A TEMPERING APPROACH FOR ITAKURA-SAITO NON-NEGATIVE MATRIX FACTORIZATION. WITH APPLICATION TO MUSIC TRANSCRIPTION

Nancy BERTIN, Cédric FÉVOTTE, Roland BADEAU

CNRS LTCI - TELECOM ParisTech (ENST)
46 rue Barrault 75634 PARIS Cedex 13, France

ABSTRACT

In this paper we are interested in non-negative matrix factorization (NMF) with the Itakura-Saito (IS) divergence. Previous work has demonstrated the relevance of this cost function for the decomposition of audio power spectrograms. This is in particular due to its scale invariance, which makes it more robust to the wide dynamics of audio, a property which is not shared by other popular costs such as the Euclidean distance or the generalized Kulback-Leibler (KL) divergence. However, while the latter two cost functions are convex, the IS divergence is not, which makes it more prone to convergence to irrelevant local minima, as observed empirically. Thus, the aim of this paper is to propose a tempering scheme that favors convergence of IS-NMF to global minima. Our algorithm is based on NMF with the beta-divergence, where the shape parameter beta acts as a temperature parameter. Results on both synthetical and music data (in a transcription context) show the relevance of our approach.

Index Terms— Non-negative matrix factorization (NMF), Itakura-Saito (IS) divergence, beta divergence, music transcription.

1. INTRODUCTION

Non-negative matrix factorization (NMF) is a now popular dimension reduction technique, employed for non-subtractive, parts-based representation of nonnegative data. Its use has dramatically grown in various signal processing applications over the last years, among which we can cite learning parts of faces and semantic features of text [1] or polyphonic music transcription [2, 3]. Given a data matrix V of dimensions $F \times N$ with non-negative entries, NMF is the problem of finding a factorization

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

where W and H are non-negative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $FK + KN \ll FN$, hence reducing the data dimension. The factorization (1) is generally obtained by minimizing a cost function defined by

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

where $d(x|y)$ is a function of two scalar variables. d is typically non-negative and takes value zero if and only if (iff) $x = y$. The most popular cost functions for NMF are the Euclidean (EUC) distance, defined as

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

and the generalized Kullback-Leibler (KL) divergence, defined as

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

Those two cost functions, as NMF itself, were particularly popularized by Lee and Seung, see, e.g, [1], which described multiplicative update rules under which $D(V|WH)$ is shown nonincreasing while ensuring non-negativity of W and H. Despite the popularity of these updates, the literature is flourishing on other algorithms among which we can cite alternating least squares, projected gradient, conjugate gradient, quasi-Newton optimization, see, e.g., [4].

We are here concerned with NMF using the Itakura-Saito (IS) cost function, expressed as

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

In [5] we have suggested the IS cost function to be better suited to the NMF of music power spectrograms than the usual EUC or KL costs, as it was shown to better bring out the semantics of audio. We have described how IS-NMF may be expressed as a maximum likelihood (ML) estimation in two different frameworks: it can either be seen as variance estimation in superimposed Gaussian components or as estimation of W and H from their product observed in multiplicative Gamma noise (see [5] and references therein). An interesting property of IS-NMF is scale-invariance, i.e, $d_{IS}(\lambda x|\lambda y) = d_{IS}(x|y)$, a property which is not shared by the two other cost functions; as such, we have $d_{EUC}(\lambda x|\lambda y) = \lambda^2\, d_{EUC}(x|y)$ and $d_{KL}(\lambda x|\lambda y) = \lambda\, d_{KL}(x|y)$. The scale invariance means that same relative weight is given to small and large coefficients of V in cost function (2), in the sense that a bad fit of the factorization for a low-power coefficient $[V]_{fn}$ will cost as much as a bad fit for a higher power coefficient $[V]_{f'n'}$. The scale invariance of the IS divergence is relevant to decomposition of audio spectra, which typically exhibit exponential power decrease along frequency f and also usually comprise low-power transient components such as note attacks, together with higher power components such as tonal parts of sustained notes.

However, a property shared by the EUC and KL costs $d(x|y)$, and not by the IS cost, is convexity. This means that the cost $D(V|WH)$ is at least convex with respect to (wrt) either W or H (see Section 2.3). In contrast, nothing can be said about IS-NMF. However, we can intuitively expect that its non-convex form

The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content KSPACE, and by the French GIP ANR under contract ANR-06-JCJC-0027-01, Décomposition en Éléments Sonores et Applications Musicales DESAM.

makes it more prone to local minima, as also empirically observed in previous work [6]. As such, the aim of this paper is to propose a tempering scheme that favors convergence of IS-NMF to global minima. Our algorithm is based on NMF with the β-divergence [7], which takes the EUC, KL and IS costs as special cases, where the shape parameter β acts as a temperature parameter. The idea is simply to start from a criterion with less local minima (such as $D_{EUC}(V|WH)$ or $D_{KL}(V|WH)$) and gradually reshape it to the correct criterion $D_{IS}(V|WH)$.

The paper is organized as follows. In Section 2, we provide a detailed study of the properties of the β-divergence wrt β (a study which was to our best knowledge not yet available) and describe a NMF algorithm under this cost that was previously proposed by [7]. In particular, we describe our novel tempering procedure in Section 2.5. Section 3 presents some results on synthetic data, on which our tempered algorithm is shown to attain more frequently lower cost values than the standard IS-NMF algorithm. Then, our algorithm is validated on a real-case music transcription problem in Section 4, where we show improved precision and recall rates. Section 5 provides conclusive remarks.

2. NMF WITH β-DIVERGENCE

2.1. The β-divergence

The β-divergence introduced by Eguchi and Kano in [8] is defined as

$$d_\beta(x|y) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\,\log\frac{x}{y} + (y-x) & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases}$$

As observed in [7], the IS divergence is a limit case of the β-divergence. [8] assume $\beta > 1$, but the definition domain can very well be extended to $\beta \in \mathbb{R}$. The β-divergence is shown to be continuous in β by using the identity $\lim_{\beta\to 0}(x^\beta - y^\beta)/\beta = \log(x/y)$. EUC distance is obtained for $\beta = 2$, so that the β-divergence is inclusive for our three choices of NMF costs: EUC, KL and IS.

2.2. Study of variations

Let us study $d_\beta(x|y)$ as a function of y (remember that x acts as data). Its first and second-order derivatives write

$$\nabla_y d_\beta(x|y) = y^{\beta-2}(y-x), \quad (6)$$

$$\nabla_y^2 d_\beta(x|y) = y^{\beta-3}\left((\beta-1)y + (2-\beta)x\right). \quad (7)$$

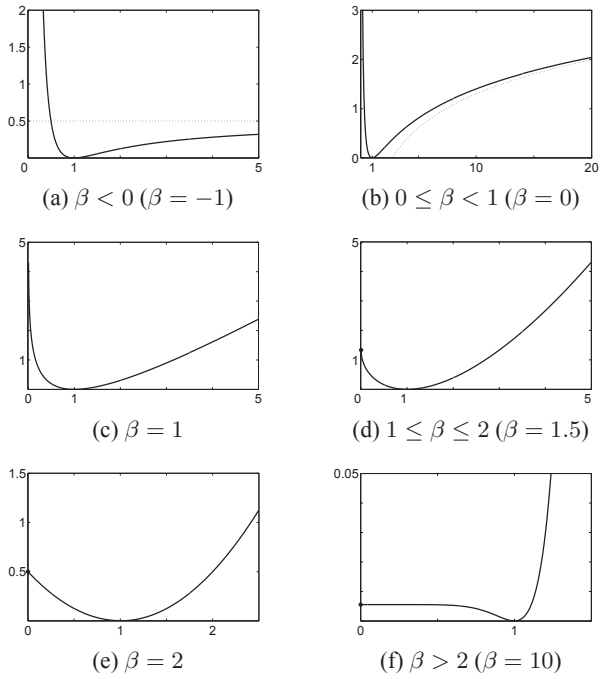
The next properties follow :

- $d_\beta(x|y)$ has a single minimum in $y = x$ and increases with $|y-x|$. This justifies its relevance as a measure of fit.
- $d_\beta(x|0)$ is finite iff $\beta \geq 1$.
- $d_\beta(x|y)$ is convex on $\mathbb{R}_+$ iff $1 \leq \beta \leq 2$.

Figure 1 represents typical behaviours of the β-divergence.

2.3. β-divergence between matrices

The cost function $D_\beta(V|WH)$ is not convex in general wrt the pair (W,H), even if the cost $d_\beta(x|y)$ is convex (wrt y). However, when $d_\beta(x|y)$ is convex, D_β is at least convex as a function of W (resp. H) with fixed V and H (resp. W), because it is expressed as a sum of convex functions composed with linear functionals (see Eq. (2)).



(a) $\beta < 0$ ($\beta = -1$) (b) $0 \leq \beta < 1$ ($\beta = 0$)

(c) $\beta = 1$ (d) $1 \leq \beta \leq 2$ ($\beta = 1.5$)

(e) $\beta = 2$ (f) $\beta > 2$ ($\beta = 10$)

Fig. 1. β-divergence as a function of y (with $x = 1$). Subfigures (e), (c) and (b) show EUC, KL and IS costs respectively.

2.4. NMF with the β-divergence

Computing the gradient $\nabla_H D_\beta(V|WH)$ (resp. $\nabla_W D_\beta(V|WH)$) using Eq. (6), and multiplying H (resp. W) at previous iteration by the ratio of the negative and positive parts of the gradient, we obtain the following alternate multiplicative algorithm [7] :

$$H \leftarrow H \otimes \frac{W^T(V \otimes (WH)^{\cdot[\beta-2]})}{W^T((WH)^{\cdot[\beta-1]})} \quad (8)$$

$$W \leftarrow W \otimes \frac{(V \otimes (WH)^{\cdot[\beta-2]})H}{((WH)^{\cdot[\beta-1]})H^T} \quad (9)$$

where $\otimes$ and $\oslash$ denote (Hadamard) entrywise product and division respectively, the fraction is also entrywise and $A^{\cdot[n]}$ denotes the matrix with entries $[A]_{ij}^n$. For $\beta = 1, 2$, we obtain Lee and Seung's original algorithm. Using convexity of $d_\beta(x|y)$, monotonicity of the criterion under the latter rules can be shown for $1 \leq \beta \leq 2$ [9]. In other cases, this monotonicity was observed in practice, though not proven.

2.5. Tempering algorithm

As we observed in practice that IS-NMF is more prone to local minima [5, 6], we now describe a tempering scheme that favors convergence of IS-NMF to global minima. It simply consists of using β as a temperature parameter, which is set to a value between 1 and 2 in the first iterations (where the cost $D_\beta(V|WH)$ is at least convex wrt to either W or H) and gradually decrease it to the target cost, i.e, IS in our case, obtained for $\beta = 0$. As such, we simply apply update rules (8) and (9), with β being a function of the iteration number. More precisely, we use the template described by Fig. 2; β takes value β_i during n_i iterations, then starts to decrease following

a cosine during n_d more iterations, until it finally reaches its target value, to which it remains fixed during the last n_e iterations. In the following, the prefix and superscript $(\beta_e \to \beta_i)$ will refer to one particular template, as described by its initial and final values of β; the other parameters n_i, n_d and n_e being fixed to arbitrary values.

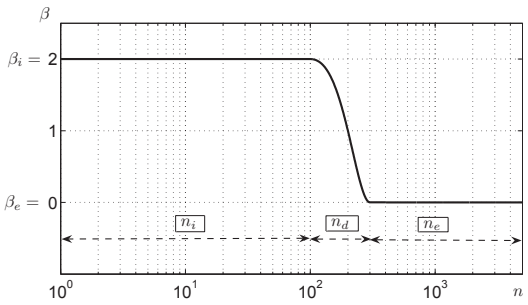


Fig. 2. β wrt the number of iterations n.

3. SIMULATIONS

In this part, we wish to investigate on local minima of β-divergences and on the convergence of various NMF algorithms, including the tempering algorithm proposed in section 2.5.

3.1. Experimental setup

We use the multiplicative noise model from [5] as a generative model, to obtain synthetic data for which minimizing IS divergence is relevant in a statistical sense. Ground truth matrices W_0 and H_0 are chosen randomly. Multiplicative noise E is generated according to the Gamma distribution with shape parameter and mean 1. The matrix V_0 to factorize is then built as $V_0 = (W_0 H_0) \otimes E$.

We then draw a random initialization (W_i, H_i) and use it to compute factorizations $(W^{\beta_i \to \beta_e}, H^{\beta_i \to \beta_e})$ for different values of β_i and β_e. For each realization V_0, we draw 100 random initializations to repeat this process. This experiment is carried out for 10 different realizations V_0.

Dimensions of matrices are chosen smaller than for a real application, but their ratios are compatible with real cases. For each realization and initialization, the following pairs (β_i, β_e) are tested: (10,0),(2,0),(1,0),(0,0). The following table sums up the parameters used in the simulations.

Parameter	F	K	N	n_i	n_d	n_e
Value	50	5	500	100	200	4700

3.2. Results

We consider one run of a given algorithm as a success if the inequality $D_{IS}^{\beta_e \to \beta_i}(V_0|WH) \leq D_{IS}^{0 \to 0}(V_0|WH)$ is verified. Table 1 presents the results of convergence of the tested algorithms.

$\beta_i \to \beta_e$	10→0	2→0	1→0
Success rate	18	100	98

Table 1. Success rate (%).

For illustration purpose, we also represent on figure 3 the evolution of IS divergence wrt the iteration number, for selected significant runs and for $\beta_e = 0$ and $\beta_i = 2, 0$.

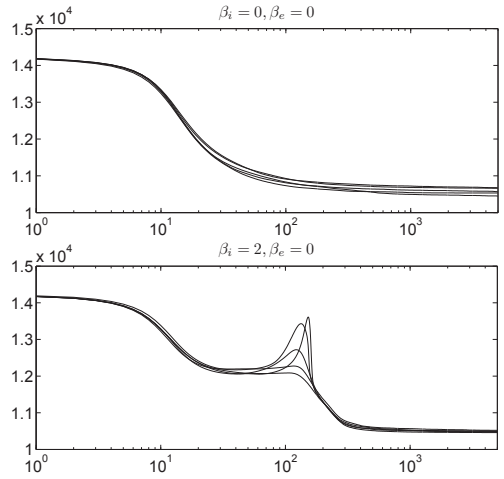


Fig. 3. IS divergence vs. number of iterations.

3.3. Discussion

As shown in Table 1, the tempering approach with $\beta_e = 2$ is particularly efficient to reach lower final error than $(0 \to 0)$-NMF (i.e., classical IS-NMF) when the data follow the underlying statistical model. Tempering with $\beta_e = 1$ is equally performant, but it is important to stress that $D_{IS}^{2 \to 0}(V_0|WH) \leq D_{IS}^{1 \to 0}(V_0|WH)$ in most cases. This relation, added to the relatively poor performance of $(10 \to 0)$-tempering, suggests the importance of the "convexity zone" $(1 \leq \beta \leq 2)$ and its use in the first iterations of the algorithm.

Though no clearly identifiable local minimum appear in this test set, final values reached by $(0 \to 0)$-NMF are more scattered than those obtained with $(2 \to 0)$-tempering. This may confirm the existence of local minima, observed for typical NMF-costs in [6].

$(10 \to 0)$-NMF performs relatively bad. Several explanations may be invoked. First, the non-convexity of the cost for $\beta > 2$ let us presume a negative effect of such a choice. The synthetic material used here has low dynamics (compared to audio signals), putting less importance on the property $d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y)$ which is mostly useful for wide-dynamics signals. Moreover, as visible on Figure 1, d_{10} diverge very quickly to infinity when y grows, making the algorithm diverge if initialization falls into too high values. Eventually, as no local minimum was observed, and considering the relatively low dimensions F, K, N, we can guess that a tempering approach with a too high temperature may be unnecessary, nay noxious.

Despite the lack of proof, we verify that $(0 \to 0)$-NMF exhibits nonincreasing IS divergence. On the contrary, we observe phases where IS divergence increases in $(2 \to 0)$-NMF, which is the expected behaviour of a tempering algorithm. We notice that the higher the "local maximum" around iteration 100, the lower the final cost. This validates the use of a tempering approach: by allowing IS-divergence to increase, it allows to find NMF solutions that are not reachable by usual approaches.

4. AN AUDIO APPLICATION

In this section, we consider the music transcription task as a "wav-to-midi" task, with single-channel music as input and a midi-file including note onsets, offsets and pitch as output.

Let us take V as a time-frequency representation of a polyphonic music signal, F being the number of frequency bins and N the number of time frames. [2] suggests that under an adequate additivity hypothesis, the NMF $V \approx WH$ may give a separation of the K notes appearing in the input signal, by interpreting W as a basis of note pseudo-spectra and H as the corresponding time envelopes. Following this idea, we proposed in [3] a full "wav-to-midi" system, by performing a monopitch estimation on columns of W and an energy thresholding on lines of H. Transcription performance, expressed in terms of precision and recall at the note level, is then a new way to evaluate NMF costs and algorithms.

We perform midi transcriptions of six 30-second excerpts of real piano music, recorded from a Yamaha DisKlavier, providing a midi reference of the piece. Each piece is factorized with 10 different random initializations. Midi transcription and reference are then compared in order to get averaged transcription scores. Table 2 sums up the results for six algorithms with fixed and variable β.

$\beta_i \to \beta_e$	10→0	2→0	1→0	0→0	2→2	1→1
Precision	**83.4**	73.6	69.7	77.2	67.8	70.5
Recall	**79.2**	79.2	73.6	70.6	73.6	65.5
F-measure	**81.3**	76.3	71.6	73.7	70.6	67.9

Table 2. Averaged transcription performance (%).

Several observations can be made. First, all algorithms involving IS divergence, with or without tempering, outperform EUC and KL-NMF. The tempering algorithm with $\beta_i = 10$ gives the best transcription scores, despite the non-convexity of the β-divergence with $\beta > 2$. A possible explanation for this fact is the property $d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y)$, meaning that first iterations of $(10 \to 0)$-NMF put a strong importance on high-energy components. Since analyzed data possess the typical wide dynamics of audio, this can be decisive, compared to synthetic data of previous section. Moreover, the dimensions of the problem make it more likely to possess numerous local minima, motivating the use of a higher initial temperature.

Tempering with $\beta_i = 2$ is the second best in our test. Yet, $(0 \to 0)$-NMF reaches better final cost values in all cases. No clear correlation between final cost and transcription performance is observed.

Another noticeable result to mention is the existence of a few severe failures for $(0 \to 0)$-NMF (for which the F-measure is below 10%). On the contrary, EUC-NMF, KL-NMF and tempering algorithms seem to be unexposed to this phenomenon. However, if we exclude pathological cases from transcription scores computation, $(0 \to 0)$-NMF performs as good as tempering.

This study confirms how suitable IS-divergence is in an audio processing task, and the improvement brought by the tempering approach.

5. CONCLUSIONS

In this paper, we motivated the choice of Itakura-Saito divergence for NMF-based audio processing, by a theoretical and experimental study that confirms its interest. We proposed a new approach, inspired from tempering, to minimize this divergence in the NMF framework.

Experiments on synthetic material confirm our intuition that the tempering approach, and in particular the exploitation of convexity of β-divergences with $1 \leq \beta \leq 2$ in the first iterations of the algorithm, allows to reach better final cost values, thus avoiding local minima of the IS-divergence. The optimal β_i seems however signal-dependent and its choice remains an open question. Good properties of IS-NMF and tempering approaches are confirmed on a real music test set.

The best music transcription performance, which is obtained by algorithms with IS divergence and/or tempering, does generally not correspond to the lowest cost value. This points out the inevitable limitations of any numerical criterion. Numerous improvements of NMF-based music transcription are obtained by constraining the solution to possess supplementary properties such as sparsity, harmonicity or temporal smoothness. We can thus expect better transcription performances by integrating such constraints to NMF with IS divergence, which was done for instance in [5] for another audio application.

6. REFERENCES

[1] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[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, pp. 177–180.

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, Apr. 2007.

[4] A. Cichocki, R. Zdunek, and S. Amari, "Nonnegative matrix and tensor factorization," *IEEE Signal Processing Magazine*, pp. 142 – 145, Jan. 2008.

[5] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis," *Neural Computation*, 2008, In press. `http://www.tsi.enst.fr/~fevotte/TechRep/techrep08_is-nmf.pdf`.

[6] N. Bertin and R. Badeau, "Initialization, distances and local minima in audio applications of the non-negative matrix factorization," in *Proc. of Acoustics'08, JASA*, 2008, `http://perso.telecom-paristech.fr/~nbertin/publis/acoustics08.pdf`.

[7] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, Charleston SC, USA, Mar. 2006, pp. 32–39.

[8] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," Tech. Rep., Tokyo Institute of Statistical Mathematics, 2001, available at `http://www.ism.ac.jp/~eguchi/pdf/RobustifyMLE.pdf`.

[9] R. Kompass, "A generalized divergence measure fon nonnegative matrix factorization," *Neural Computation*, vol. 19(3), pp. 780–791, 2003.