JENSEN–RÉNYI DIVERGENCE FOR SOURCE SEPARATION ON THE TIME–FREQUENCY PLANE

Zeyong Shan and Selin Aviyente

Department of Electrical and Computer Engineering
Michigan State University, East Lansing, MI 48824, USA
e-mail: {shanzeyo, aviyente@egr.msu.edu}

ABSTRACT

Blind source separation aims at recovering the original source signals given only observations of their mixtures. Some common approaches to the source separation problem include second or higher order statistics based methods, and independent component analysis. Most of these methods are developed in the time domain, and thus, inherently assume the stationarity of the underlying signals. Since most real life signals of interest are non–stationary, there have been efforts to perform source separation in the time–frequency domain. In this paper, we propose a new approach for source separation on the time–frequency plane using an information–theoretic cost function. Jensen–Rényi divergence, as adapted to time–frequency distributions, is introduced as an effective cost function to extract sources that are disjoint on the time–frequency plane. The sources are extracted through a series of Givens rotations and the optimal rotation angle is found using the steepest descent algorithm. The performance of the proposed method is illustrated and quantified through examples.

1. INTRODUCTION

Blind source separation (BSS) is an important and fundamental problem in signal processing with a broad range of applications. A number of BSS algorithms have been proposed based on the instantaneous mixture model, in which the observed signals are linear combinations of the source signals. Among these methods, the most common ones are second order statistics based methods [1], and information–theoretic approaches which utilize cost functions such as mutual information or divergence measures, e.g. independent component analysis (ICA) [2, 3]. These methods in general assume a certain structure for the underlying source signals. For example, higher–order statistics based methods assume non–Gaussian and i.i.d source signals, whereas ICA assumes the independence of the source signals.

Most real life signals are non–stationary, and thus do not obey the underlying assumption of stationarity that is embedded in the current methods. For this reason, recently various methods have been introduced to exploit the non–stationarity of the source signals. Researchers have resorted to the powerful tool of time–frequency signal representations to solve the source separation problem. For non–stationary signals, a blind separation approach using a spatial time-frequency distribution is proposed in [4] and the separation is achieved by joint diagonalization of the auto–terms in the spatial time-frequency distributions.

In this paper, we introduce a new approach to the source separation problem combining time–frequency representations with information–theoretic measures. An information–theoretic criterion, Jensen–Rényi divergence as adapted to the time–frequency distributions, is used as the objective function to separate the sources. The underlying sources are assumed to be disjoint on the time–frequency plane and it is shown that this new cost function achieves its maximum when the signals are disjoint. With the assumption that the source signals are disjoint on the time–frequency plane, signal separation is performed through a rotation transformation using a steepest descent algorithm.

2. BACKGROUND ON TIME-FREQUENCY DISTRIBUTIONS AND INFORMATION MEASURES

A time-frequency distribution (TFD), $X(t,\omega)$, from Cohen's class can be expressed as [1] [5]:

$$X(t,\omega)=\int\int\int \phi(\theta,\tau)s(u+\frac{\tau}{2})s^*(u-\frac{\tau}{2})e^{j(\theta u-\theta t-\omega\tau)}du\,d\theta\,d\tau, \tag{1}$$

where $\phi(\theta,\tau)$ is the kernel function and s is the signal. Some of the most desired properties of TFDs are the energy preservation and the marginals. They are satisfied when $\phi(\theta,0)=\phi(0,\tau)=1\ \ \forall\tau,\theta$ and are given as follows:

$$\begin{aligned}\int\int X(t,\omega)\,dt\,d\omega &= \int |s(t)|^2\,dt=\int |S(\omega)|^2\,d\omega,\\ \int X(t,\omega)\,d\omega &= |s(t)|^2\ \ ,\int X(t,\omega)\,dt=|S(\omega)|^2.\end{aligned} \tag{2}$$

The formulas given above evoke an analogy between a TFD and the probability density function (pdf) of a two–dimensional random variable. This analogy has inspired the adaptation of information–theoretic measures such as entropy to the time–frequency plane [6]. Although entropy measures have proven to be useful in quantifying the complexity of individual signals, they cannot be used directly to quantify the difference between signals. For this reason, well–known divergence measures from information theory have been adapted to the time–frequency plane [7, 8]. One such distance measure is the Jensen–Rényi divergence based on the Jensen difference. Jensen–Rényi divergence is the modification of Jensen–Shannon divergence from an arithmetic to a geometric mean introduced by Michel [7]. For time–frequency distributions, Jensen–Rényi divergence can be defined as:

$$G_{12}^{\alpha}(X_1,X_2)=H_\alpha(\sqrt{X_1X_2})-\frac{H_\alpha(X_1)+H_\alpha(X_2)}{2}, \tag{3}$$

[1] All integrals are from $-\infty$ to ∞ unless otherwise stated.

where H_α represents Rényi entropy defined on the time–frequency plane as:

$$H_\alpha(X) = \frac{1}{1-\alpha}\log_2 \int\int \left(\frac{X(t,\omega)}{\int\int X(u,v)du\,dv} \right)^\alpha dt\,d\omega, \tag{4}$$

where $\alpha > 0$. Jensen–Rényi divergence is equal to zero when $X_1 = X_2$, and its positivity can be proven using the Cauchy–Schwartz inequality. This measure has some desired properties such as being symmetric and monotonically increasing as the overlap between the two distributions decreases, i.e. $G^\alpha_{12}(X_1, X_2) \to \infty$ as $X_1(t,\omega) \times X_2(t,\omega) \to 0$. Therefore, maximizing this measure corresponds to obtaining disjoint time–frequency representations.

3. PROBLEM FORMULATION AND METHOD

3.1. Problem Statement in The Time–Frequency Domain

In this paper, we consider the problem of determining the source signals when the number of observed mixtures is equal to or greater than the number of the source signals. Assume that the M mixtures, $\{s_1(t), s_2(t), \cdots, s_M(t)\}$, of the N non-stationary complex source signals are given ($M \geq N$). Each mixture, $s_i(t)$, is first transformed to the time-frequency plane as:

$$X_i(n,\omega;\psi) = \sum_m \sum_l \psi(n-l,m) s_i\left(l+\frac{m}{2}\right) s_i^*\left(l-\frac{m}{2}\right) e^{-j\omega m}. \tag{5}$$

The time-frequency distribution corresponding to each mixture is vectorized and a matrix of time–frequency distributions is formed:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_M \end{bmatrix} = \begin{bmatrix} X_1(1) & \cdots & X_1(Q) \\ X_2(1) & \cdots & X_2(Q) \\ & \vdots & \\ X_M(1) & \cdots & X_M(Q) \end{bmatrix}, \tag{6}$$

where $\mathbf{X}_i$ is a vector of length $Q = K \times L$ points, K and L are the number of time and frequency points, respectively. The signals to be separated on the time–frequency plane are defined as:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_N \end{bmatrix} = \begin{bmatrix} Y_1(1) & \cdots & Y_1(Q) \\ Y_2(1) & \cdots & Y_2(Q) \\ & \vdots & \\ Y_N(1) & \cdots & Y_N(Q) \end{bmatrix}. \tag{7}$$

In order to make the following discussions simpler, we concentrate on the case where $M = N$. The discussions can be generalized for $M > N$ as illustrated through an example in Sect. 4.

The sources $\mathbf{Y}$ are extracted by applying a rotation transform $\mathbf{R}(\theta)$ in N–dimensions:

$$\mathbf{Y} = \mathbf{R}(\theta)\mathbf{X}. \tag{8}$$

Rotation matrix is used for extracting the sources since any unitary transform can be written in terms of rotation matrices and it provides a convenient parametrization of the problem. The rotation angle θ is adapted to maximize the following cost function:

$$G_\alpha \triangleq \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \left[H_\alpha(\sqrt{\mathbf{Y}_i\mathbf{Y}_j}) - \frac{H_\alpha(\mathbf{Y}_i) + H_\alpha(\mathbf{Y}_j)}{2} \right]. \tag{9}$$

Maximizing this cost function will ensure that the extracted components do not overlap with each other on the time–frequency plane.

3.2. Cost Function

The Jensen–Rényi divergence between two time–frequency distributions is defined as:

$$G^\alpha_{ij} = H_\alpha(\sqrt{\mathbf{Y}_i\mathbf{Y}_j}) - \frac{H_\alpha(\mathbf{Y}_i) + H_\alpha(\mathbf{Y}_j)}{2}. \tag{10}$$

This expression can be further simplified as:

$$\begin{aligned} G^\alpha_{ij} &= \frac{1}{1-\alpha}\log\left[\sum_{k=1}^{Q}\left(\sqrt{Y_i(k)Y_j(k)}\right)^\alpha\right] \\ &\quad - \frac{1}{2(1-\alpha)}\left[\log\left(\sum_{k=1}^{Q} Y_i^\alpha(k)\right) + \log\left(\sum_{k=1}^{Q} Y_j^\alpha(k)\right)\right] \\ &= \frac{1}{1-\alpha}\log\left[\frac{\sum_{k=1}^{Q}\left(\sqrt{Y_i(k)Y_j(k)}\right)^\alpha}{\sqrt{\left(\sum_{k=1}^{Q} Y_i^\alpha(k)\right)\left(\sum_{k=1}^{Q} Y_j^\alpha(k)\right)}}\right], \end{aligned} \tag{11}$$

which represents the ratio of the energy of the overlap between the two TFDs to the product of the energy of the individual TFDs. Let

$$J^\alpha_{ij} = \frac{\sum_{k=1}^{Q}\left(\sqrt{Y_i(k)Y_j(k)}\right)^\alpha}{\sqrt{\left(\sum_{k=1}^{Q} Y_i^\alpha(k)\right)\left(\sum_{k=1}^{Q} Y_j^\alpha(k)\right)}}, \tag{12}$$

and

$$J_\alpha = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J^\alpha_{ij}. \tag{13}$$

Since log is a monotonous function, maximizing G_α is equivalent to minimizing J_α for $\alpha > 1$, or maximizing J_α for $\alpha < 1$. This means that we can equivalently use J_α as our cost function. In this paper, we will consider orders of $\alpha > 1$. The results are similar for $\alpha < 1$. One special case of $\alpha > 1$ is the quadratic one when $\alpha = 2$. When $\alpha = 2$, the cost function J_α simplifies to:

$$J_2 = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left[\frac{\sum_{k=1}^{Q} Y_i(k)Y_j(k)}{\sqrt{\left(\sum_{k=1}^{Q} Y_i^2(k)\right)\left(\sum_{k=1}^{Q} Y_j^2(k)\right)}}\right]. \tag{14}$$

In this paper, we will use $\alpha = 2$ since the Rényi entropy will be well–defined for this order even when the distributions are non-positive.

3.3. Rotation

In N–dimensional space the simplest rotation is in the two–dimensional plane. If a rotation is through an angle θ_{ab} in the $a-b$ plane, then the rotation matrix $\mathbf{R}_{ab}(\theta_{ab})$ is:

$$\mathbf{R}_{ab}(\theta_{ab}) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\theta_{ab}) & \cdots & \sin(\theta_{ab}) & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -\sin(\theta_{ab}) & \cdots & \cos(\theta_{ab}) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}, \tag{15}$$

where $\mathbf{R}_{ab}(\theta_{ab})$ equals the $N \times N$ identity matrix $\mathbf{I}_N$ except that the elements $I_N(a,a), I_N(a,b), I_N(b,a)$, and $I_N(b,b)$ are replaced by $\cos(\theta_{ab}), \sin(\theta_{ab}), -\sin(\theta_{ab})$, and $\cos(\theta_{ab})$, respectively, where $I_N(a,b)$ is the element of $\mathbf{I}_N$ located at the ath row and bth column. From [9], we know that any N–dimensional rotation matrix can be written as the product of $N(N-1)/2$ two–dimensional–plane N–dimensional rotation matrices,which is:

$$\mathbf{R}(\theta) = \mathbf{R}_{12}(\theta_{12}) \cdots \mathbf{R}_{ab}(\theta_{ab}) \cdots \mathbf{R}_{(N-1)N}(\theta_{(N-1)N}), \quad (16)$$

where $\theta = [\ \theta_{12}, \cdots, \theta_{ab}, \cdots, \theta_{(N-1)N}\ ]^T$, and $a < b$.

3.4. Proposed Algorithm

The goal of the proposed algorithm is to determine the optimal rotation transform such that the total pairwise divergence measure is maximized to achieve signal separation. We use the gradient adaptation algorithm also known as the steepest descent [10] to update the rotation angles.

The overall update equation for stochastic gradient descent is:

$$\theta(n+1) = \theta(n) - \mu \frac{\partial J_2}{\partial \theta}, \quad (17)$$

where μ is the step size parameter. The gradient of the cost function J_2 with respect to the rotation angle θ_{ab} is derived as:

$$\frac{\partial J_2}{\partial \theta_{ab}} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\partial J_{ij}^2}{\partial \theta_{ab}}, \quad (18)$$

where

$$\begin{aligned} \frac{\partial J_{ij}^2}{\partial \theta_{ab}} = & \frac{\sum_{k=1}^{Q} \left(\frac{\partial \mathbf{R}_i}{\partial \theta_{ab}} \mathbf{X}(k) Y_j(k) + Y_i(k) \frac{\partial \mathbf{R}_j}{\partial \theta_{ab}} \mathbf{X}(k) \right)}{\sqrt{\left(\sum_{k=1}^{Q} Y_i^2(k)\right)\left(\sum_{k=1}^{Q} Y_j^2(k)\right)}} \\ & - \frac{\sum_{k=1}^{Q} Y_i(k) Y_j(k)}{\left(\sqrt{\left(\sum_{k=1}^{Q} Y_i^2(k)\right)\left(\sum_{k=1}^{Q} Y_j^2(k)\right)}\right)^3} \times \\ & \left[\left(\sum_{k=1}^{Q} Y_i(k) \frac{\partial \mathbf{R}_i}{\partial \theta_{ab}} \mathbf{X}(k) \right) \left(\sum_{k=1}^{Q} Y_j^2(k) \right) \right. \\ & \left. + \left(\sum_{k=1}^{Q} Y_i^2(k) \right) \left(\sum_{k=1}^{Q} Y_j(k) \frac{\partial \mathbf{R}_j}{\partial \theta_{ab}} \mathbf{X}(k) \right) \right], \end{aligned} \quad (19)$$

where $\mathbf{R}_i$ is the ith row of $\mathbf{R}(\theta)$, and $\mathbf{X}(k)$ is the kth column of $\mathbf{X}$.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In order to evaluate the effectiveness of the proposed method, we consider the following source separation examples. The sources are assumed to be approximately disjoint on the time–frequency plane, and issues regarding the accuracy of the extracted sources, convergence rate and robustness to noise are discussed.

Example 1: Separation of a chirp signal and two Gabor logon signals

In this example, we consider the separation of three source signals. A chirp signal is added to the mixture of two Gabor logons. The linear chirp signal has an initial normalized frequency of -0.2 and its instantaneous frequency increases to a normalized frequency of 0.2. The first Gabor logon is centered at the time sample point 50 and normalized frequency of 0.7, and the second Gabor logon is centered at the time sample point 150 and normalized frequency of -0.7. Three mixtures of these three source signals are given. Each combination is transformed to the time–frequency domain using a binomial kernel [5] with $K = 50$ time samples and $L = 64$ frequency samples. Each TFD is vectorized to form a TFD observation matrix of size 3×3200. It is known that the chirp signal overlaps with these two Gabor logons in the time domain, so it is not possible to separate them using time domain decomposition approaches. However, it is illustrated in Fig. 1 that these three signals can be effectively extracted using the proposed method on the time–frequency plane. Moreover, the convergence rate is high as shown in Fig. 2.

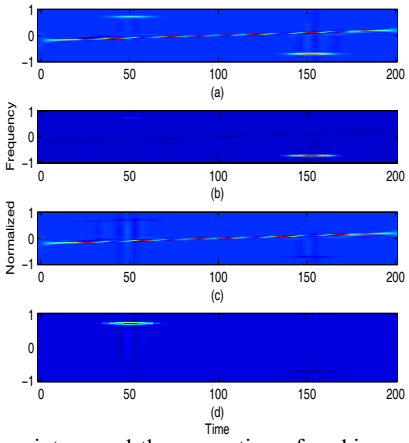

Fig. 1. The mixture and the separation of a chirp and two Gabor logons: (a) the mixture, (b) and (d) the two extracted Gabor logons, (c) the extracted chirp

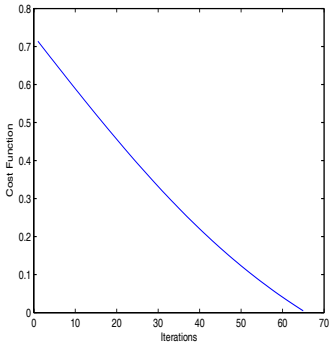

Fig. 2. The cost function versus the number of iterations

Example 2: Separation of two crossing chirp signals

In this example, we consider the separation of two signals overlapping in the time–frequency domain. A mixture of two linear chirp signals is used for source separation. One of the chirp signals has an initial normalized frequency of -0.8 and its instantaneous frequency increases to a normalized frequency of 0.8. The other one has an initial normalized frequency of 0.8 and its instantaneous frequency decreases to a normalized frequency of -0.8. Obviously, these two chirp signals overlap with each other in both the time and frequency domains. Typical time domain or frequency domain separation methods can not be used to perfectly recover them. Fig. 3

shows that using the proposed approach, we can successfully separate these two chirp signals from their mixtures.

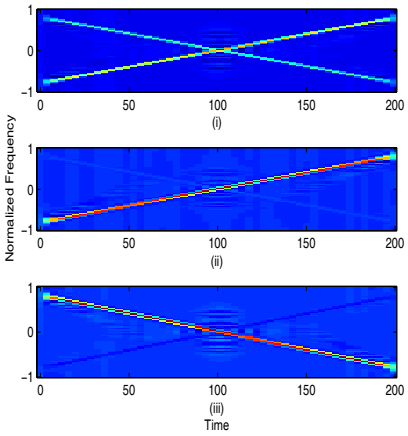

Fig. 3. The mixture and the separation of two crossing chirp signals: (i) the mixture, (ii) and (iii) the separated signals

Example 3: Performance comparison with FastICA

In order to evaluate the performance of the proposed approach, we compare it with FastICA on the time–frequency plane. The Mean Squared Error (MSE) is applied as the performance criterion, which is defined as:

$$\varepsilon_{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\| \hat{\mathbf{Y}}_i - \mathbf{Y}_i \|^2}{\| \mathbf{Y}_i \|^2}, \qquad (20)$$

where $\mathbf{Y}_i$ is the original source signal, $\hat{\mathbf{Y}}_i$ an estimate of this source signal, and N is the number of the source signals. We compare the MSE of the proposed algorithm with FastICA in the time–frequency domain for the signals discussed in Example 1 by adding white Gaussian noise over a SNR range of 2–18 dB. We use 100 Monte Carlo simulations for each noise level. It is evident from Fig. 4 that the proposed method has smaller MSE compared to FastICA. The difference in performance is due to the fact that the given sources are not necessarily independent and thus do not fit the assumptions underlying ICA.

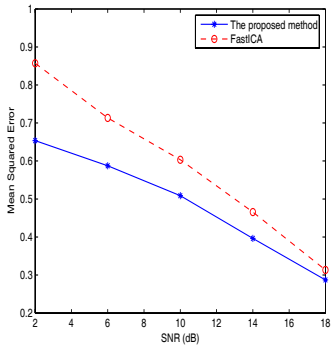

Fig. 4. Error performance of the proposed method and FastICA versus SNR

5. CONCLUSIONS

In this paper, a new approach is presented for the separation of non–stationary signals on the time–frequency plane using an information-theoretic cost function. The proposed algorithm performs an N–dimensional rotation to separate the source signals. Using Jensen–Rényi divergence as the cost function, a steepest descent algorithm is implemented to update the rotation angles. Several examples are given to illustrate the performance of the proposed algorithm. Issues regarding convergence rate and robustness under noise are investigated. The results illustrate that maximizing the divergence on the time–frequency plane can separate sources that are disjoint in the time–frequency domain, and is better than the mutual information cost function used in ICA in terms of fidelity to the original sources.

Future work includes investigation of the effect of order α in the Jensen–Rényi divergence on the performance of the source separation algorithm, and extending the algorithm to a more challenging case, i.e., the number of mixtures is smaller than the number of sources. Another area of future work is using signal synthesis methods to transform the extracted sources from the time–frequency domain to the time domain.

6. REFERENCES

[1] A. Belouchrani, K. Abed-Meraim anf J-F. Cardoso, and E. Moulines, "A blind source separation technique using second order statistics," *IEEE Trans. on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, 2001.

[3] K.E. Hild II, D. Erdogmus, and J. C. Principe, "Blind source separation using Rényi's mutual informaiton," *IEEE Signal Processing Lett.*, vol. 8, pp. 174–176, 2001.

[4] A. Belouchrani and M. G. Amin, "Blind source separation based on time–frequency signal representations," *IEEE Trans. on Signal Processing*, vol. 46, pp. 2888–2897, 1998.

[5] L. Cohen, *Time–Frequency Analysis*, Prentice Hall, New Jersey, 1995.

[6] R. G. Baraniuk, P. Flandrin, A. J. E. M. Janssen, and O. Michel, "Measuring time–frequency information content using the Rényi entropies," *IEEE Trans. on Info. Theory*, vol. 47, no. 4, pp. 1391–1409, May 2001.

[7] O. Michel, R. G. Baraniuk, and P. Flandrin, "Time–frequency based distance and divergence measure," in *Proc. IEEE Int. Symp. Time–Frequency and Time–Scale Analysis*, 1994, pp. 64–67.

[8] S. Aviyente, "Information processing on the time–frequency plane," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2004, vol. 2, pp. 617–620.

[9] F. D. Murnaghan, *The Unitary and Rotation Groups*, Spartan Books, Washington D.C., 1962.

[10] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, 1985.