# CAUCHY NONNEGATIVE MATRIX FACTORIZATION

*Antoine Liutkus*[1] *Derry Fitzgerald*[2] *Roland Badeau*[3]

[1]Inria, Speech Processing Team, Villers-lès-Nancy, France
[2]NIMBUS Centre, Cork Institute of Technology, Ireland
[3]Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, France

**ABSTRACT**

Nonnegative matrix factorization (NMF) is an effective and popular low-rank model for nonnegative data. It enjoys a rich background, both from an optimization and probabilistic signal processing viewpoint. In this study, we propose a new cost-function for NMF fitting, which is introduced as arising naturally when adopting a Cauchy process model for audio waveforms. As we recall, this Cauchy process model is the only probabilistic framework known to date that is compatible with having additive magnitude spectrograms for additive independent audio sources. Similarly to the Gaussian power-spectral density, this Cauchy model features time-frequency nonnegative scale parameters, on which an NMF structure may be imposed. The Cauchy cost function we propose is optimal under that model in a maximum likelihood sense. It thus appears as an interesting newcomer in the inventory of useful cost-functions for NMF in audio. We provide multiplicative updates for Cauchy-NMF and show that they give good performance in audio source separation as well as in extracting nonnegative low-rank structures from data buried in very adverse noise.

***Index Terms*—NMF, audio, Cauchy distribution, robust estimation, probabilistic modeling**

## I. INTRODUCTION

When facing tabular data, gathered as a large $F \times T$ matrix $V$, a recurring approach is to decompose it using a low-rank model, i.e. decompose it as the interaction of only a small number of patterns or *components*. In practice, the topic of Matrix Factorization (MF) aims to approximate $V$ as the product

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

where $W$ and $H$ are of dimensions $F \times K$ and $K \times T$, respectively, and $K \ll \max(F, T)$ is the number of components. This problem has been addressed by the scientific community for decades, and depending on how we define or constrain $W$ and $H$, a whole new range of solutions and approaches emerge. For instance, truncated singular value decomposition or QR factorization, and all related algorithms [1] provide an efficient solution to (1).

MF has found a renewed interest in the last decade in the particular case where both the input matrix $V$ and the latent factors $W$ and $H$ are taken as nonnegative, i.e. with positive entries. In that case, model (1) is called a Nonnegative MF (NMF), pioneered in [2]. Many matrices encountered have positive entries, such as images, histograms or audio spectrograms [3] and having nonnegative $W$ and $H$ forces the decomposition to only feature additive contributions of a few elements that often bear a physical meaning.

From an optimization perspective, learning the factors $W$ and $H$ of the model $V \approx WH$ is done by minimizing a data-fit *cost-function* $d(V \mid WH)$. Whereas early studies such as [2] considered the squared error and the Kullback-Leibler (KL) divergence, many others where proposed later, notably the family of $\beta$-divergences $d_\beta$, subsuming them as special cases [4], [5]. In another vein, robust cost-functions were also proposed recently, enabling the estimation to behave well in the presence of outliers in $V$ (see, e.g. [6], [7] and references therein). Additionally, a regularization term may be included, that favors some features that may be expected in the latent factors $W$ and $H$, such as smoothness [8], [9], [10] or sparsity [11], [12].

Whereas NMF was initially settled in such an optimization perspective, it was early realized that depending on the criteria considered, different probabilistic interpretations would arise, actually whenever the cost function is amenable to a negative log-likelihood. This happens for instance with the KL cost function, that emerges naturally in a Poisson probabilistic framework [13]. Similarly, the Itakura-Saito divergence is equivalent to variance fitting in a centered complex isotropic Gaussian model [14], [15]. In this paradigm, the regularization terms concerning $W$ and $H$ are often interpreted as providing prior distributions on those parameters [16]. Interestingly, these probabilistic interpretations not only lead to a better understanding of the NMF model, but also to alternative ways of estimating the parameters different from those considered in the initial study [2]. For instance, exploiting the Expectation-Maximization algorithm was found useful in this context [17], as well as Markov Chain Monte Carlo (MCMC) sampling [18], [19].

In this study, we propose a new particular probabilistic interpretation of NMF modeling: the cost function we consider is the negative log-likelihood in an isotropic Cauchy distribution. As we show, this particular and original choice has many interesting features. First, it comes as the only way we are aware of to justify NMF models for signals having additive *magnitude* spectrograms, if we want the probabilistic interpretation to be coherent all the way down to the actual waveforms. Choosing the Cauchy cost function provides the corresponding optimal way to learn a NMF model. As we show, this cost-function differs from the KL divergence commonly used in the same setting, notably in the Probabilistic Latent Component Analysis literature (PLCA, see e.g. [20]). The Cauchy-NMF thus comes as an interesting alternative to KL-NMF.

Beyond the mere theoretical appeal of providing a new coherent framework for audio, we also show that the Cauchy model brings in a second and more practical advantage, which is robustness to outliers. Indeed, Cauchy is a special case of $\alpha$-stable distribution, which have been a central topic in the field of robust statistics [21], [22]. Cauchy corresponds to $\alpha = 1$. From this viewpoint, Cauchy-NMF permits in practice to build low-rank approximations of large matrices that feature heavy outliers. It thus comes as an interesting alternative to, say, robust Principal Component Analysis (RPCA [23]) in the sense that the underlying factors $W$ and $H$ are further constrained to be nonnegative. We show that it is competitive to other such robust NMF algorithms [6], [7], and behaves remarkably well in the presence of very adverse $\alpha$-stable noise in a signal enhancement context.

This paper is structured as follows. In section II, we motivate the Cauchy NMF framework for audio processing applications. We

This work was partly supported under the research programme EDiSon3D (ANR-13-CORD-0008-01) funded by ANR, the French State agency for research.

show that it comes as a principled model for time-series with a large dynamic range, which is common in music. In section III, we propose a computationally effective algorithm for Cauchy NMF. Finally, we evaluate the method in section IV.

## II. CAUCHY NMF FOR AUDIO

### II-A. Cauchy processes

Let $\tilde{x}$ be the complete waveform of an audio signal in the time domain. This waveform is split into overlapping frames and the Fourier transform of each one of them is taken, yielding the so-called Short-Term Fourier Transform (STFT) $x$ of $\tilde{x}$. It is an $F \times T$ matrix with complex entries $x(f,t)$. $F$ is the number of frequency bands and $T$ is the number of times frames. In audio, where the waveforms are real, we assume that only the non-redundant frequency information is kept in the STFT $x$.

In [24], the $\alpha$-harmonizable model was proposed, that includes Cauchy as the special case $\alpha = 1$. First, all frames are assumed independent as commonly done in audio, notwithstanding the overlap between adjacent frames. Then, the waveform of each frame is assumed to be the outcome of a stationary Cauchy process. Whereas assuming local stationarity is common in audio, assuming a Cauchy distribution is less usual: the well-documented Gaussian assumption is more frequent there.

While Gaussian processes come with a very elegant theory (see e.g. [25], [26]), Cauchy processes also have many desirable properties. First, adopting a Cauchy process model permits large deviations in the observed outcomes, thus allowing for data with a large dynamic range such as audio [27], [28], [24]. Second, we claim that they address one important drawback of the Gaussian model: its lack of robustness from a parameter estimation perspective. Indeed, an observation that is far from the current belief will induce significant changes in a Gaussian model. However, this spurious observation may be caused not only by model discrepancies that would justify the update, but also by outliers in the data, such as those caused by sources interferences in a source separation perspective. A more regularized pace for model updates may thus be preferable not to get stuck in sub-optimal estimates. Such a regularization can be induced by the heavy-tail Cauchy distribution.

Interestingly, assuming both a stationary and Cauchy distributed waveform $\tilde{x}$ is equivalent to assuming independent STFT entries $x(f,t)$, with each one distributed with respect to an isotropic complex Cauchy distribution (noted $\mathcal{C}_c$). We call this a Cauchy-harmonizable model, or a Cauchy process for short, as a special case of the $\alpha$-harmonizable family introduced in [24]:

$$\tilde{x} \text{ Cauchy (locally stationary) process} \Leftrightarrow \begin{cases} \text{all } x(f,t) \text{ independent} \\ x(f,t) \sim \mathcal{C}_c\left(\sigma(f,t)\right) \end{cases} \tag{2}$$

In (2), $\sigma(f,t)$ is called a *scale parameter*. This equivalence for $\alpha$-stable processes between stationarity and an independently distributed isotropic spectral representation is demonstrated e.g. in [29, th. 6.5.1]. The Cauchy distribution is only a special case of this result.

An interesting property of Cauchy processes is their *stability* property. It means that if $J$ signals $s_j$ are Cauchy proccesses, so will be their sum. More precisely, if

$$\forall j,\ s_j(f,t) \sim \mathcal{C}_c\left(\sigma_j(f,t)\right)$$

are the STFTs of $J$ independent Cauchy (harmonizable) processes called *sources*, then their sum $x$, the *mixture*, is distributed as[1]:

$$x(f,t) \triangleq \sum_j s_j \sim \mathcal{C}_c\left(\sum_j \sigma_j(f,t)\right), \tag{3}$$

[1] $\triangleq$ stands for a definition.

so that its scale parameters $\sigma$ are given by:

$$\sigma(f,t) = \sum_j \sigma_j(f,t). \tag{4}$$

In other words, the scale parameters $\sigma_j$ of Cauchy sources add up to form the scale parameters $\sigma$ of their mixture. Let $p$ and $p_j$ denote the modulus of the observed STFTs of $x$ and $s_j$ respectively, also called the *magnitude spectrograms*. Skipping the details, it can basically be shown [30] that they form asymptotically unbiased estimates of the Cauchy scale parameters $\sigma$ and $\sigma_j$ in (4), up to a multiplicative constant independent of the signal. This leads to:

$$p(f,t) \approx \sum_j p_j(f,t), \tag{5}$$

which is often taken as a starting point in many audio processing studies (see e.g. [31]). To put it simply, (5) means that the magnitude spectrograms of the sources add up to form that of the mixture. As far as we know, only Cauchy processes have this property. Interestingly enough, we also have [24]:

$$\mathbb{E}\left[s_j(f,t) \mid x(f,t), \{\sigma_j\}_j\right] = \frac{\sigma_j(f,t)}{\sum_{j'} \sigma_{j'}(f,t)} x(f,t), \tag{6}$$

which means that if we know only the mixture and the scale parameters, we can estimate the sources through a soft TF masking. Doing so is furthermore the optimal way to proceed in a posterior expectation sense. A practical estimate $\hat{s}_j$ of $s_j$ is hence obtained by replacing the true scale parameters by the magnitude spectrograms $p_j$. This strategy has long been known to provide excellent performance in many audio processing studies, even if no theoretical interpretation of this fact was available until recently.

To summarize, the Cauchy process model puts together robust signal processing tools through the use of the heavy-tail Cauchy distribution [21], [27], and the efficiency of TF masking, that was known to be theoretically grounded only for wide-sense stationary signals until recently [24].

### II-B. The Cauchy NMF model

In the literature, the $\alpha$-harmonizable model with $\alpha = 2$ is called the Local Gaussian Model (LGM [32], [33], [26]). The scale parameters, termed Power Spectral Densities (PSDs) are in that case denoted as $\sigma_j^2$. They correspond to variances and are thus nonnegative. A popular model for audio sources is to express their PSDs as single spectral nonnegative *patterns* $W_j(f)$, each of dimension $F \times 1$, modulated over time through activation vectors $H_j(t)$, of dimension $1 \times T$:

$$\sigma_j^2 = W_j H_j, \tag{7}$$

yielding $\sigma^2 = \sum_j W_j H_j$ for the PSD of the mixture. This can be expressed in a concise matrix form as:

$$\sigma^2 = WH, \tag{8}$$

where $W_j$ are the columns of $W$ and $H_j$ are the rows of $H$. The NMF has been very popular in audio and has encountered successful applications in both music information retrieval [9] and audio processing tasks [17], [31]. In essence, those approaches boil down to fitting the empirical power-spectrogram $p^2$ of the mixture, by minimizing the Itakura-Saito divergence $d_0$ (IS):

$$\left\{\hat{W}, \hat{H}\right\} \leftarrow \underset{W,H}{\operatorname{argmin}} \sum_{f,t} d_0\left(p^2(f,t) \mid \sum_j W_j(f) H_j(t)\right).$$

However, many studies have also reached excellent performance by fitting the magnitude spectrogram $p$ of the mixture instead of $p^2$. This is notably the case in PLCA studies [31], where

the KL divergence $d_1$ is often used as a cost-function. In audio processing, the choice of this divergence is *in fine* justified by good empirical performance, but as we discussed above, only the Cauchy model over waveforms leads to additive magnitude spectrograms. Logically, we thus propose to study the performance of the Cauchy-NMF model:

$$\sigma = WH, \tag{9}$$

where the scale parameters $\sigma(f,t)$ now pertain to Cauchy instead of Gaussian random variables $x(f,t)$.

## III. ESTIMATION OF THE PARAMETERS

On probabilistic grounds, a natural idea to estimate the parameters $\{W,H\}$ of the Cauchy NMF model is to adopt a maximum likelihood approach. Since all TF bins $x(f,t)$ are independent, this leads to[2]:

$$\left\{\hat{W},\hat{H}\right\} \leftarrow \underset{W,H}{\operatorname{argmin}} \left\{ D(\sigma) \triangleq \sum_{f,t} -\log \mathbb{P}\left(x(f,t) \mid W,H\right) \right\}, \tag{10}$$

where $D(\sigma)$ is the global cost function to be minimized and the closed-form expression of $\mathbb{P}\left(x(f,t) \mid W,H\right)$ is given by the Complex isotropic Cauchy distribution [29, ex. 2.5.6 p. 81]:

$$\mathbb{P}\left(x(f,t) \mid W,H\right) = \frac{\sigma(f,t)}{2\pi\left(p(f,t)^2+\sigma(f,t)^2\right)^{3/2}}. \tag{11}$$

It is straightforward to show that $D(\sigma)$ in (10) is given by:

$$D(\sigma) \stackrel{c}{=} \sum_{f,t} \left[ \frac{3}{2}\log\left(p(f,t)^2+\sigma(f,t)^2\right) - \log\sigma(f,t) \right], \tag{12}$$

where $\stackrel{c}{=}$ denotes equality up to an additive constant independent of $\{W,H\}$. The common methodology here is to proceed by iteratively updating $W$ and $H$ so as to decrease $D(\sigma)$, while the other one is kept fixed. We considered two approaches for this purpose. We call the first one the naive update, presented in section III-A, and the second one is called the Majorization-Equalization (ME) update, presented in section III-B.

### III-A. Naive multiplicative updates

A first straightforward but heuristic approach involves the derivative of the global cost function $D(\sigma)$ in (12) with respect to any parameter —written $\theta$— to be updated ($\theta$ is either $W$ or $H$):

$$\frac{\partial D(\sigma)}{\partial\theta} = \sum_{f,t}\left(\frac{3\sigma(f,t)}{p(f,t)^2+\sigma(f,t)^2} - \frac{1}{\sigma(f,t)}\right)\frac{\partial\sigma(f,t)}{\partial\theta}. \tag{13}$$

Since this derivative can be expressed as the difference $G_+(\theta) - G_-(\theta)$ between two nonnegative terms:

$$\begin{aligned} G_+(\theta) &= \sum_{f,t}\frac{3\sigma(f,t)}{p(f,t)^2+\sigma(f,t)^2}\frac{\partial\sigma(f,t)}{\partial\theta} \\ G_-(\theta) &= \sum_{f,t}\sigma(f,t)^{-1}\frac{\partial\sigma(f,t)}{\partial\theta}, \end{aligned}$$

we can adopt the now classical multiplicative update procedure pioneered in [2] and update $\theta$ through:

$$\theta \leftarrow \theta\cdot\frac{G_-(\theta)}{G_+(\theta)},$$

[2]The notation $\mathbb{P}(z)$ here denotes the probability density function of the random variable $z$.

All updates use the latest available versions of all parameters for computing $\sigma$:

- $W \leftarrow W \cdot \frac{(\sigma^{\cdot-1})H^\top}{zH^\top}$
- $H \leftarrow H \cdot \frac{W^\top(\sigma^{\cdot-1})}{W^\top z}$

The notations $a\cdot b$ and $\frac{a}{b}$ denote element-wise multiplications and divisions, respectively, while $a^{\cdot p}$ denotes element-wise exponentiation for $p\in\mathbb{Z}$. $z$ is defined in (14).

**Table I**. Naive multiplicative updates for Cauchy NMF.

where $a\cdot b$ and $\frac{a}{b}$ denote element-wise multiplications and divisions, respectively. Provided $W$ and $H$ have been initialized as nonnegative, they remain so throughout iterations. The procedure is summarized in table I, where $z$ is defined as:

$$z(f,t) \triangleq \frac{3\sigma(f,t)}{p(f,t)^2+\sigma(f,t)^2}. \tag{14}$$

### III-B. Majorization-equalization update

Even if the updates found in table I are derived straightforwardly using classical non-negative methodology, they do not guarantee a non-increasing cost-function $D(\theta)$. An alternative way to derive update rules for the parameters is to adopt the Majorization-Equalization (ME) approach presented in [5]. In essence, the strategy first requires identifying a majorization of the cost-function (12), which is of the form:

$$\forall(\hat{\sigma},\sigma),\ D(\hat{\sigma}) \le g(\hat{\sigma},\sigma)$$

with $\forall\sigma, D(\sigma) = g(\sigma,\sigma)$. Then, given some current parameter, we look for a new different value, such that the new model $\hat{\sigma}$ obeys $g(\hat{\sigma},\sigma) = D(\sigma)$. This approach guarantees that the cost function will be non-increasing over the iterations, and it is known to provide a faster convergence rate than the "majorize-minimize" approach [34]. Besides, remember that in the case of $\beta$-divergences, this strategy leads to the regular NMF multiplicative update rules [5]. Due to space constraints, the complete details of the ME derivation we propose will be given in a further study. Here, we simply mention that the majorization we used is:

$$\begin{aligned} \forall(\sigma,\hat{\sigma}),\ D(\widehat{\sigma}) \le D(\sigma) \\ + \sum_{f,t}\frac{3}{2}\frac{\widehat{\sigma}(f,t)^2-\sigma(f,t)^2}{\sigma(f,t)^2+p(f,t)^2} + \frac{\sigma(f,t)}{\widehat{\sigma}(f,t)} - 1. \end{aligned} \tag{15}$$

We provide the corresponding updates for the model parameters in table II and highlight that the Cauchy cost function (12) is guaranteed to be non-increasing over iterations using these updates.

## IV. EVALUATION

### IV-A. Separation performance on musical signals

To test the performance of Cauchy NMF for sound source separation, a database of 25 single channel mixtures each containing 2 monophonic instruments was used. Details of the dataset can be found in [35]. The naive Cauchy updates as well as the ME updates were tested against KL-NMF (with magnitudes of STFT) and IS-NMF (with power spectrograms). The rank of the factorization was chosen as 10, and basis functions were clustered using the oracle clustering approach described in [36]. The tests were ran 10 times, with the same random initialization used for all algorithms tested. Figure 1 then shows the Signal to Distortion Ratio (SDR) as obtained using [37]. It can be seen that the performance of both Cauchy algorithms is competitive with that of the KL divergence

All updates use the latest available versions of all parameters for computing $\sigma$. Update each matrix $\theta$ (either $W$ or $H$) through:

$$\theta \leftarrow \theta \cdot \frac{b_\theta}{a_\theta + \sqrt{a_\theta^{\cdot 2} + 2b_\theta \cdot a_\theta}}$$

where $a_\theta$ and $b_\theta$ are of the same size as $\theta$ and given by:

| $\theta$ | $a_\theta$ | $b_\theta$ |
|---|---|---|
| $W$ | $\frac{3}{4}\frac{\sigma}{\sigma^{\cdot 2}+p^{\cdot 2}}H^\top$ | $\sigma^{\cdot -1}H^\top$ |
| $H$ | $\frac{3}{4}W^\top\frac{\sigma}{\sigma^{\cdot 2}+p^{\cdot 2}}$ | $W^\top\sigma^{\cdot -1}$ |

The notations $a \cdot b$ and $\frac{a}{b}$ denote element-wise multiplications and divisions, respectively, while $a^{\cdot p}$ denotes element-wise exponentiation for $p \in \mathbb{Z}$.

**Table II**. Majorization-equalization updates for Cauchy NMF.

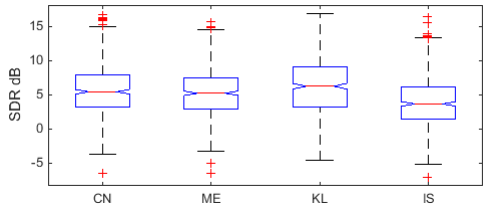


**Fig. 1**. SDR for Cauchy Naive (CN), Cauchy ME (ME), KL and IS NMF algorithms (10 runs). Higher is better

and outperforms that of IS-NMF, with informal listening tests suggesting the separation quality is perceptually similar for both Cauchy algorithms and the KL divergence. This demonstrates that Cauchy NMF is useful for audio source separation. Figure 2 shows the convergence of both Cauchy NMF algorithms against the iteration number. Both algorithms converge well, with the naive updates initially converging faster than the ME updates. In practice, the naive updates were always observed to converge even if it is not theoretically guaranteed by the update rules.

### IV-B. Denoising performance on synthetic signals

To test the denoising ability of Cauchy NMF, synthetic test data were created using 5 component pairs for $W$ and $H$ generated by taking the 4th power of random Gaussian noise, resulting in sparse components. The product $WH$, of dimension $F \times T$ was then used as the scale parameters of independent symmetric $\alpha$-stable random observations, for various values of $\alpha$ in the range 0.2 - 2. Fitting parameters on this observation permits to test the robustness of the proposed algorithms against adversity of noise, because small $\alpha$ in essence lead to observations corrupted by very adverse impulsive noise. Performance of model parameters estimation was tested for the Cauchy NMF algorithms, as well as KL-NMF, IS-NMF and RPCA. Here the rank of the NMF decompositions was set to 5.

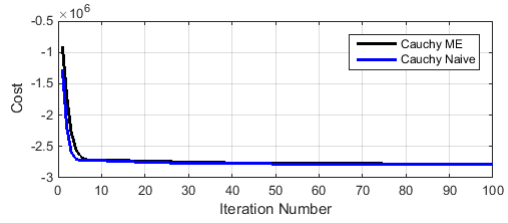


**Fig. 2**. Cauchy Cost Function vs. Iteration Number (10 runs).

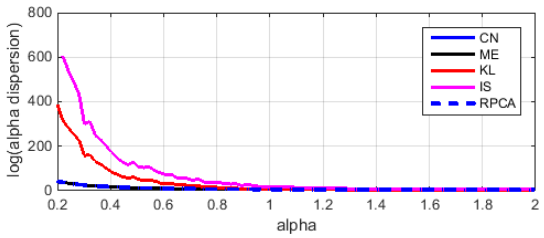


**Fig. 3**. Reconstruction of original data measured using Log($\alpha$-dispersion) as a function of $\alpha$

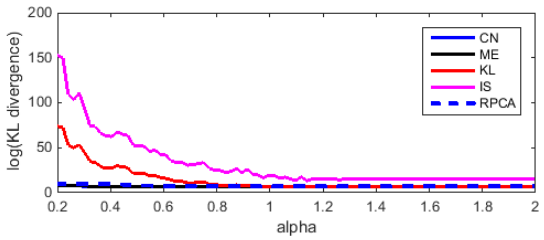


**Fig. 4**. Reconstruction of original data measured using Log(KL divergence) as a function of $\alpha$

Figure 3 shows the average results obtained over 100 independent runs when using the $\alpha$-dispersion to measure reconstruction of the original clean data as a function of $\alpha$. The $\alpha$-dispersion is defined as

$$L_\alpha = \sum_{ft} |\sigma(f,t) - \hat{\sigma}(f,t)|^{1/\alpha}, \tag{16}$$

where $\sigma = WH$ and $\hat{\sigma} = \hat{W}\hat{H}$. Due to the large range of the data, the log of the $\alpha$-dispersion is plotted.

Remarkably, the Cauchy NMF algorithms show a very similar reconstruction to that obtained using RPCA, with all three coinciding in the plots shown. It can be seen that the Cauchy NMF algorithms are more robust to noise than KL-NMF and IS-NMF, demonstrating the usefulness of Cauchy NMF in adverse noise conditions.

Also measured was the quality of the reconstruction of the data in terms of the KL divergence, shown in figure 4. Here the absolute value of the RPCA low rank matrix is used as a proxy for the actual RPCA reconstruction due to the negative values allowable in RPCA. It can be seen that, in terms of reconstruction measured using the KL divergence, Cauchy NMF algorithms are considerably more robust than KL-NMF and IS-NMF for impulsive noise ($\alpha \leq 1$), while showing improved robustness at low $\alpha$ compared to RPCA. This demonstrates that Cauchy NMF is a suitable algorithm for robust denoising of data.

## V. CONCLUSION

We have introduced new algorithms for NMF based on the complex Cauchy distribution, and shown that it is a natural fit for audio signals where the magnitude spectrograms are assumed to be additive. We provide two methods for implementing Cauchy NMF, the first one based on naive multiplicative updates and the second one based on a majorization-equalization approach, for which the cost function is guaranteed to reduce at each iteration. It is then shown that in practice both algorithms converge well and that they are competitive with existing NMF-based separation algorithms, while having the benefit of being theoretically justified. Furthermore, Cauchy NMF is demonstrated to be more robust to noise than KL and IS-NMF, while demonstrating similar robustness to RPCA.

## VI. REFERENCES


[1] G.H. Golub and C.F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[2] D. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[3] P. Smaragdis and J. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180. IEEE, 2003.

[4] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Non-negative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley Publishing, September 2009.

[5] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011.

[6] Deguang Kong, Chris Ding, and Heng Huang. Robust non-negative matrix factorization using l21-norm. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 673–682. ACM, 2011.

[7] N. Dobigeon and C. Févotte. Robust nonnegative matrix factorization for nonlinear unmixing of hyperspectral images. In *Proc. IEEE GRSS Workshop Hyperspectral Image Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2013.

[8] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *Audio, Speech, and Language Processing, IEEE Trans. on*, 15(3):1066–1074, 2007.

[9] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *Audio, Speech, and Language Processing, IEEE Trans. on*, 18(3):538–549, 2010.

[10] C. Févotte. Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 1980–1983. IEEE, 2011.

[11] A. Lefevre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 21–24. IEEE, 2011.

[12] C. Joder, F. Weninger, D. Virette, and B. Schuller. A comparative study on sparsity penalties for NMF-based speech separation: Beyond Lp-norms. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conf. on*, pages 858–862. IEEE, 2013.

[13] T. Virtanen, A.T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1825–1828. IEEE, 2008.

[14] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[15] 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)*, volume 6684, pages 102–115, Malaga, Spain, 2010., 2010. Springer.

[16] M. N. Schmidt and H. Laurberg. Non-negative matrix factorization with Gaussian process priors. *Computational Intelligence and Neuroscience*, ID 361705, 2008.

[17] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *Audio, Speech, and Language Processing, IEEE Trans. on*, PP(99):1, 2011.

[18] O. Dikmen and A.T. Cemgil. Unsupervised single-channel source separation using Bayesian NMF. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 93–96. IEEE, 2009.

[19] U. Simsekli and A.T. Cemgil. Markov chain Monte Carlo inference for probabilistic latent tensor factorization. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

[20] P. Smaragdis, C. Févotte, G.J. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, May 2014.

[21] G. Arce. *Nonlinear signal processing: a statistical approach*. John Wiley & Sons, 2005.

[22] G. Samoradnitsky and M. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.

[23] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[24] A. Liutkus and R. Badeau. Generalized Wiener filtering with fractional power spectrograms. In *40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015. IEEE.

[25] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[26] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155 –3167, July 2011.

[27] P. Kidmose. *Blind separation of heavy tail signals*. PhD thesis, Technical University of Denmark, Denmark, 2001.

[28] P. Georgiou, P. Tsakalides, and C. Kyriakakis. Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise. *IEEE Transactions on Multimedia*, 1(3):291–301, September 1999.

[29] G. Samoradnitsky and M. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.

[30] G.A. Tsihrintzis, P. Tsakalides, and C.L. Nikias. Spectral methods for stationary harmonizable alpha-stable processes. In *European signal processing conference (EUSIPCO)*, pages 1833–1836, Rhodes, Greece, September 1998.

[31] P. Smaragdis, C. Févotte, G.J. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, May 2014.

[32] A.T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill. Prior structures for Time-Frequency energy distributions. In *Proc. of the 2007 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA)*, pages 151–154, NY, USA, 2007.

[33] N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *Audio, Speech, and Language Processing, IEEE Trans. on*, 18(7):1830 –1840, sept. 2010.

[34] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011.

[35] D. FitzGerald, M. Cranitch, and E. Coyle. Extended non-negative tensor factorisation models for musical sound source separation. *Comp. Intelligence and Neuroscience*, 2008.

[36] T. Barker and T. Virtanen. Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation. In *Proceedings of Interspeech*, 2013.

[37] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(4):1462 –1469, July 2006.