Blind Source Separation Using Mixtures of Alpha-Stable Distributions
Nicolas Keriven, Antoine Deleforge, Antoine Liutkus

To cite this version:

HAL Id: hal-01633215
https://inria.hal.science/hal-01633215v3
Submitted on 9 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT

We propose a new blind source separation algorithm based on mixtures of α-stable distributions. Complex symmetric α-stable distributions have been recently showed to better model audio signals in the time-frequency domain than classical Gaussian distributions thanks to their larger dynamic range. However, inference with these models is notoriously hard to perform because their probability density functions do not have a closed-form expression in general. Here, we introduce a novel method for estimating mixtures of α-stable distributions based on characteristic function matching. We apply this to the blind estimation of binary masks in individual frequency bands from multichannel convolutive audio mixtures. We show that the proposed method yields better separation performance than Gaussian-based binary-masking methods.

Index Terms— Blind Source Separation, Binary Masking, Alpha-Stable, Generalized Method of Moments

1. INTRODUCTION

This paper is concerned with source separation, which is a topic in applied mathematics that aims at processing mixture signals so as to recover their constitutive components, called sources [1]. It is a field of important and widespread practical applications, notably in audio. It is traditionally exemplified by the cocktail party problem, which consists in isolating some specific discussion within the recording of a crowd [2, 3]. Apart from such speech processing scenarios, source separation also enjoyed much interest in the music processing literature, due to its important applications in the entertainment industry [4].

From the perspective of this paper, it is worth mentioning that a significant portion of the research on source separation first makes some assumptions on the source signals and then picks some mixing model. While the former usually stands on probabilistic grounds, the latter often comes from physical assumptions and explains how the observed mixtures are generated from the sources.

Historically, the overdetermined linear case was considered, i.e., more mixtures than sources are available [1]. The interesting fact about such mixing models is they can be inverted easily, allowing to recover the sources from the mixtures, provided their parameters are known. The breakthrough brought in by source separation is to allow identification of such mixing parameters with only very general assumptions about the sources. These assumptions are mostly either that sources are independent, identically distributed (iid.) and non-Gaussian, as in Independent Components Analysis (ICA, [5]), or that they are Gaussian but not iid. as in Second-Order Blind Identification (SOBI [6]). Going in the frequency domain allowed to extend such approaches to convolutive mixtures, i.e. for which the sensors capture the sources after some acoustic propagation whose duration is not negligible.

The validity of the mixing model and its invertibility is crucial for applying separation methods that make only broad assumptions on the sources. When such assumptions are violated, those approaches are not applicable. This typically happens in the underdetermined scenario, where fewer mixtures than sources are available, which is common in audio. In that case, separation may only be achieved through more involved source models and time-varying filtering procedures [4]. For this reason, it is natural that research in underdetermined separation focused on highly parameterized and tractable source models. In short, a huge part of the models proposed in the literature stands on Gaussian grounds, where one wants to estimate time-varying power-spectral densities and steering vectors for building the corresponding multichannel Wiener filters. In that framework, estimation is typically achieved through maximization of likelihoods, for instance using the Expectation-Maximization (EM) algorithm [9]. This line of thought leaves room for much flexibility and a large community strived to provide effective audio spectrogram models, from sophisticated linear factorization [10] to recent developments in deep learning [11].

An intrinsic weakness of Gaussian processes for modelling audio sources is to require many parameters to faithfully represent sophisticated signals. This is made unavoidable by their light-tails, which only allow for small explorations around averages and standard deviations. One typically has to pick a different Gaussian distribution for each time-frequency bin to obtain a good model [6], and precise estimation of all parameters is required for good performance. This inevitably makes all related estimation methods very sensitive to initialization. Using distributions with heavier-tails than the Gaussian for underdetermined separation has been less explored [12] although it is common practice in the overdetermined case [13]. Among such distributions, the α-stable distribution [15] enjoyed some interest in signal processing [16] and more particularly in source separation recently, because it was shown to straightforwardly yield effective filters with better perceived audio quality than the more classical Wiener [17, 18].

However, the delicate question of how to estimate the parameters of α-stable source models remains quite an open issue. It appears to be very challenging because such distributions do not provide an analytical expression for their likelihood, which prevents the use of classical inference methods. Two alternative options were considered so far. First, Markov Chain Monte Carlo methods [19] are applicable and effective at yet a high computational cost. Second, classical moment-matching methods were proposed [20] that are effective, but somewhat ad-hoc and hard to translate into the multichannel case of several mixtures.
In this paper, we use a variant of the recent algorithm introduced in [21] for the estimation of mixture models by generalized moment matching (GeMM), to exploit mixtures of multivariate α-stable distributions in the context of audio source separation. This algorithm, referred to as Compressive Learning-Orthogonal Matching Pursuit with Replacement (CL-OMPR), is a greedy, heuristic method that was initially used in the context of sketching [21] to estimate mixture models on large-scale databases using only a collection of generalized moments computed in one pass. Sketching enjoyed several successful applications in machine learning [22], but also in source localization [23]. Here, we exploit instead the capacity of CL-OMPR to estimate an α-stable mixture model whose probability density function does not enjoy an analytical expression.

2. ALPHA-STABLE UNMIXING

2.1. Alpha-stable mixture model

Let us consider a mixture of $K$ sound sources observed through $M$ channels. We denote by $\{s_k(f,t)\}_{k=1}^K$ the emitted source spectrograms and by $\{x_{o,k}(f,t)\}_{k=1}^M$ the observed channel spectrograms in the complex short-time Fourier domain, where $f \in [1 \ldots F]$ and $t \in [1 \ldots T]$ denote the discrete frequency and time indexes. Assuming time-domain convolutive filters from sources to channels which are short compared to the Fourier windows, the mixing model at $(f,t)$ can be written

$$x(f,t) = \sum_{k=1}^K a_k(f)s_k(f,t)$$

where $x(f,t) \in \mathbb{C}^M$ is the observed vector, $s(f,t) \in \mathbb{C}^K$ the source vector and $a_k(f) \in \mathbb{C}^M$ source $k$’s steering vector.

Now, we choose an original probabilistic model for the source signals, inspired by recent research on α-harmonizable processes [17, 23]. For each $f$, all $\{s_k(f,t)\}_{k=1}^K$ are assumed independent and identically distributed (iid.) with respect to (wrt.) a symmetric complex and centered α-stable distribution of unit scale parameter and characteristic exponent $\alpha_{k,f}$, which we write:

$$p(s_k(f,t); \alpha_{k,f}) = \mathcal{S}_\alpha(s_k(f,t); \alpha_{k,f}).$$

In short, the symmetric centered α-stable distribution generalizes the Gaussian isotropic one [24], while providing significantly heavier tails as its characteristic exponent $\alpha_{k,f} \in [0, 2]$ gets small, $\alpha_{k,f} = 2$ corresponding to the Gaussian case. Contrary to classical Gaussian audio source models [78] the parameters of the proposed model are time-invariant, drastically reducing its size. This is permitted by the fact that the distribution $\mathcal{S}_\alpha$ enables important dynamics for $s_k(f,t)$. In other words, [2] corresponds to a model for the marginal distribution of the sources. Such ideas have already been considered in [23]. The particularity of our approach in this regard is to feature a frequency-dependent characteristic exponent $\alpha_{k,f}$ for increased expressive power. The choice of a unit scale for the distribution comes with no loss of generality: any frequency-dependent scaling of the sources is incorporated in the steering vectors $a_k(f)$.

We highlight that the probability density function (pdf) of $s_k(f,t)$ in [2] does not have a closed-form expression except for $\alpha_{k,f} = 1$ (Cauchy) and $\alpha_{k,f} = 2$ (Gaussian). However, its characteristic function, defined as the Fourier transform of its pdf does. We have [15, 17]:

$$\forall \omega \in \mathbb{C}, \mathbb{E}\{\exp(i\Re[\omega^*s_k(f,t)])\} = \exp(-|\omega|^\alpha_{k,f}).$$

At this point, we make one important simplifying assumption: we suppose only one source is significantly active at each time-frequency (TF) point. More specifically, let $z(f,t)$. be the index of the source that has the strongest magnitude $|s_k(f,t)|$ at TF bin $(f,t)$. Our assumption is that all other sources have a magnitude close to $0$. This is less strong than the so called W-disjoint orthogonality assumption [3] where a single source is assumed to be active. This allows us to assume weak sources are approximately distributed wrt a Gaussian distribution. Indeed, even if it lacks an analytical expression, the pdf for a symmetric α-stable distribution is infinitely differentiable close to the origin [15], justifying this second order approximation for weak sources.

As a result of these assumptions, we take our mixture as:

$$x(f,t) = \sum_{k=1}^K z(f,t) = k\{a_k(f)s_k(f,t) + e_k(f,t)\},$$

where $I$ is the indicator function and $e_k(f,t) \in \mathbb{C}^M$ is a residual Gaussian term containing all non-dominating signals (other than $k$) and possible additional noise. For convenience, we neglect the inter-channel correlations coming from weak sources, to simply assume that $e_k$ is composed of iid. entries with variance $\sigma^2_{e_k}$:

$$p(e_k(f,t); z(f,t) = k; \sigma^2_{e_k}) = \mathcal{N}_{c}(e_k(f,t); 0, \sigma^2_{e_k}I_M) \tag{5}$$

where $\mathcal{N}_c$ denotes the multivariate complex circular-symmetric Gaussian distribution [24]. $I_M$ is the $M$-dimensional identity matrix and $\sigma^2_{e_k}$ is the residual variance at frequency $f$ when source $k$ dominates. Furthermore, the indexes $z(f,t)$ of the strongest source for each TF bin are modelled as iid. multinomial variables:

$$p(z(f,t) = k; \pi_{f}) = \pi_{k,f} \tag{6}$$

where $\pi_{k,f}$ is the probability of source $k$ dominating in frequency band $f$, and $\sum_{k} \pi_{k,f} = 1$.

From all the preceding assumptions and dropping the indexes $(f,t)$ for convenience, we deduce the characteristic functions of $a_k, s_k, e_k$ and $x|z = k$, where $\omega \in \mathbb{C}^M$:

$$\psi_{a_k}(\omega) = \exp(-|\omega|^\alpha_{k}) \tag{7}$$

$$\psi_{e_k}(\omega) = \exp(-\sigma^2_{e_k}|\omega|^2) \tag{8}$$

$$\psi_{x|z = k}(\omega) = \exp(-|\omega|^\alpha_{k} - \sigma^2_{e_k}|\omega|^2) \tag{9}$$

Combining [7] and [8], we deduce that $\{x(f,t)\}$, follows a mixture model parametrized by

$$\theta_f = \{\alpha_{k,f}, \sigma^2_{e_k}, a_k(f), \pi_{k,f}\}_{k=1}^K \tag{10}$$

Following the two-stage approach of [25], the proposed blind source separation method consists in clustering observations $x(f,t)$ independently at each frequency according to this mixture model. The resulting classical source permutation ambiguity across frequencies is left aside here (see Section 2.4), and a binary mask is then obtained for each source [23]. The special Gaussian case $\alpha_{k,f} = 2$ is discussed in Section 2.2 while a parameter estimation method for the general case is given in Section 2.3.

2.2. Special case $\alpha_{k,f} = 2$

Let us consider the special Gaussian case where $\alpha_{k,f} = 2$ for all $f,k$. The observation model at each frequency becomes

$$p(x|z = k; \theta) = \mathcal{N}_{c}(x|f,t; 0, a_k a^*_k + \sigma^2_{e_k}I_M) \tag{11}$$
where frequency indexes have been dropped for convenience. The parameters $\theta$ of this mixture model can be straightforwardly estimated via an expectation-maximization (EM) procedure \cite{Dempster1977}. Interestingly, using the re-parameterization $a_k \leftarrow \sigma_k a_k$ and $\sigma_k \leftarrow 2 \sigma_k^2$, it turns out that these EM updates match those of the blind source separation model proposed in \cite{Hyvarinen2000}, up to a small additive constant for $\sigma_k^2$. A key difference is that in \cite{Hyvarinen2000}, the observations are normalized so that $\|x_t\|_2 = 1$. As such, \cite{Hyvarinen2000} belongs to the class of spatial-feature clustering-based methods, similarly to DUET \cite{Batista2015}, while our method operates in the signal domain.

2.3. Parameter estimation via generalized moment matching

In the general case $\alpha \neq 2$, estimation is done by generalized moment matching, that is, minimizing the difference between the empirical and theoretical values of a finite number of generalized moments, which are here samples of the empirical characteristic function of the data at some frequency vectors $\omega_j \in \mathbb{C}^M$, $j \in [1, \ldots, J]$, to be matched with their analytical expression \cite{Nowak1994}. Following the methodology in \cite{Oja1983}, the vectors $\omega_j$ are drawn randomly according to some probability distribution $\Lambda$, in practice designed automatically from the data using the method prescribed in \cite{Oja1983}.

More precisely: given the data points to cluster $x_1, \ldots, x_T \in \mathbb{C}^M$ (where the index $f$ has been dropped), the estimation is performed as follows:

1. Draw $m$ random vectors $\omega_j \overset{iid.}{\sim} \Lambda$ for $j \in [1, \ldots, J]$;
2. Compute the empirical characteristic function at these frequencies $y = \frac{1}{T} \sum_{t=1}^T e^{i \Re(\omega_j^T x_t)} \in \mathbb{C}^J$;
3. Estimate the model parameters \cite{Nowak1994} by (approximately) solving
   \[
   \min_{\theta} \left\| y - \left[ \psi(x|z=k, \omega_j) \right]_{j=1}^J \right\|_2^2 \quad (12)
   \]
   where $\psi(x|z=k, \omega_j)$ is defined by \cite{Nowak1994}, parameterized by $\theta$.

**CL-OMP**. The minimization \cite{Oja1983} is carried out by a modified version of the CL-OMP algorithm \cite{Oja1983} adapted to our model. It is a greedy, heuristic algorithm precisely designed to perform mixture model estimation by generalized moment matching. Although it offers limited theoretical guarantees except for very particular settings \cite{Nowak1994}, it has been empirically shown to perform well for a large variety of models \cite{Oja1983}. In particular, it is applicable as soon as the considered mixture model has a closed-form characteristic function with respect to the parameters of the model, which is the case for mixture of $\alpha$-stables distributions. Although it was initially designed to perform mixture model estimation on large databases, we use it here mostly because the probability density function of the proposed model \cite{Oja1983} does not enjoy an analytical expression. This forbids the use of classical methods such as EM. To our knowledge, there is no other algorithm capable of estimating mixtures of multivariate $\alpha$-stable distributions in the literature.

The CL-OMP algorithm is a variant of Orthogonal Matching Pursuit (OMP), a classical greedy algorithm in compressive sensing. Like OMP, it iteratively adds a component to the mixture model by maximizing its correlation to the residual signal. Since the space of parameters is continuous, this is done here with a gradient ascent randomly initialized. Furthermore, CL-OMP alternates this greedy step with a non-convex, global gradient descent on \cite{Oja1983} initialized with the current support. This additional step adjust the whole support when a component is added. Finally, it also performs more iterations than OMP and includes a hard thresholding step to maintain the number of components at $K$, to allow for replacing spurious components.

The CL-OMP algorithm is described in detail in \cite{Oja1983}, where it is applied to Gaussian Mixture Model (GMM) estimation. Replacing the GMM by our $\alpha$-stable model is easily implemented and only requires computation of the gradient of $\psi(x|z=k)$ with respect to the different parameters. The code is available at https://github.com/nkeriven/alpha_stable_bss

**Approximate clustering**. A drawback of the $\alpha$-stable model to investigate in future work is that the pdf $p(x|z=k)$ does not have an explicit expression. Therefore, the clustering of data points $x_t$ cannot be done by exactly maximizing the posterior $p(x_t|z=k)$ with respect to $k$. In other words, once we have estimated the mixture of $\alpha$-stable distributions, it is difficult to actually assign each point to a component of the mixture.

Although a few methods may exist to approximately compute this posterior using approximate numerical integration \cite{Nowak1994}, in practice we found them to be extremely unstable and time consuming. Instead, we decided to cluster the data as if the model was Gaussian, i.e. with $\alpha_k = 2$, since the likelihood is then computable. Hence, the “clustering” part (and therefore the final source separation step) of both EM (Section 2.2) and the $\alpha$-stable model are in fact the same. The difference between the two lies in the estimation of parameters $(\alpha_k, \sigma_k^2, \pi_k)$. Our hope is that by using the more realistic $\alpha$-stable source model, steering vectors $\alpha_k$ will be estimated more precisely.

2.4. Frequency permutation ambiguity

Once clustering is performed at each frequency, a permutation ambiguity remains as the assignment of frequency masks to sources is not known. This is a classical problem in blind source separation referred to as permutation alignment. It notably occurs when using ICA \cite{Hyvarinen2000} and clustering-based methods \cite{Batista2015} \cite{Oja1983}. A number of techniques have been proposed to tackle it, based on temporal activation patterns \cite{Oja1983}, steering vector models \cite{Batista2015} or adjacent frequency bands similarity \cite{Oja1983}. The selection and tuning of a specific permutation technique highly depends on the type of signal and mixing model considered, which is out of the scope of this study. For this reason and for fairness, all methods evaluated in the next section benefited from the same oracle permutation scheme. At each frequency, the permutation minimizing the mean-squared error between estimated and true source images is selected.

3. EVALUATION AND RESULTS

We use two datasets for evaluation. First, a subset of the QUASI database\footnote{www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/} consisting in 10 musical excerpts of 30s. For each excerpt, we produced stereo ($M = 2$) mixes of $K = 4$ musical tracks (vocals, bass, drums, electric guitar, keyboard...) using random pure gains and delays. Second, the TIMIT speech database\footnote{catalog.ldc.upenn.edu/ldc93s1} from which we created 10 tracks of 30s. For each experiment we mix $K = 3$ of them selected at random into $M = 2$ channels, again with random pure gains and delays. In all cases, the gain difference between the two channels are at most 5dB and the delay is at most 20 samples. Note that none of the tested methods make assumption on the specific convolutive filters used for mixing, as long as they are relatively small compared to the Fourier analysis window. The STFT
Table 1: Separation results with $K$ sources and $M = 2$ channels, for the four clustering algorithms as well as oracle and mixture results. Each slot contains the mean and standard deviation over the 100 trials and $K$ sources, i.e. over 100$K$ values.

<table>
<thead>
<tr>
<th></th>
<th>SDR (dB)</th>
<th>SIR (dB)</th>
<th>MER (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mix</td>
<td>-5.96 ± 4.96</td>
<td>-5.49 ± 4.85</td>
<td>N/A</td>
</tr>
<tr>
<td>Oracle</td>
<td>8.33 ± 3.16</td>
<td>18.3 ± 4.13</td>
<td>N/A</td>
</tr>
<tr>
<td>[25]</td>
<td>1.26 ± 2.44</td>
<td>2.88 ± 3.82</td>
<td>10.5 ± 9.84</td>
</tr>
<tr>
<td>EM</td>
<td>3.50 ± 2.87</td>
<td>9.04 ± 4.92</td>
<td>12.3 ± 11.0</td>
</tr>
<tr>
<td>CF-GMM</td>
<td>3.80 ± 2.53</td>
<td>8.60 ± 3.62</td>
<td>12.3 ± 9.90</td>
</tr>
<tr>
<td>CF-$\alpha$</td>
<td>4.11 ± 2.59</td>
<td>9.17 ± 3.51</td>
<td>12.7 ± 9.73</td>
</tr>
</tbody>
</table>

(a) QUASI database (music), $K = 4$

<table>
<thead>
<tr>
<th></th>
<th>SDR (dB)</th>
<th>SIR (dB)</th>
<th>MER (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mix</td>
<td>-3.14 ± 1.91</td>
<td>-3.13 ± 1.90</td>
<td>N/A</td>
</tr>
<tr>
<td>Oracle</td>
<td>11.9 ± 0.98</td>
<td>25.9 ± 1.05</td>
<td>N/A</td>
</tr>
<tr>
<td>[25]</td>
<td>2.16 ± 1.33</td>
<td>4.90 ± 2.54</td>
<td>22.0 ± 6.57</td>
</tr>
<tr>
<td>EM</td>
<td>0.54 ± 0.50</td>
<td>1.44 ± 1.21</td>
<td>12.0 ± 3.64</td>
</tr>
<tr>
<td>CF-GMM</td>
<td>1.60 ± 1.10</td>
<td>4.13 ± 2.46</td>
<td>14.8 ± 3.32</td>
</tr>
<tr>
<td>CF-$\alpha$</td>
<td>2.70 ± 1.74</td>
<td>6.11 ± 3.31</td>
<td>18.9 ± 2.72</td>
</tr>
</tbody>
</table>

(b) TIMIT database (speech), $K = 3$

Separation results. In Table 1 we show the separation results for all algorithms. Recall that [25] performs separation purely based on spatial clustering while EM, CF-GMM and CF-$\alpha$ also rely on statistical source models. The results suggest that first, using source models is more beneficial in heavily underdetermined scenarios, e.g. Table 1(a), where source signals are less sparse and more numerous. Second, the proposed $\alpha$-stable model is better suited than Gaussian models for both speech and music sources. Finally, the proposed approach blindly estimates mixing filters in a more stable way than the EM approach of [25] despite its multiple initialization, as showed by the lower standard deviations of MER.

Relevance of log-likelihood. A somewhat surprising observation is that CF-GMM significantly outperforms EM on speech data, despite the fact that both estimate a GMM. In Fig. 1 we compare the log-likelihood results obtained with the three algorithms during the clustering phase subsequent to the estimation of the parameters (recall that all three algorithms have the same clustering phase that do not use the estimated $\alpha_k$). As expected, EM significantly outperforms the two other algorithms on this criterion. This is not surprising since EM aims at maximizing the log-likelihood while the two CF algorithms consider only the characteristic function. Since the CF approaches outperform EM in terms of separation results, we conclude that maximization of the log-likelihood, while natural, might not be the most appropriate approach to estimate the mixture parameters in this case, which is an interesting lead for future research.

4. CONCLUSION

We presented a novel method for multichannel blind separation of audio sources using an $\alpha$-stable model for source signals, combined with the assumption that only one source dominates each ($t, f$) point. The parameters of the proposed model, including distinct scale and $\alpha$ values for each source, are estimated at each frequency using a novel method based on random generalized moment matching. Results show that using oracle permutations, the proposed model performs better than Gaussian models, and that the proposed estimation method outperforms EM even using the same Gaussian model. Future work will further investigate the $\alpha$ and scale values estimated by our method. In particular, it would be interesting to see if they can be constrained or exploited to resolve permutation ambiguities. The potential of random generalized moment matching versus maximum likelihood methods in source separation should also be further studied.
5. REFERENCES


