbb{R}_+^K$. We resort to a surrogate auxiliary function for iteratively solving this subproblem. The $\mathbb{R}_+^K \times \mathbb{R}_+^K \to \mathbb{R}_+$ mapping $G(\mathbf{h}|\tilde{\mathbf{h}})$ is said to be an *auxiliary function* to $C(\mathbf{h})$ if and only if 1) $\forall \mathbf{h} \in \mathbb{R}_+^K$, $C(\mathbf{h}) = G(\mathbf{h}|\mathbf{h})$, and 2) $\forall (\mathbf{h}, \tilde{\mathbf{h}}) \in \mathbb{R}_+^K \times \mathbb{R}_+^K$, $C(\mathbf{h}) \leq G(\mathbf{h}|\tilde{\mathbf{h}})$. The optimization of $C(\mathbf{h})$ can be replaced by iterative optimization of $G(\mathbf{h}|\tilde{\mathbf{h}})$. Indeed, any iterate $\mathbf{h}^{(i+1)}$ satisfying $G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)})$ produces a monotone algorithm (i.e., an algorithm which decreases the objective function at every iteration) as we have $C(\mathbf{h}^{(i+1)}) \leq G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) = C(\mathbf{h}^{(i)})$.

An auxiliary function to $C(\mathbf{h})$ is proposed in [10], where nonnegative linear regression with the IS divergence is considered for an image restoration problem. The auxiliary function is constructed by majorizing the convex part of the criterion (terms in $1/y$) using Jensen's inequality and majorizing the concave part of the criterion (terms in $\log y$) by their tangent. The resulting auxiliary function reads

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_k \frac{\tilde{h}_k^2}{h_k}\left(\sum_f w_{fk}\frac{v_f}{\tilde{v}_f^2}\right) + h_k\left(\sum_f \frac{w_{fk}}{\tilde{v}_f}\right) + cst \tag{12}$$

where $\tilde{v}_f = [\mathbf{W}\tilde{\mathbf{h}}]_f$ and cst denotes constant terms w.r.t $\tilde{\mathbf{h}}$. The auxiliary function separates into functions of its individual variables and its minimization w.r.t to $\mathbf{h}$ leads to the following MM update

$$h_k = \tilde{h}_k \sqrt{\frac{\sum_f w_{fk} v_f/\tilde{v}_f^2}{\sum_f w_{fk}/\tilde{v}_f}}. \tag{13}$$

As a matter of fact, removing the square root from the latter expression still leads to a monotone algorithm, though it does not correspond to a MM algorithm anymore, see [10].

3.2. Penalized case

We now get back to our original problem of minimizing the objective function $C_2(\mathbf{W}, \mathbf{H})$ defined at Eq. (10), and more precisely to its minimization w.r.t $\mathbf{H}$, given $\mathbf{W}$. We propose to update the columns $\mathbf{h}_n$ of $\mathbf{H}$ sequentially. For $n = 2, \ldots, N-1$, the individual contribution of $\mathbf{h}_n$ to the objective function writes

$$C_P(\mathbf{h}_n) = D_{IS}(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n) + L(\mathbf{h}_n; \mathbf{h}_{n-1}, \mathbf{h}_{n+1}), \tag{14}$$

$$L(\mathbf{h}_n; \mathbf{h}_{n-1}, \mathbf{h}_{n+1}) = \lambda \sum_k d(h_{k(n-1)}|h_{kn}) + d(h_{kn}|h_{k(n+1)}). \tag{15}$$

An auxiliary function to the penalized objective function $C_P(\mathbf{h}_n)$ is readily obtained as

$$G_P(\mathbf{h}_n|\tilde{\mathbf{h}}_n) = G(\mathbf{h}_n|\tilde{\mathbf{h}}_n) + L(\mathbf{h}_n; \mathbf{h}_{n-1}, \mathbf{h}_{n+1}). \tag{16}$$

Algorithm 1 Smooth IS-NMF

Input : nonnegative matrix $\mathbf{V}$
Output : nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$
Initialize $\mathbf{W}$ and $\mathbf{H}$ with nonnegative values
Compute $\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$
for $i = 1 : n_{iter}$ **do**
  `%% Update H %%`
  $\mathbf{G}^- = \mathbf{W}^T(\mathbf{V}.\hat{\mathbf{V}}^{\cdot -2})$; $\mathbf{G}^+ = \mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1})$
  Update $\mathbf{h}_1$ (requires solving K order 2 polynomials)
  for $n = 2 : N-1$ **do**
  $\mathbf{h}_n^{(i)} = \sqrt{[\mathbf{g}_n^-.(\mathbf{h}_n^{(i-1)})^{.2} + \lambda \mathbf{h}_{n-1}^{(i)}]./[\mathbf{g}_n^+ + \lambda./\mathbf{h}_{n+1}^{(i-1)}]}$
  end for
  Update $\mathbf{h}_N$ (requires solving K order 2 polynomials)
  Compute $\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$
  `%% Update W %%`
  $\mathbf{W} \leftarrow \mathbf{W}.[(\hat{\mathbf{V}}^{\cdot -2}.\mathbf{V})\,\mathbf{H}^T]./[\hat{\mathbf{V}}^{\cdot -1}\,\mathbf{H}^T]$
  Compute $\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$
  Normalize $\mathbf{W}$ and $\mathbf{H}$ together
end for

Again, the auxiliary function separates into functions of its individual variables and its minimization leads to

$$h_{kn} = \sqrt{\frac{\tilde{h}_{kn}^2 \sum_f w_{fk} v_{fn}/\tilde{v}_{fn}^2 + \lambda h_{k(n-1)}}{\sum_f w_{fk}/\tilde{v}_{fn} + \lambda/h_{k(n+1)}}}. \tag{17}$$

The latter expression only holds for $n = 2, \ldots, N-1$. At the borders, for $n = 1, N$, updating h_{kn} is easily shown to amount to solving an order 2 polynomial with only one nonnegative root, see the MATLAB code available at [9]. Algorithm 1 recapitulates the MM approach to smooth IS-NMF. Most operations can be efficiently vectorized, leading to fast and simple implementations.

4. RESULTS

Like in [4] we consider for illustration the decomposition of a 108 seconds-long music excerpt from *My Heart (Will Always Lead Me Back To You)* recorded by Louis Armstrong and His Hot Five in the twenties. The band features a trumpet, a clarinet, a trombone, a piano and a double bass. A STFT $\mathbf{X} = [x_{fn}]$ of the original signal x (sampled at 11kHz) was computed using a sinebell analysis window of length $L = 256$ (23 ms) with 50 % overlap, leading to $F = 129$ frequency bins and $N = 9312$ frames. To illustrate the effect of the regularization of the rows of $\mathbf{H}$ we perform the following experiment. First we run unpenalized IS-NMF with $K = 10$ and 5000 iterations, retaining the solution with lowest final cost value among ten runs from different random initializations. Then we run smooth IS-NMF with $\mathbf{W}$ and $\mathbf{H}$ respectively *fixed* and *initialized* to the unpenalized solution. Fig. 1 reports results with different values of λ.

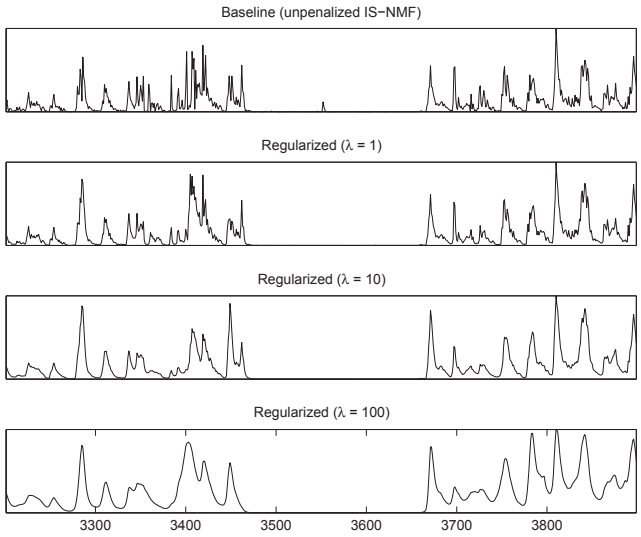

Fig. 1. Effect of regularization for $\lambda = \{1, 10, 100\}$. We display a segment of one of the rows of $\mathbf{H}$, corresponding to the activations of the accompaniment (piano and double bass). A trumpet solo occurs between frames 3470 and 3660, where the accompaniment vanishes; the regularization smoothes out coefficients with small energies that remain in unpenalized IS-NMF.

Besides this experiment, we ran a full decomposition of the signal with smooth IS-NMF (including the estimation of $\mathbf{W}$), using $K = 10$ and $\lambda = 25$. 1000 iterations of unpenalized and smooth IS-NMF take respectively 516 and 546 seconds (CPU time) on an average PC. Component STFT estimates were reconstructed through Wiener filtering $\hat{c}_{k,fn} = [w_{fk}h_{kn}/\hat{v}_{fn}]x_{fn}$ and then inverted to time domain with overlap-add. Audio samples and representations of the Wiener masks are available at [9]. It is shown that the algorithm singles out most of the accompaniment and hiss noise in separate (large band) components. The remaining components have a pitched structure and, when added together, reconstruct most of the leading instruments (trumpet, clarinet). The inclusion of the smoothness penalty was found to lead to more significant decompositions, and in turn more pleasant to listen to. The decomposition may be used for music editing tasks such as denoising (elimination of the hiss noise) and upmix (mono to stereo conversion), visit [9] for such audio examples.

5. CONCLUSIONS

In this paper we have proposed a MM algorithm for smooth IS-NMF. The algorithm is computationally friendly and was shown to produce satisfying results for music editing tasks. As for perspective we intend to incorporate our algorithm to nonnegative tensor factorization settings for multichannel source separation [11]. This will allow thorough evaluation of the system on a specific task (for which standard test data and evaluation criteria exist) and will in particular allow to quantify the influence of λ on the results.

6. REFERENCES

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

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, Oct. 2003.

[3] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, Mar. 2007.

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

[5] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, 2004.

[6] 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 ICASSP*, Toulouse, France, May 2006.

[7] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. IEEE ICASSP*, Las Vegas, Nevada, USA, Apr. 2008.

[8] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *In Proc. Interspeech 2008*, Brisbane, Australia, Sep. 2008.

[9] C. Févotte, "ICASSP'2011 companion webpage," `http://perso.telecom-paristech.fr/~fevotte/Samples/icassp11/`.

[10] Y. Cao, P. P. B. Eggermont, and S. Terebey, "Cross Burg entropy maximization and its application to ringing suppression in image reconstruction," *IEEE Trans. Image Processing*, vol. 8, no. 2, Feb. 1999.

[11] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues," in *Proc. 7th International Symposium on Computer Music Modeling and Retrieval (CMMR'2010)*, Malaga, Spain, June 2010.