Median Based Clustering for Underdetermined Blind Signal Processing

Aditya Puranik, Vaibhav Sachdeva

June 15, 2020

Abstract

In underdetermined blind source separation, more sources are to be extracted from less observed mixtures without knowing both sources and mixing matrix. K-means-style clustering algorithms are commonly used to do this algorithmically given sufficiently sparse sources,but in any case other than deterministic sources, this lacks theoretical justification. After establishing that mean-based algorithms converge to wrong solutions in practice, we propose a median-based clustering scheme. Theoretical justification as well as algorithmic realizations (both online and batch) are given and illustrated by some examples.

1 Introduction

Blind Source Separation (BSS) is the separation of a set of source signals from a set of mixed signals, without the aid of information (or with very little information) about the source signals or the mixing process. The goal here is to identify the mixing matrix **A** by **Blind Mixing model Recovery** (BMMR) and the **n**-dimensional source random vector **s** by **Blind Source Recovery** (BSR) from an observed mixture random vector **x** where **x** is:

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

where,

x = m-dimensional mixture random vector
s = n-dimensional source random vector with component density p(s)
A = mixing matrix having independent columns each of unit norm

More sources are to be extracted from less observed mixtures in **Underdetermined Blind Source Separation** (n < m) without knowing both the sources and the mixing matrix. Blind Source Separation (BSS) is predominantly based on independent source assumptions. Assuming that statistically independent sources have at most one Gaussian variable, it is known that **A** is uniquely determined by **x**. The most commonly used algorithms are based on sparse sources that are correctly identified by a clustering of k-means. But, mean-based clustering can only classify the appropriate **A** if the data density approaches a delta

distribution. Mean-based clustering has no equivariance property which suggests that the performance would be independent of **A**. Thus, a **median-based** approach has been chosen.

Blind Source Separation follows a two step approach:

1. Geometric matrix recovery of **A** through Blind Mixing Model recovery (BMMR).
2. Source extraction through Blind source recovery (BSR).

2 Background

The most commonly used overcomplete algorithms rely on sparse sources , which can be identified by clustering, usually by k-means or some extension. However, apart from the fact that mean-based clustering lacks theoretical justifications, it only identifies the correct A if the data density approaches a delta distribution and hence, at times converges to the wrong solution. Also, mean based clustering does not possess any equivariance property which suggests that performance will be independent of **A**. However, apart from the fact that theoretical justifications have not been found, mean-based clustering only identifies the correct **A** if the data density approaches a delta distribution. In Fig. 1, we illustrate the deficiency of mean-based clustering; we get an error of up to 5 per mixing angle, which is rather substantial considering the sparse density and the simple, complete case of m=n=2. Moreover, the figure indicates that median-based clustering performs much better. Indeed, mean-based clustering does not possess any equivariance property (performance independent of **A**).

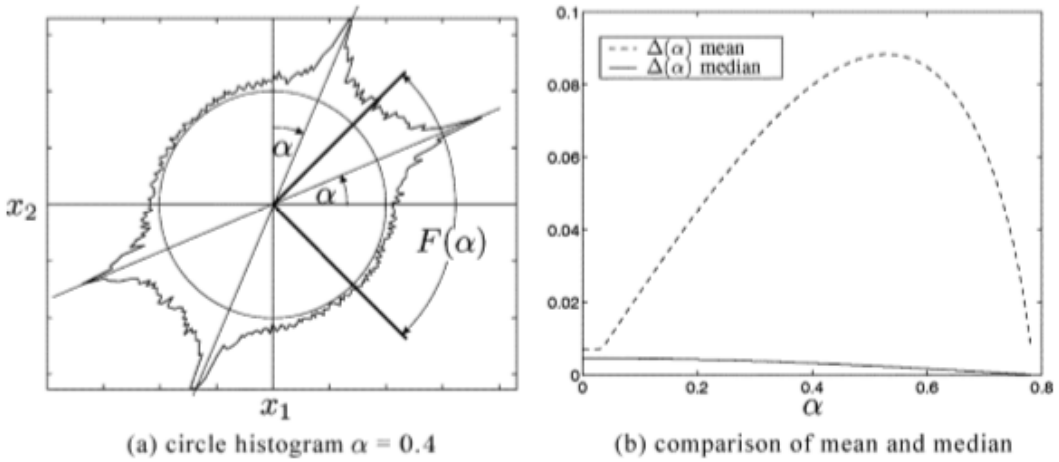

Figure 1: Mean- versus median-based clustering. We consider the mixture x of two independent gamma-distributed sources ($\gamma = 0.5, 10^5$ samples) using a mixing matrix A with columns inclined by α and ($\pi/2$ - α), respectively. (a) Mixture density for $\alpha = 0.4$ after projection onto the circle. (b) For α [0; $\pi/4$), we compare the error when estimating **A** by the mean and the median of the projected density in the receptive field F(α)=(-$\pi/4$; $\pi/4$) of the known column a of **A**. The former is the k-means convergence criterion.

3 Methods

3.1 Geometric Matrix Recovery

Without loss of generality, we assume that $\mathbf{A}$ has pairwise linearly independent columns, and $m < n$. BMMR tries to identify $\mathbf{A}$ in $\mathbf{x} = \mathbf{As}$ given $\mathbf{x}$, where $\mathbf{s}$ is assumed to be statistically independent. Obviously, this can only be done up to equivalence, where $\mathbf{B}$ is said to be equivalent to $\mathbf{A}$, $\mathbf{B} \sim \mathbf{A}$, if $\mathbf{B}$ can be written as $\mathbf{B} = \mathbf{APL}$ with an invertible diagonal matrix $\mathbf{L}$ (scaling matrix) and an invertible matrix $\mathbf{P}$ with unit vectors in each row (permutation matrix). Hence, we may assume the columns a_i of $\mathbf{A}$ to have unit norm. For geometric matrix-recovery, we use a generalization of the geometric independent component analysis (ICA) algorithm. Let $\mathbf{s}$ be an independent n-dimensional, Lebesgue-continuous, random vector with density p_s describing the sources. As $\mathbf{s}$ is independent, p_s factorizes into $p_s(s_1, ..., s_n) = p_{s1}(s_1).p_{s2}(s_2)..p_{sn}(s_n)$ with the one-dimensional marginal source density function p_{si}. We assume symmetric sources, i.e., for $s \in \Re$ and $i \in [1 : n] :=$ 1,...,n , in particular $\mathbf{E(s)} = 0$. The *geometric blind mixing model recovery* (BMMR) algorithm for symmetric distributions goes as follows :

Pick *2n* starting vectors on the unit sphere $w_1, w_1', ..., w_n, w_n'$ such that w_i) and w_i' are opposite each other, i.e.w_i=$-w_i'$ for i=1,.....,n and such that the w_i are pairwise linearly independent vectors. These w_i are called **neurons**.

All the neurons are updated using the following iteration rule using **k-means** algorithm.

$$w_i(t+1) = \pi(w_i(t) + \eta(t)\pi(y(t) - w_i(t)))$$

where,

$$y(t) = x(t)/|x(t)|$$

learning rate: $\eta : N \to \Re$ such that $\eta(t) > 0$.

$$\pi(x) := x/|x|$$

such that $w_i'(t+1) = -w_i'(t+1)$. We will mean by median in the above algorithm and prove the equivariance and convergence property of the performance.

3.2 BSR

Using the results from the BMMR step, we can assume that an estimate of $\mathbf{A}$ has been found. In order to solve the overcomplete BSS problem, we are therefore left with the task of reconstructing the sources using the mixtures $\mathbf{x}$ and the estimated matrix (BSR). Since $\mathbf{A}$ has full-rank, the equation yields the $n - m$-dimensional affine vector space as solution space for $\mathbf{s(t)}$. Hence, if $n > m$, the source-recovery problem is ill-posed without further assumptions. Using a maximum-likelihood approach, an appropriate assumption can be derived.

The maximum-likelihood algorithm states that the probability of observing $\mathbf{x}$ given $\mathbf{A}$ and $\mathbf{s}$ can be written as $P(\mathbf{x}|\mathbf{s}; \mathbf{A})$.

Using Bayes Theorem the posterior probability of $\mathbf{s}$ is -

$$P(\mathbf{s}|\mathbf{x}, \mathbf{A}) = \frac{P(\mathbf{x}|\mathbf{s}, \mathbf{A}) \times P(\mathbf{s})}{P(\mathbf{x})}$$

such that $\mathbf{x} = \mathbf{A}\mathbf{s}$.
For reconstructing $\mathbf{s}$, we will maximize the posterior probabilities

$$\mathbf{s} = argmax\ P(\mathbf{s}|\mathbf{x}, \mathbf{A}) = argmax\ P(\mathbf{x}|\mathbf{s}, \mathbf{A}) \times P(\mathbf{s})$$

4 Progress

4.1 Tasks completed

1. Two mixtures **X1 and X2** were generated using three source signals **S1, S2, and S3.**
2. Implemented the weight update rule mentioned in the paper and acquired the optimum weights, thus finding neurons associated with it.
3. Obtained the source signals back using Pseudo Inverse Method.

4.2 Tasks to be completed

1. Extracting **A** by means of Mean Based Clustering, and then by Median Based Clustering and collating them.
2. The **A** obtained from both the methods will be compared.
3. The source signals **S1, S2, and S3** will then be recovered from the obtained matrix **X** through **A** by ML rule.