World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:5, No:3, 2011

Learning an Overcomplete Dictionary using a Cauchy Mixture Model for Sparse Decay

E. S. Gower and M. O. J. Hawksford

***Abstract*—**An algorithm for learning an overcomplete dictionary using a Cauchy mixture model for sparse decomposition of an under-determined mixing system is introduced. The mixture density function is derived from a ratio sample of the observed mixture signals where 1) there are at least two but not necessarily more mixture signals observed, 2) the source signals are statistically independent and 3) the sources are sparse. The basis vectors of the dictionary are learned via the optimization of the location parameters of the Cauchy mixture components, which is shown to be more accurate and robust than the conventional data mining methods usually employed for this task. Using a well known sparse decomposition algorithm, we extract three speech signals from two mixtures based on the estimated dictionary. Further tests with additive Gaussian noise are used to demonstrate the proposed algorithm's robustness to outliers.

***Keywords*—**expectation-maximization, Pitman estimator, sparse decomposition

I. Introduction

An overcomplete dictionary (OD) is a collection of basis vectors such that their number exceeds the dimensionality of the data [1-5]. ODs can model more intricate data structures than complete dictionaries such as principal component analysis (PCA) [6] and independent component analysis (ICA) [7]. PCA is usually used to find the basis vectors in the directions of greatest data variation by modeling the data as a multivariate Gaussian density. The basis vectors, called principal components, are restrained to be orthogonal. The limitation is that if the data is non-Gaussian the model can predict the data where none occur. Unlike PCA, ICA basis vectors can be non-orthogonal.Applications of PCA and ICA are limited to cases where the basis vectors are as many as the observed mixture signals, which mean that the source signals are equal in number to the mixtures. To identify more sources than mixtures requires the use of an OD which can be used to learn overcomplete representations (finding a representation of the data in which only a few components are significant at the same time) often called sparse representations. Most algorithms developed for learning basis vectors use all the available sample values with equal weight to estimate their optimal directions with respect to some statistical assumption (such as reduced mutual information), as is the case in [1-7]. This means that they are susceptible to noise effects because a slight change of a few sample values (which might be due to additive noise) will invariably affect the concerned basis vectors, leading to an erroneous OD.

In this paper we derive an algorithm for learning an OD using a Cauchy mixture model (CMM) by variable weighting of the available sample values for a more robust and accurate learning of the basis vectors. In the CMM density function, the location parameters of the Cauchy mixture components coincide with the optimal directions of the basis vectors. We show that these location parameters are adequately defined by the ratio sample values at time instances where only one of the sparse source signals is observed. This avoids equal use of all sample values of which most might not offer valuable traction of the optimal directions of the basis vectors and reduces the deleterious effects of additive noise.

This paper is structured as follows: In Section II we derive a CMM density function from the ratio sample of a given pair of mixture signals, where the location parameters of the Cauchy mixture components coincide with the directions of the OD's basis vectors. In Section III we derive a maximization technique based on the Pitman estimator to optimize the location parameters, hence the basis vectors. In Section IV, we acknowledge the current sparse decomposition algorithms for using the estimated OD to obtain sparse representations. The Simulation results are in Section V. The discussion in Section VI summarizes the advantages of the proposed algorithm as well as its limitations.

II. Problem Modeling

For a set of M observed mixture signals $\{x_1(n),\ldots,x_M(n)\}$ and K statistically independent and sparse source signals $\{s_1(n),\ldots,s_K(n)\}$, with $K \geq M$, let the matrix $\mathbf{X}$ be given by

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S},$$

which can be expanded to

$$\begin{bmatrix} X_1 \\ \vdots \\ X_M \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1K} \\ \vdots \ddots \vdots \\ a_{M1} \cdots a_{MK} \end{bmatrix} \begin{bmatrix} S_1 \\ \vdots \\ S_K \end{bmatrix},$$

where $n \in [1, N]$ is the discrete time index, the variable X_m is realized by the data points of $x_m(n)$, for $m \in [1, M]$, S_k is the variable for $s_k(n)$, for $k \in [1, K]$, and $\mathbf{A}$ is the $M \times K$ mixing matrix or overcomplete dictionary. Using only $\mathbf{X}$, the

E. S. Gower is with the Department of Electrical Engineering, Faculty of Engineering, University of Botswana, Gaborone, Botswana, email: ephraim.gower@mopipi.ub.bw.

M. O. J. Hawksford is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom, email: mjh@essex.ac.uk.

task is to retrieve $\mathbf{S}$, but to do so we must first infer the overcomplete dictionary $\mathbf{A}$. Since the sources are assumed to be sparse, let $n \in \mathrm{T}_n(S_k)$ be the time instances when $S_{j\neq k} = 0$ and $S_k \neq 0$, for $j,k \in [1,K]$. Therefore, given a pair of variables X_m and X_r, $r \neq m$, we have

$$\begin{bmatrix} X_r \\ X_m \end{bmatrix} = \begin{bmatrix} a_{rk} \\ a_{mk} \end{bmatrix} S_k, \quad \text{for} \quad n \in \mathrm{T}_n(S_k), \tag{1}$$

from which we can form the ratio sample

$$Q_{mr} = \frac{X_m}{X_r} = \frac{a_{mk}}{a_{rk}}, \quad \text{for} \quad n \in \mathrm{T}_n(S_k). \tag{2}$$

As a result of (2), the basis vectors of the OD are adequately determined by the sample values of Q_{mr} at the time instances $n \in \mathrm{T}_n(S_k)$, $k \in [1,K]$. The sample values where $n \notin \mathrm{T}_n(S_k)$ are unnecessary for determining the OD.For sparse sources, it is most likely that the values of Q_{mr} for $n \in \mathrm{T}_n(S_k)$ are the modes of its density function. But without any knowledge of the type of this density, one cannot ascertain that the mode is the minimum variance estimator of the points of interest. However, the expected heavy tailed nature of the density function of Q_{mr} (due to division by the sample values of X_r close to zero) discourages the use Gaussian and platykurtic densities and makes the Cauchy density function attractive since it is thicker tailed than most densities.Fig.1 shows an example scatter plot for X_r and X_m for $r,m \in [1,M]$ and $r \neq m$. The black dots are the scatter plot points and it is assumed that both variables are centered. Let $\mathbf{A} = [\mathbf{a}_1,\ldots,\mathbf{a}_k]$, where $\mathbf{a}_k$ is the column vector k of the overcomplete dictionary of which we are interested in the value of (2) for it. Let there be the q_{mr}-axis orthogonal to the unknown vector $\mathbf{a}_k$ as shown. The magnitude of the vector from the origin to the q_{mr}-axis is u_k and the point of intersection of $q_{mr} = c_k$. A line is drawn from the origin to an arbitrary scatter plot point and it crosses the q_{mr}-axis at the arbitrary value q_{mr}. This line makes an angle θ_{mr} with respect to the unknown and wanted direction of $\mathbf{a}_k$ (i.e. the direction is given by (2) as a ratio of the two elements of $\mathbf{a}_k$). Under this model, the values of the q_{mr}-axis at the intersection points are the realizations of the ratio sample Q_{mr}, where the subscript k in u_k and c_k is used to illustrate the relation to the vector $\mathbf{a}_k$.

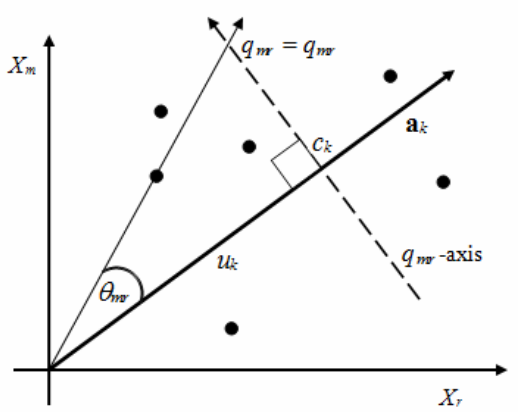

Fig. 1 A scatter plot for X_r and X_m with $\mathbf{a}_k$ as the column vector k of the mixing matrix $\mathbf{A}$. u_k is the distance from the origin to the intersection point of $\mathbf{a}_k$ with the q_{mr}-axis at the point $q_{mr} = c_k$. The angle θ_{mr} is made by the line from the origin to an arbitrary scatter plot point with respect to $\mathbf{a}_k$. This line crosses the q_{mr}-axis at the point $q_{mr} = q_{mr}$.

The angle θ_{mr} is such that $\tan(\theta_{mr}) = (q_{mr} - c_k)/u_k$, which means that

$$\frac{d\theta_{mr}}{\pi} = \frac{1}{\pi} \cdot \frac{u_k}{[u_k^2 + (q_{mr} - c_k)^2]},$$

and integrating over all possible values of Q_{mr} evaluates to one. This means that

$$\mathrm{p}(q_{mr} \mid c_k, u_k) = \frac{u_k}{\pi\left[u_k^2 + \left(q_{mr} - c_k\right)^2\right]}, \tag{3}$$

where $u_k \in (0,\infty)$ and $c_k (-\infty,\infty)$ are the scale and location parameters respectively, and $\mathrm{p}(q_{mr} \mid c_k, u_k)$ is known as the Cauchy or Lorentzian density function [8]. From (3), even though X_r and X_m may not be Gaussian (i.e. the ratio of two centered Gaussians is a Cauchy), by derivation the Cauchy density appears to be a good approximation of the underlying density function. Moreover, X_r and X_m are the sum of statistically independent variables $S_1,\ldots,S_K$ and by the central limit theorem [9] the mixture variables may be approximately Gaussian, further making the estimation of Q_{mr} by a Cauchy variable reasonable.The model in (3) is for one column vector. For K column vectors there are K independent mixture components (assuming that the original sources are statistically independent). Thus the probability of the sample value $Q_{mr} = q_{mr}$ is given by the CMM

$$\mathrm{p}_{mix}\left(q_{mr} \mid \varphi_1,\cdots,\varphi_K\right) = \sum_{k=1}^{K} \mathrm{p}\left(q_{mr} \mid \varphi_k\right) \mathrm{P}(k), \tag{4}$$

where $\mathrm{p}(q_{mr} \mid \varphi_k)$ is the mixture component k with the parameter set $\varphi_k = \{c_k, u_k\}$ and the relative weight $\mathrm{P}(k)$ such that

$$\sum_{k=1}^{K} \mathrm{P}(k) = 1.$$

To estimate the directions of the basis vectors of $\mathbf{A}$, all we need to do is to find the location parameters $\{c_1, \cdots, c_K\}$ of all the mixture components.

III. Learning Overcomplete Dictionaries

Q_{mr} is drawn from K mixture components and it is unknown which component of the CMM generated which sample values. Therefore, let the *n-th* sample value of Q_{mr} be replaced by the (K+1)-tuplets $\langle v_n, w_{n1}, \cdots, w_{nK} \rangle$ where $w_{nk} = 1$ if v_n was generated by the mixture component k, for $n \in [1,N]$. The expectation-maximization (EM) algorithm [10] is a general way of iteratively estimating the tuplets and the parameters $\{\varphi_1, \ldots, \varphi_K\}$. The EM algorithm is divided into two steps:

A. Expectation Step

The expected values of the tuplets are evaluated using the current or initialized estimates of the parameters $\{\varphi_1, \ldots, \varphi_K\}$ and the weights $\mathrm{P}(k)$, $k \in [1,K]$. The expected value of w_{nk} is the probability that v_n was generated by the mixture component k, which means that

$$\mathrm{E}(w_{nk}) = \frac{\mathrm{p}(v_n \mid \varphi_k)\mathrm{P}(k)}{\mathrm{p}_{mix}(v_n \mid \varphi_1, \cdots, \varphi_K)}.$$

Thus,

$$\mathrm{E}(w_{nk}) = \frac{\mathrm{P}(k)\dfrac{u_k}{\pi\left[u_k^2 + (v_n - c_k)^2\right]}}{\displaystyle\sum_{j=1}^{K} \mathrm{P}(j)\frac{u_j}{\pi\left[u_j^2 + (v_n - c_j)^2\right]}}, \quad (5)$$

for $k \in [1,K]$ and $n \in [1,N]$.

B. Maximization Step

The current expected values of the tuplets serve as prior knowledge about the possibility that v_n was generated by the mixture component k. This information, form the expectation step, is used to optimize the parameters $\{\varphi_1, \ldots, \varphi_K\}$ using an appropriate method such as the maximum-likelihood estimation [11]. The choice of the maximization algorithm depends on the type of density being optimized. However, the weights are given by

$$\mathrm{P}(k) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}(w_{nk}). \quad (6)$$

The Cauchy density function has an undefined mean due to its infinite variance. As a result, different methods have been developed for the estimation of its scale and location parameters [12-15]. If the scale is known, the robust Pitman estimator [14, 15] is often used for point estimation of the location parameter because it is the minimum variance estimator. Therefore, in Appendix A we derive the maximization algorithm for the K location parameters using the Pitman estimator. If $\hat{c}_k$ is the Pitman estimator for the mixture component k then

$$\hat{c}_k = \sum_{j=1}^{N} \mathrm{E}(w_{jk}) v_j \frac{\Re(\Psi_{jk})}{\displaystyle\sum_{j=1}^{N} \mathrm{E}(w_{jk})\Re(\Psi_{jk})}, \quad \text{for } k \in [1,K], \quad (7)$$

where $\Re(\Psi_{jk})$ denotes the real part of

$$\Psi_{jk} = \prod_{n \neq j}\left[\frac{\mathrm{E}(w_{jk})}{(v_j - v_n)^2 + 4u_k^2}\right]\left[1 - \frac{2u_k}{(v_j - v_n)}\sqrt{-1}\right]. \quad (8)$$

The Pitman estimator is quite robust to outliers and this can be understood by considering the divisor term $(v_j - v_n)^2 + 4u_k^2$ in Ψ_{jk}. If v_j is much smaller/larger than the other sample values, its effect in the optimization of $\hat{c}_k$ is negligible. Thus Ψ_{jk} acts as some form of statistical filter placing more emphasis on the most frequent values which are also close to each other, and for a Cauchy density function these are located around the location parameter c_k. The scale parameter u_k controls the effect of the difference between a given v_j and v_n. If u_k is small then $(v_j - v_n)^2$ is dominant in $(v_j - v_n)^2 + 4u_k^2$ making Ψ_{jk} small and therefore contributing less to the optimization of $\hat{c}_k$. If u_k is larger, then $(v_j - v_n)^2$ is swamped by the $4u_k^2$ term. This reduces the selectivity of Ψ_{jk} and thus sample values larger/smaller than the most frequent ones might affect the result of $\hat{c}_k$. In fact, when the scale parameter $u_k \to 0$, the Pitman estimator acts like the sample median and inherits its superior robustness to outliers. But the median is too discriminating and can miss some important information from the "useless" sample values unlike the Pitman estimator which gives minimal uncertainty [15]. For $u_k \to \infty$ the Pitman estimator acts like the sample mean and is non-ideal in this case given thick tails of the ratio sample Q_{mr}. As a result, it is necessary to choose a small value of u_k to inherit the robustness of the sample median yet

achieving a more accurate result.Based on the analysis of (7) and (8), it is clear that the estimator $\hat{c}_k$ is focused mostly on sample values of Q_{mr} for $n \in \mathrm{T}_n(S_k)$. The uncertainty or variance of the Pitman estimators is derived in Appendix B. The developed EM algorithm steps are:

1. Initialize the parameters $\{\varphi_1, \cdots, \varphi_K\}$ and the weights $\mathrm{P}(k)$, where $\varphi_k = \{c_k, u_k\}$.
2. Expectation step: Evaluate the tuplets $\mathrm{E}(w_{nk})$ based on the current values of the parameters and weights via (5).
3. Maximization step: Optimize the location parameters $\{c_1, \cdots, c_K\}$ using (7) and the weights using (6) using the current values of the tuplets.
4. Repeat steps 2 and 3 until convergence or some condition such as the number of iterations is met.

According to (2), $\hat{c}_k$ is the ratio of a_{mk} to a_{rk}, therefore

$$a_{mk} = a_{rk}\hat{c}_k \quad \text{for} \quad r, m \in [1, M] \text{ and } k \in [1, K]. \tag{9}$$

It is necessary to choose a value for a_{rk} in order to get a_{mk}, for all k. For example, if the r-*th* row of **A** is set to $a_{r1} = a_{r2} = \cdots = a_{rK} = \lambda$, for some scalar λ, then by (9) $a_{m1} = \lambda a_{r1}$ up to $a_{mK} = \lambda a_{rK}$. Inevitably there is a scaling ambiguity between the estimated OD **A** and the unknown mixing OD. It is essential to set all elements of **A** is the row of X_r to the same value λ so that the scaling ambiguity is constant for all column vectors, otherwise the estimated OD will be incorrect.

For dimensionalities M>2, to maintain a constant scaling ambiguity for all the elements of **A**, the variable X_r corresponding to the initialized row is used repeatedly with all the other M-1 variables. With X_r as the reference row variable, all the M-1 row elements of **A** are scaled with respect to its *r-th* row elements. This ensures that **A** is just an ambiguously scaled version of the unknown mixing dictionary, which is usually unavoidable [7].

IV. Sparse Decomposition

Signal decomposition under an overcomplete dictionary is not unique due to the under-determined data space. This degeneracy can be circumvented by making a statistical assumption about the nature of the underlying sources. If the sources are sparse, the problem can be solved by minimizing the ℓ^0-norm resulting in an method known as the SL0 algorithm [16, 17] which is usually faster and more accurate than the well known basis pursuit [18], matching pursuit [19] and FOCUSS [20]. After estimating the overcomplete dictionary **A**, we use the SL0 algorithm to find the sparse representations in the simulations.

V. Simulation Results

Fig. 2 shows K=3 speech signals which are mixed down to M=2 instantaneous mixtures given in Fig. 3 using the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 4 & 6 \end{bmatrix}. \tag{10}$$

After forming the ratio sample $Q_{21} = X_2 / X_1$, the CMM algorithm is used to learn the ratios of the elements of the column vectors of **A** via estimation of the location parameters of the Cauchy mixture components.

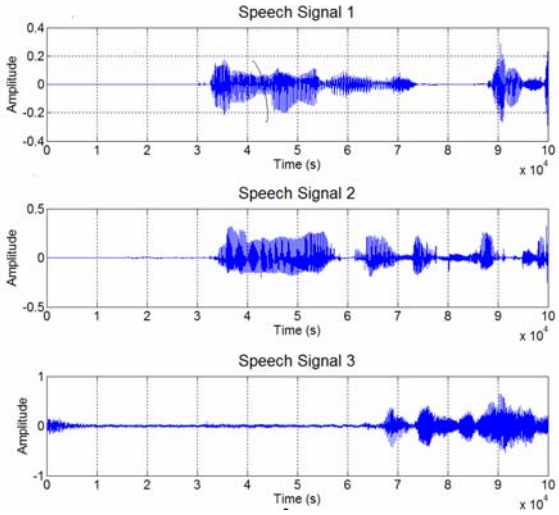

Fig. 2 Speech signals are sparse and occasionally only one is observed at some instance. These signals are used to give two mixtures thus forming an under-determined signal space

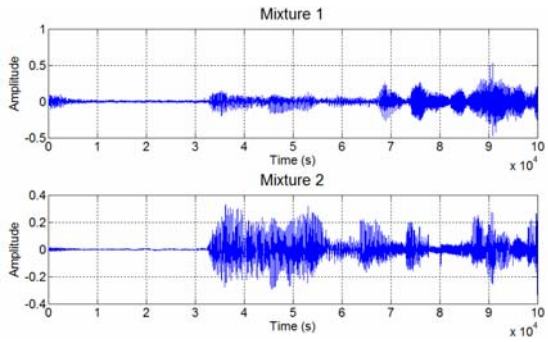

Fig. 3 Two instantaneous mixtures obtained using three speech signals and an OD of dimensions $M = 2$ and $K = 3$. In each mixture signal there are three sources and therefore the system is under-determined

The CMM results are plotted in Fig. 4 for three different scale parameters of choice. The values of (2) for the mixing matrix given in (10) are $Q_{21} = -1$ for $n \in \mathrm{T}_n(S_1)$, $Q_{21} = 4$ for $n \in \mathrm{T}_n(S_2)$ and $Q_{21} = 6$ for $n \in \mathrm{T}_n(S_3)$. These values coincide with the location parameters of each of the three learned CMMs. As the scale parameter is reduced, the resolution of the CMM about the location parameters is increased due to the

high selectivity of Ψ_{jk}. In all three cases, it is clear that the CMM algorithm has accurately determined the overcomplete dictionary **A**.

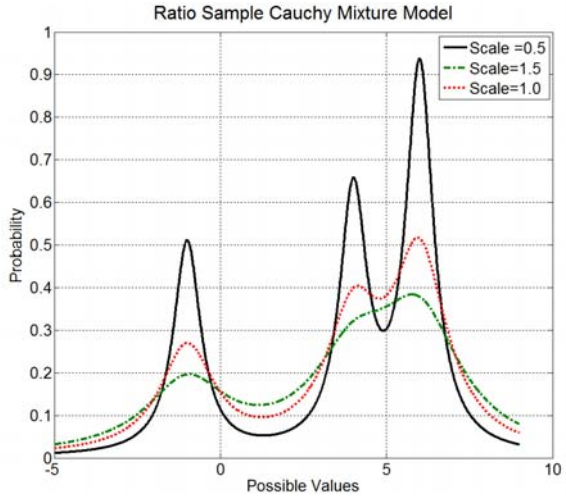

Fig. 4 Each Cauchy mixture model has three modes corresponding to the ratios of the column vector elements. In each case, the estimated location parameters are at $\hat{c}_1 = -1$, $\hat{c}_2 = 4$ and $\hat{c}_3 = 6$. As the scale parameter is reduced, the resolution of the algorithm increases

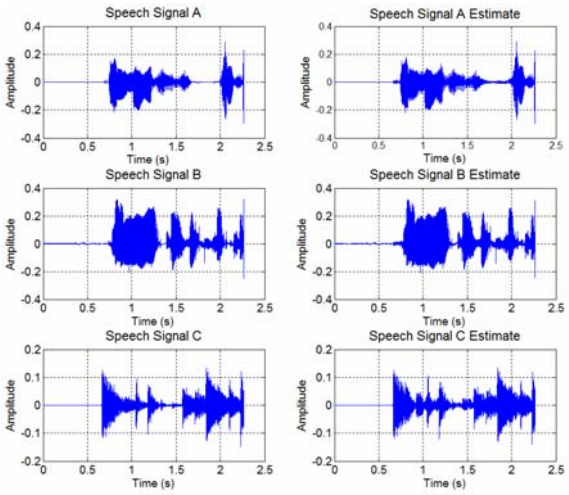

Fig. 5 The SL0 algorithm is used to infer the original sources using the overcomplete dictionary learned by the CMM algorithm. The obtained representations resemble the original sources

Fig. 5 shows the resulting source estimates alongside the original sources after learning **A** and using the SL0 algorithm for sparse source separation. Although there are some slight discrepancies between a given source and its estimate, this is a limitation of the SL0 algorithm since the OD has been accurately determined by the CMM method.

In [1], Lewicki and Sejnowski derive an algorithm for learning an overcomplete dictionary by maximizing the data likelihood over the basis functions. The Lewicki-Sejnowski (LS) algorithm optimization steps interchange between estimating the dictionary and inferring the sparse sources. All sample values from the available mixtures weigh an equal amount in the estimation process unlike in the CMM algorithm where the Pitman estimators place more emphasis on the sample values for the time instances $n \in \mathrm{T}_n(S_k)$.

To illustrate the efficiency of the CMM approach, some Gaussian noise is added to the mixture signals where in table I the signal-to-noise ratio is 12dB and in table II it is reduced to 9dB. In table II, the new OD is such that $K = 4$ (i.e. four source signals) while maintaining $M = 2$ mixture signals to compare the accuracy of the methods with an increase in uncertainty.

TABLE I
COLUMN VECTOR ESTIMATION SNR=12 DECIBELS

Column Vector Ratio	LS Error	CMM Error $(u_k = 0.01)$
$\mathbf{a}_1 = -2$	0.38	0
$\mathbf{a}_2 = 3$	0.14	0.05
$\mathbf{a}_3 = -1$	0.23	0

TABLE II
COLUMN VECTOR ESTIMATION SNR=9 DECIBELS

Column Vector Ratio	LS Error	CMM Error $(u_k = 0.01)$
$\mathbf{a}_1 = -2$	1.67	0.11
$\mathbf{a}_2 = 3$	3.44	0.05
$\mathbf{a}_3 = -1$	2.18	0.18
$\mathbf{a}_4 = 4$	6.77	0.13

In tables I and II, the error is given by the modulus of the ratio deviation to eliminate scaling ambiguities. That is, from the estimated OD the ratio of the elements of each column vector is computed and subtracted from the true column ratio of the mixing matrix. In table I, for $K = 3$ sources with and SNR of 12 dB the LS algorithm performs reasonably well but the CMM method produced better results. This is a consequence of using all the samples equally to estimate the location parameters or column ratios. The CMM is only slightly affected by the additive Gaussian noise due to its focus on sample values close to each other, much like the sample median.It is observed in table II that as the number of sources and additive noise increase the LS algorithm performance degrades significantly, while the CMM algorithm continues to be robust. Besides the increased noise, from [16]-[20] it is known that estimating the sources under an overcomplete dictionary is sensitive to an increase in the number of basis vectors. Since the LS method requires the estimation of the sparse sources in order to optimize the dictionary, it suffers from this computational uncertainty. In contrast, the CMM algorithm determines the OD first, and then a sparse decomposition algorithm like the SL0 technique is used for sparse representations. The result is a better source separation result albeit noisy extracted sources.

VI. Discussion

A new algorithm for learning an OD has been introduced with the aim of increased robustness to outliers which often affect the accuracy of data mining methods using all the sample values equally in the optimization process. By forming a ratio sample from the observed mixture signals, the ratios of the mixing dictionary column elements coincide with the location parameter of the CMM. By study, these ratios (i.e. basis vector directions) are fully determined by the sample values when only one of the sparse sources is active. The Pitman estimator places more emphasis on the most frequent samples values closer to the Cauchy location parameters and less on those constituting the heavy tails. This allows the CMM algorithm to be robust to outliers much like the sample median (although not as robust) but with better accuracy. On the downside, to form a ratio sample it is mandatory to have at least two mixture signals but not necessarily more, and by derivation the algorithm is limited to sparse and statistically independent source signals.

Appendix

A. Pitman Estimators for the CMM Location Parameters

Let the maximum a posterior estimator [15] for the location parameter of the mixture component k be $\hat{c}_k$, then

$$\hat{c}_k = \int_{-\infty}^{\infty} c_k \, \mathrm{p}\left(c_k \mid v, u_k\right) dc_k \, .$$

Assuming a non-informative prior $\mathrm{p}(c_k)$ and leaving out the normalization factors, then by Bayes' rule

$$\begin{aligned} \hat{c}_k &= \int_{-\infty}^{\infty} c_k \, \mathrm{p}\left(v \mid c_k, u_k\right) dc_k \\ &= \int_{-\infty}^{\infty} c_k \prod_{n=1}^{N} \mathrm{p}\left[v_n \mid c_k, u_k, \mathrm{E}\left(w_{nk}\right)\right] dc_k \end{aligned},$$

since the sources are statistically independent and $\mathrm{p}[v_n \mid c_k, u_k, \mathrm{E}(w_{nk})] = \mathrm{E}(w_{nk})\,\mathrm{p}(v_n \mid c_k, u_k)$, with $\mathrm{E}(w_{nk})$ as the prior knowledge that v_n is drawn from mixture component k. Let

$$\begin{aligned} \mathrm{I}_{rk}\left(v\right) &= \int_{-\infty}^{\infty} c_k^r \prod_{n=1}^{N} \mathrm{E}\left(w_{nk}\right) \mathrm{p}\left(v_n \mid c_k, u_k\right) dc_k \\ &= \int_{-\infty}^{\infty} c_k^r \prod_{n=1}^{N} \mathrm{E}\left(w_{nk}\right) \frac{u_k}{\pi\left[u_k^2 + \left(v_n - c_k\right)^2\right]} dc_k \qquad (11) \\ &= \left(\frac{u_k}{\pi}\right)^N \int_{-\infty}^{\infty} \mathrm{g}_r\left(c_k\right) dc_k , \end{aligned}$$

for $r \in [0, 2N-1)$, where

$$\mathrm{g}_r\left(c_k\right) = c_k^r \prod_{n=1}^{N} \frac{\mathrm{E}\left(w_{nk}\right)}{\left(c_k - c_{nk}^{+}\right)\left(c_k - c_{nk}^{-}\right)},$$

with $c_{nk}^{\pm} = v_n \pm u_k\sqrt{-1}$. Using contour integration [21],

$$\mathrm{I}_{rk}\left(v\right) = \left(\frac{u_k}{\pi}\right) \lim_{R\to\infty} \int_{C_R} \mathrm{g}_r\left(c_k\right) dc_k \, ,$$

where $C_R : z = R \cdot \exp(-\phi\sqrt{-1})$. The integral is evaluated along the counter-clockwise contour spanning the range [-*R*,*R*] followed by the upper half circumference of a circle of radius *R* centered at (0,0), the origin. By the residue theorem [22],

$$\mathrm{I}_{rk}\left(v\right) = \left(\frac{u_k}{\pi}\right)^N 2\pi\sqrt{-1} \sum_{j=1}^{N} \mathrm{Res}\left[\mathrm{g}_r\left(c_k\right), c_{jk}^{+}\right], \qquad (12)$$

where only the poles in the upper half plane are considered due to the Cauchy integral theorem [23]. The order of the poles in $\mathrm{g}_r(c_k)$ is $m = 1$, thus

$$\begin{aligned} \mathrm{Res}\left[\mathrm{g}_r\left(c_k\right), c_{jk}^{+}\right] &= \lim_{c_k \to c_{jk}^{+}} \left[\left(c_k - c_{jk}^{+}\right) c_k^r \prod_{n=1}^{N} \frac{\mathrm{E}\left(w_{nk}\right)}{\left(c_k - c_{nk}^{+}\right)\left(c_k - c_{nk}^{-}\right)}\right] \\ &= \frac{\left(c_{jk}^{+}\right)^r \mathrm{E}\left(w_{nk}\right)}{\left(c_{jk}^{+} - c_{jk}^{-}\right)} \prod_{n\neq j} \frac{\mathrm{E}\left(w_{nk}\right)}{\left(c_{jk}^{+} - c_{nk}^{+}\right)\left(c_{jk}^{+} - c_{nk}^{-}\right)} . \end{aligned}$$

As $(c_{jk}^{+} - c_{jk}^{-}) = 2u_k\sqrt{-1}$, $(c_{jk}^{+} - c_{nk}^{+}) = (v_j - v_n)$ and $(c_{jk}^{+} - c_{nk}^{-}) = (v_j - v_n + 2u_k\sqrt{-1})$,

$$\begin{aligned} \mathrm{Res}\left[\mathrm{g}_r\left(c_k\right), c_{jk}^{+}\right] &= \frac{\left(c_{jk}^{+}\right)^r \mathrm{E}\left(w_{jk}\right)}{\left(c_{jk}^{+} - c_{jk}^{-}\right)} \\ &\prod_{n\neq j} \left[\frac{\mathrm{E}\left(w_{nk}\right)}{\left(v_j - v_n\right)^2 + 4u_k^2}\right] \left[1 - \frac{2u_k\sqrt{-1}}{\left(v_j - v_n\right)}\right]. \end{aligned} \qquad (13)$$

The first moment or Pitman estimator of the mixture component k is given by

$$\hat{c}_k = \frac{\mathrm{I}_{1k}\left(v\right)}{\mathrm{I}_{0k}\left(v\right)} . \qquad (14)$$

Noting that the integrals in (11) are real and after substituting (13) into (12),

$$\begin{aligned} \mathrm{I}_{0k}\left(v\right) &= \left(\frac{u_k}{\pi}\right)^{N-1} \sum_{j=1}^{N} \mathrm{E}\left(w_{jk}\right) \Re\left(\Psi_{jk}\right), \\ \mathrm{I}_{1k}\left(v\right) &= \left(\frac{u_k}{\pi}\right)^{N-1} \sum_{j=1}^{N} \mathrm{E}\left(w_{jk}\right) v_j \Re\left(\Psi_{jk}\right), \end{aligned} \qquad (15)$$

where,

$$\Psi_{jk} = \prod_{n \neq j} \left[\frac{\mathrm{E}(w_{nk})}{(v_j - v_n) + 4u_k^2} \right] \left[1 - \frac{2u_k\sqrt{-1}}{(v_j - v_n)} \right]. \quad (16)$$

Therefore by (14), the Pitman estimator for mixture component k is given by

$$\hat{c}_k = \sum_{j=1}^{N} \mathrm{E}(w_{jk}) v_j \frac{\Re(\Psi_{jk})}{\sum_{j=1}^{N} \mathrm{E}(w_{jk}) \Re(\Psi_{jk})}. \quad (17)$$

B. Variance of the CMM Pitman Estimators

For the number of samples $N \geq 3$, the Pitman estimator is an unbiased estimator of the location parameter of the Cauchy density function [15]. That is

$$\mathrm{E}(\hat{c}_k - c_k) = \int_{-\infty}^{\infty} (\hat{c}_k - c_k)\, \mathrm{p}(v \mid c_k, u_k)\, dv = 0.$$

Differentiate with respect to c_k using the product rule and rearrange to give

$$\int_{-\infty}^{\infty} (\hat{c}_k - c_k)^2 \sqrt{\mathrm{p}(v \mid c_k, u_k)} \cdot \sqrt{\mathrm{p}(v \mid c_k, u_k)} \frac{\partial}{\partial c_k} \ln \mathrm{p}(v \mid c_k, u_k)\, dv \geq \int_{-\infty}^{\infty} \mathrm{p}(v \mid c_k, u_k)\, dv.$$

By the Cauchy-Schwartz inequality,

$$\int_{-\infty}^{\infty} (\hat{c}_k - c_k)^2\, \mathrm{p}(v \mid c_k, u_k)\, dv \geq \frac{\left[\int_{-\infty}^{\infty} \mathrm{p}(v \mid c_k, u_k)\, dv \right]^2}{\int_{-\infty}^{\infty} \mathrm{p}(v \mid c_k, u_k) \left[\frac{\partial}{\partial c_k} \ln \mathrm{p}(v \mid c_k, u_k) \right]^2 dv},$$

or,

$$\mathrm{var}(\hat{c}_k) \geq \frac{\left[\int_{-\infty}^{\infty} \mathrm{p}(v \mid c_k, u_k)\, dv \right]}{\Gamma_N(c_k)}, \quad (18)$$

where $\Gamma_N(c_k)$ is known as the *N*-sample Fisher information [15], and $\mathrm{var}(\hat{c}_k)$ is the variance or uncertainty associated with the estimator. The inequality suggests that the precision to which we can estimate c_k is fundamentally limited by the reciprocal of the *N*-sample Fisher information multiplied by the square of the sum of the possible probabilities for the mixture component k on the sample values of the ratio sample Q_{mr}. Evaluating the numerator,

$$\int_{-\infty}^{\infty} \mathrm{p}(v \mid c_k, u_k)\, dv = \int_{-\infty}^{\infty} \prod_{n=1}^{N} \mathrm{E}(w_{nk})\, \mathrm{p}(v_n \mid c_k, u_k)\, dv.$$

The variance of $\hat{c}_k$ is evaluated after convergence, which means that we can substitute for $\mathrm{p}(v_n \mid c_k, u_k)$ using

$$\mathrm{p}(v_n \mid c_k, u_k) = \frac{\mathrm{E}(w_{nk})}{\mathrm{P}(k)} \mathrm{p}_{\mathrm{mix}}(v_n \mid \varphi_1, \ldots, \varphi_K).$$

As a result,

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{p}(v \mid c_k, u_k)\, dv &= \int_{-\infty}^{\infty} \prod_{n=1}^{N} \mathrm{E}(w_{nk})\, \mathrm{p}(v_n \mid c_k, u_k)\, dv \\ &= \prod_{n=1}^{N} \frac{\mathrm{E}^2(w_{nk})}{\mathrm{P}(k)} \int_{-\infty}^{\infty} \prod_{n=1}^{N} \mathrm{p}_{\mathrm{mix}}(v_n \mid \varphi_1, \ldots, \varphi_K)\, dv \\ &= \prod_{n=1}^{N} \frac{\mathrm{E}^2(w_{nk})}{\mathrm{P}(k)}, \end{aligned} \quad (19)$$

for $k \in [1,K]$. Since $\mathrm{p}(v \mid c_k, u_k)$ factors (i.e. the sources are statistically independent), it can be shown that the *N*-sample Fisher information is given by

$$\Gamma_N(c_k) = \sum_{n=1}^{N} \Gamma_n(c_k), \quad (20)$$

for the single-sample Fisher information $\Gamma_n(c_k)$. Thus

$$\begin{aligned} \Gamma_n(c_k) &= \int_{-\infty}^{\infty} \mathrm{E}(w_{nk})\, \mathrm{p}(v_n \mid c_k, u_k) \frac{\partial}{\partial c_k} \ln \left[\mathrm{E}(w_{nk})\, \mathrm{p}(v_n \mid c_k, u_k) \right] dv \\ &= \frac{4u_k}{\pi} \mathrm{E}(w_{nk}) \int_{-\infty}^{\infty} \frac{(v_n - c_k)^2}{\left[u_k^2 + (v_n - c_k)^2 \right]^3} dv. \end{aligned}$$

Using contour integration and applying the residue theorem,

$$\begin{aligned} \Gamma_n(c_k) &= \frac{4u_k}{\pi} \mathrm{E}(w_{nk}) \lim_{R \to \infty} \int_{C_R} \mathrm{g}(v_n)\, dv \\ &= \frac{4u_k}{\pi} \mathrm{E}(w_{nk}) \cdot 2\pi\sqrt{-1} \cdot \mathrm{Res}\left[\mathrm{g}(v_n), v_k^+ \right], \end{aligned} \quad (21)$$

where

$$\mathrm{g}(v_n) = \frac{(v_n - c_k)^2}{\left[u_k^2 + (v_n - c_k)^2 \right]^3} = \frac{(v_n - c_k)^2}{(v_n - v_k^+)^3 (v_n - v_k^-)^3},$$

for $v_k^{\pm} = c_k \pm u_k\sqrt{-1}$. The poles of $\mathrm{g}(v_n)$ are of order $m = 3$, thus

$$\begin{aligned}&\mathrm{Res}\left[\mathrm{g}\left(v_n\right), v_k^+\right]\\&= \lim_{v_n \to v_k^+}\left[\frac{1}{2}\cdot\frac{d^2}{dv_n^2}\left(v_n - v_k^+\right)^3 \cdot \frac{\left(v_n - c_k\right)^2}{\left(v_n - v_k^+\right)^3\left(v_n - v_k^-\right)^3}\right] \quad (22)\\&= -\frac{\sqrt{-1}}{16u_k^3}.\end{aligned}$$

Substituting (22) into (21) and evaluating (20),

$$\Gamma_N\left(c_k\right) = \sum_{n=1}^{N}\Gamma_n\left(c_k\right) = \sum_{n=1}^{N}\frac{\mathrm{E}\left(w_{nk}\right)}{2u_k^2}. \quad (23)$$

From (18), (19) and (23)

$$\mathrm{var}\left(\hat{c}_k\right) \geq \frac{2u_k^2}{\mathrm{P}^{2N}\left(k\right)}\cdot\frac{\mathrm{E}^{(2N+4)}\left(w_{nk}\right)}{\sum_{n=1}^{N}\mathrm{E}\left(w_{nk}\right)}, \quad (24)$$

for $k \in [1, K]$.

REFERENCES

[1] M. S. Lewicki and T. J. Sejnowski, "Learning Overcomplete Representations, " *Neural Computations*, vol. 12, No. 2, pp. 337-365, February 2000.
[2] M. Zhong, H. Tang, H. Cheng and Y. Tang, "An EM Algorithm for Learning Sparse and Overcomplete Representations, " *Neurocomputing*, vol. 57, pp. 467-476, 2004.
[3] M. Aharon, M. Elad and A. Bruckstein, "K-SVD: An algorithm for Designing Overcomplete Dictionaries for Sparse Representations," *IEEE Transactions on Signal Processing*, vol. 54, No. 11, November 2006.
[4] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Egan, T. Lee and T. J. Sejnowski, "Dictionary Learning Algorithms for Sparse Representations, " *eural Computations*, vol. 15, No. 2, pp. 349-396, 2003.
[5] K. Egan, S. O. Aase and J. H. Hakon-Husoy, "Method of Optimal Directions for Frame Design, "in *IEEE International Conference of Acoustic, Speech and Signal Processing*, Vol. 5, pp. 2443-2446, 1999.
[6] I. T. Jollife, *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd Edition, .
[7] P. Common, "Independent Component Analysis: A new Concept?, " *Signal Processing*, vol. 36, pp. 287-314, 1994.
[8] G. Marsaglia, "Ratios of Normal Variables, " *Journal of Statistical* Software, vol. 16, No. 4, May 2006.
[9] W. Feller, *An Introduction to Probability Theory and Applications*, 2nd Edition, New York: Wiley, pp. 229-235, 1968.
[10] P. G. Hoel, *Introduction to Mathematical Statistics*, 3rd Edition, New York: Wiley, pp. 57-62, 1962.
[11] J. W. Harris and H. Stoker, "Maximum Likelihood Method, " *in Handbook of Mathematics and Computational Science*, New York: Springer-Verlag, pp. 824-835, 1998.
[12] A. Koutrouvelis, "Estimation of the Location and Scale of the Cauchy Distribution using the Empirical Characteristic Function, " *Biometrica*, Vol. 69, pp. 205-213, 1982.
[13] F. Nagy, "Parameter Estimation of the Cauchy Distribution in Information Theory Approach, "*Journal of Universal Computer Science*, Vol. 12, No. 9, pp. 1332-1344, May 2006.
[14] G. B. Freue, "The Pitman Estimator of the Cauchy Location Parameter, " *Journal of Statistical Planning and Inference*, vol. 137, pp. 1900-1913, May 2006.
[15] K. M. Hanson and D. R. Wolf, "Estimators for the Cauchy Distribution, " *In Maximum Entropy and Bayesian Methods*, pp. 255-263, 1996.
[16] C. Jutten, G. H. Mohimani and M. B. Zadeh, "Fast Sparse Representations based on the l0 Norm, " In *Pro. ICA'07*, London, UK, 2007.
[17] G. H. Mohimani, M. B. Zaden and C. Jutten, "Complex Valued Sparse Representations based on Smoothed l0 Norm, " In *ICASSP'08*, 2008.
[18] S. S. Chen, D. L. Donoho and M. A. Saunders, "Atomic Decomposition by Basis Pursuit, " *SIAM Journal of Computing*, Vol. 20, No. 1, pp. 33-61, 1999.
[19] S. G. Mallat and Z. Zhang, "Matching Pursuit with Time-Frequency Dictionaries, " *IEEE Transactions on Signal Processing*, pp. 3397-3415, December 1993.
[20] I. F. Gorodnitsky and B. D. Rao, "Sparse Signal Reconstruction from Limited Data using FOCUSS: A Re-weighted Minimum Norm Algorithm, " *IEEE Transactions on Signal Processing*, Vol. 45, No. 3, March 1997.
[21] J. E. Marsden and J. M. Hoffman, *Basic Complex Analysis*, 3rd Edition, W. H. Freeman, 1998, ISBN: 978-0716728771.