Maximum Likelihood Performance over Higher-Order Statistics for Blind Source Separation in Wireless Systems

Syed A. Hassan and Bin Yang
Chair of System Theory and Signal Processing
University of Stuttgart, Germany
E-mail: alihassan@gatech.edu

***Abstract*- Blind source separation (BSS) has recently become an area of prime interest. Conventional adaptive source separation systems use a training sequence to estimate and separate sources with the help of predefined optimization criteria. In BSS, the key idea is to use the data statistics to get apriori knowledge and thus separate the sources blindly. Two important approaches to this regime are the maximum likelihood (ML) estimation and higher-order statistical (HOS) estimation. This paper presents the BSS problem in separating sources for a dual antenna communication system using the aforementioned algorithms. It has been shown that ML estimation outperforms HOS estimation for a wireless medium with noisy data transmission.**

I. INTRODUCTION

In signal processing, communications and controls, blind source separation is generally based on a wide class of unsupervised learning and filtering algorithms and it founds potential applications in many areas of engineering. Typically three unsupervised learning algorithms form the basis of BSS problem as formulated by Haykin in [1]. One of the most commonly used in signal processing area is the instantaneous BSS problem also known as independent component analysis (ICA). In ICA, a set of n unknown sources, s_i i Є 1,2,...,n, are linearly mixed in an unknown environment to produce an n-by-1 observation vector **x** such that

$$\mathbf{x} = \mathbf{A}\mathbf{s} \qquad (1)$$

where

$$\mathbf{s} = [\, s_1,\ s_2\,, \ldots, s_n]^T$$
$$\mathbf{x} = [\, x_1, x_2\,, \ldots\ , x_n]^T$$

and **A** is a nonsingular mixing matrix of dimension n-by-n. i.e. the number of observed signals is equal to that of number of sources. The solution to this problem is feasible, under certain conditions, except for an arbitrary scaling and permutation. In other words, given observed vector **x**, it is possible to find a demixing matrix **W** defined ideally as follows:

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} = \mathbf{D}\mathbf{P}\mathbf{s} \qquad (2)$$

where **y** is the output signal vector, **D** is nonsingular diagonal matrix and **P** is a permutation matrix. An illustration of the process can be given in Figure 1.

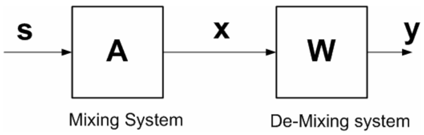

Fig. 1 Illustration of the basic BSS problem

For the given ICA model to be estimated, there are certain restrictions that should be taken into account. i.e., the source signal s_i are assumed statistically *independent* and they must have nongaussian distributions. At most one of the sources can be Gaussian. Throughout this paper, we also assume that the unknown mixing matrix is square.

A variety of successful algorithms have been developed for above and related source separation applications. An important application of BSS is in speech processing, where multiple observations from microphones located at different room locations have to be separated. This process is blind because we have no apriori knowledge about the measured observations. Thus using data statistics, people have devised algorithms, which separate the sources. Another important application is the separation of wireless signals in multiple antenna systems, where signals received from each antenna have first to be separated for each user. Many algorithms uses time domain analysis [2] while some uses the time-frequency domain approach to solve this problem [3]. A very popular approach for estimating the source components is maximum likelihood (ML) approach [4-7]. An interpretation of ML estimation is that we take those parameter values as estimates that give the highest probability for the observations. Another important algorithm is Kurtosis maximization which includes higher order statistics of the data [1,4]. Using kurtosis we devise practical algorithms by gradient methods. This paper basically involves the design of a digital communication system with multiple antennas. The goal is to show the performance evaluation and superiority of ML estimation over kurtosis maximization for a 2x2 multiple-input multiple-output (MIMO) system with QPSK modulated data transmitted over an AWGN channel. In basic BSS problem, there always arises a problem of permutation and scaling [4]. To solve these problems, a priori information about first two symbols of transmitted data has been taken.

The scheme of the paper is as follows. Section II presents an overview of algorithms for solving the BSS problem particularly for instantaneous mixtures. Section III deals, in details, with the simulated system model followed by solution of ambiguities. Simulations parameters and results have been discussed in Section IV and the paper concludes with certain recommendations in Section V.

II. Algorithms for ICA

Two most commonly used algorithms for ICA are the kurtosis maximization and maximum likelihood approach for estimating components. The following subsections present an overview of both of these techniques.

A. Kurtosis Maximization

Nongaussianity is of paramount importance in ICA estimation. According to central limit theorem, sum of two independent random variables results in another random variable that has distribution closer to the Gaussian distribution. Thus, to estimate one of the independent components, we can consider a linear combination of the x_i's. Let

$$y = \mathbf{b}^T\mathbf{x} = \mathbf{b}^T\mathbf{A}\mathbf{s} = \mathbf{q}^T\mathbf{s} \tag{3}$$

where vector **b** has to be determined. If **b** were one of the rows of the inverse of **A**, this linear combination $\mathbf{b}^T\mathbf{A}$ would actually equal one of the independent components. The question is now: How could we use the central limit theorem to determine **b** so that it would equal one of the rows of the inverse of **A**? In practice, we cannot determine such a **b** exactly, because we have no knowledge of matrix **A**, but we can find an estimator that gives a good approximation [4].

Let us vary the coefficients in **q**, and see how the distribution of y= $\mathbf{q}^T\mathbf{s}$ changes. The fundamental idea here is that since a sum of even two independent random variables is more Gaussian than the original variables, y is usually more Gaussian than any of the s_i and becomes least Gaussian when it in fact equals one of the s_i. In this case, obviously only one of the elements q_i of **q** is nonzero.

Typically nongaussianity is measured by the absolute value of kurtosis. Kurtosis can be estimated by fourth order moment of the sample data. The kurtosis of y, denoted by kurt(y), for a zero mean variable is defined by

$$kurt(y) = E\left\{y^4\right\} - 3\left(E\{y^2\}\right)^2 \tag{4}$$

Absolute or mean values of kurtosis are zero for a Gaussian variable, and greater than zero for most nongaussian random variables. To maximize the absolute value of kurtosis, we would start from some vector **w** which is the optimized vector for reconstruction filter, compute the direction in which the absolute value of the kurtosis of y = $\mathbf{w}^T\mathbf{z}$ is growing most strongly, based on the available sample **z**, where **z** is the whitened data obtained from **x**, and then move the vector **w** in that direction.

Thus denoting y = $\mathbf{w}^T\mathbf{z}$, the gradient of the absolute value of kurtosis of $\mathbf{w}^T\mathbf{z}$ can be simply computed as:

$$\frac{\partial\,|\,kurt(\mathbf{w}^T\mathbf{z})\,|}{\partial\mathbf{w}} = 4sign(kurt(\mathbf{w}^T\mathbf{z}))\left[E\{z(\mathbf{w}^T\mathbf{z})^3\} - 3\mathbf{w}\,\|\,\mathbf{w}\,\|^2\right] \tag{5}$$

For whitened w, $\|\mathbf{w}\|^2$=1, thus the numerical optimization problem reduces to the following update equation for kurtosis maximization

$$\partial\mathbf{w} \propto (kurt(\mathbf{w}^T\mathbf{z}))E\{z(\mathbf{w}^T\mathbf{z})^3\} \tag{6}$$

$$\mathbf{w} \leftarrow \mathbf{w}/\|\mathbf{w}\| \tag{7}$$

B. Maximum Likelihood Estimation

Another very popular approach for estimating the independent components is the maximum likelihood (ML) approach. For the ICA problem, ML estimation corresponds to minimizing the Kullback Leibler (KL) divergence between the distribution of **As** and the density of **s** by adjusting the matrix **A**. The KL divergence is a natural measure of the deviation for two pdfs and it gives how far away two densities are in terms of Euclidean distance. In order to obtain a good estimate **y**=**Wx** of the source signals **s**, we introduce an objective function or a contrast function l(**y**,**W**) in terms of the estimated **y** and **W.** Its expectation with respect to **y** is

$$L(\mathbf{W}) = E\{l\{\mathbf{y},\mathbf{W}\}\} \tag{8}$$

and it should be a function of **W** that represents the performance demixing by **W**. In other words L(**W**) should be minimized when the components of **y** are as independent as possible, that is, **W** is a rescaled permutation of $\mathbf{A}^{-1}$. We use Kullback-Leibler divergence for this purpose. Let $f_y(\mathbf{y};\mathbf{W})$ be the pdf of **y**=**Wx**=**Was** and let q(**y**) denotes another pdf of **y**, from which all y_i are statistically independent. In this case q(**y**) can be decomposed into a product form as:

$$q(\mathbf{y}) = \prod_{i=1}^{n} q_i(y_i) \tag{9}$$

The independent distribution is a reference function. We use KL divergence between the distribution $f_y(\mathbf{y};\mathbf{W})$ of y obtained by **W** and the reference distribution q(**y**),

$$\begin{aligned} D_{fq}(\mathbf{W}) &= D\left[f_y(\mathbf{y};\mathbf{W})\,\|\,q(\mathbf{y})\right] \\ &= \int f_y(\mathbf{y};\mathbf{W})\log\frac{f_y(\mathbf{y};\mathbf{W})}{q(\mathbf{y})}d\mathbf{y} \\ &= \int f_y(\mathbf{y};\mathbf{W})\log f_y(\mathbf{y};\mathbf{W})d\mathbf{y} \\ &\quad -\int f_y(\mathbf{y};\mathbf{W})\log(q(\mathbf{y}))d\mathbf{y} \end{aligned} \tag{10}$$

The KL divergence is a natural measure of the deviation for two pdfs. Hence $D_{fq}(\mathbf{W})$ shows how far the distribution $f_y(\mathbf{y};\mathbf{W})$ is from the reference distribution. Thus after applying stochastic gradient approach, the update rule for the demixing matrix can be formulated as

$$\Delta\mathbf{W} = -\eta \frac{D_{fq}}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} \tag{11}$$

where η is the learning parameter. In summary, the ML contrast function is the Kullback mismatch:

$$\phi_{ML} = D[\mathbf{x}][\mathbf{As}] \tag{12}$$

Considering output **y**, the ML principle is seen to correspond to minimize the KL mismatch

$$\phi_{ML}[\mathbf{y}] = D[\mathbf{y}][\mathbf{s}] \tag{13}$$

between the distribution of **y** and the distribution of a hypothetical source vector **s**.

III. Simulation Layout

After having gone through the techniques for signal separation, this section presents the layout in which simulations are being carried. It describes the block diagram of simulated system followed by the removal of ambiguities and finally design of algorithms for source separation problem.

A. System Model

For demonstrating the BSS algorithm for instantaneous mixtures applicable to a communication system, a 2x2 multiple input multiple output (MIMO) system has been simulated with different parameters. An i.i.d. bipolar source of data with uniform distribution generates binary bits. After serial to parallel conversion, a QPSK mapper maps the bit sequences into symbols according to QPSK constellation diagram. The symbols are then transmitted over the channel with a linear mixing matrix generated randomly. The linearly mixed symbols are then received by the receiving antenna and are then processed by the demixing matrix **W** to yield the outputs $\mathbf{y_1}$ and $\mathbf{y_2}$. The block diagram of the process is shown in Figure 2.

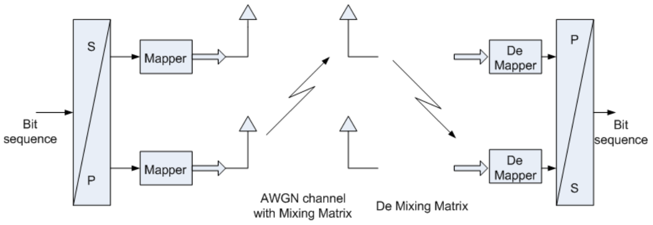

Fig. 2 Block diagram of the MIMO system with mixing channel

B. Permutation and scaling Ambiguities

It is well known that ICA/BSS algorithms can separate the source signals mixed linearly by a mixing matrix upto an unknown scaling and permutation. The same case was visible here. For that purpose, we took the first two symbols to get apriori knowledge about the *phase* and the *permutation* ambiguity. Here is how it works.

The problem arises because the mixing and demixing causes scaling and phase ambiguity in a sense that if e.g. a symbol with amplitude 1 and phase π/4 is transmitted, it could happen that after demixing it is received with a phase of π /4 + π /2, thus with an additional phase shift of π /2. The solution to this problem is simple: rotate the whole constellation of received symbols by an angle of π /2 radians. This is how the phase ambiguity in the received symbols can be avoided. The situation is shown in Figure 3.

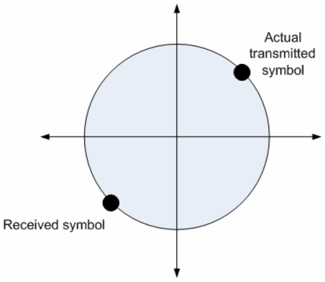

Fig. 3 Illustration of the scaling problem

For the permutation problem, since it is a 2 x 2 MIMO system, so this may be probable that the output y_2 is actually the reconstructed version of source s_1 and vice versa. To solve this problem of permutation, the knowledge from the first two symbols is sufficient. i.e., if we know that the phase difference between first two symbols of the first stream (first source) are π apart and the symbols of the other stream are π /2 apart, then using this apriori knowledge we can handle the permutation problem. The situation is clear in Figure 4.

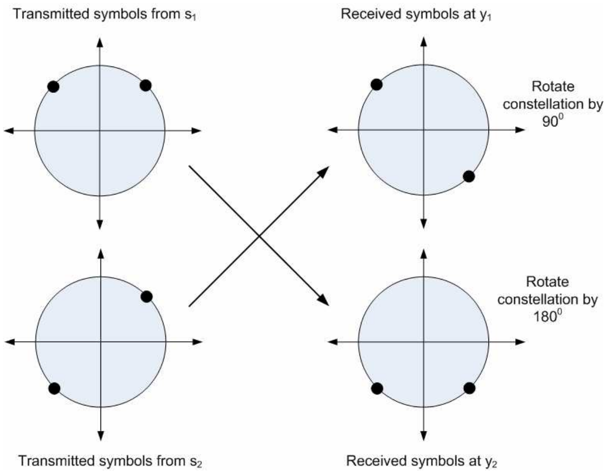
Transmitted symbols from s_1
Received symbols at y_1
Rotate constellation by 90^0
Rotate constellation by 180^0
Transmitted symbols from s_2
Received symbols at y_2

Fig. 4 Illustration of the permutation problem

C. Simulating the Algorithms

Simulating the ICA kurtosis maximization algorithm is straight forward. Taking gradients at each point, according to equations (6,7), gives the solution. The major problem arises in simulating ML receiver because it contains a term involving source densities, which is quite difficult to estimate. But in QPSK the information is encoded in terms of unit power and orthogonal phase rotation between every two bits. The phase can thus take any value, and since the amplitude remains constant, the baseband signal distribution is a circle on the complex plane. A smooth cumulative density function (cdf) **F** of such a circular distribution at unit circle is given as

$$F = \tanh\left(w(|y|-1)\right) \tag{14}$$

The differentiation of above cdf gives the probability density function (pdf) as

$$f(y) = w\left(1-\tanh^2\left(w(|y|-1)\right)\right) \tag{15}$$

where y =a + ib is a complex valued variable, and the parameter and w controls the steepness of the slope of the tanh function. When the steepness w approaches infinity, the densities approach the ideal density of a QPSK source, i.e., a unit circle at each symbol location j on QPSK constellation map. Of particular interest is the fact that in communications the signals are artificial, as the case here, thus the properties are known exactly as shown above. Now using the ML update rule from section II, we get

$$\begin{aligned}\Delta\mathbf{W} &\propto -\frac{d\mathbf{F}}{d\mathbf{W}}\mathbf{W}^T\mathbf{W} \\ &= \left(\left(\frac{f_j'\left(y_j,w_j\right)}{f_j\left(y_j,w_j\right)}\right)_j y^H + \mathbf{I}\right)\mathbf{W}\end{aligned} \tag{16}$$

where **y**=**Wx** are the sources separated from mixtures **x**, $f_j(y_j,w_j)$ is the pdf of source j parameterized by w and H is the hermitian operator. Now we can use the circular pdfs from equation (15) and inserting the result in equation (16), the final update rule is given by

$$\Delta\mathbf{W} = \left(\mathbf{I} - 2\left(\frac{w_j\tanh\left(w_j\left(|y_j|-1\right)\right)y_j}{|y_j|}\right)_j y^H\right)\mathbf{W} \tag{17}$$

As the pdf's are predetermined in this special case and they need not to be estimated, so they can be exploited in BSS with faster convergence and more accurate results.

IV. Simulation Results

Different scenarios have been simulated using the two main BSS algorithms for instantaneous mixtures, i.e. the kurtosis maximization and the maximum likelihood. The analysis shows the bit error rate (BER) with respect to signal to noise ratio (SNR) for stationary as well as non-stationary channel. For stationary channel, the mixing matrix remains the same throughout the simulation period for different SNRs. The cases are discussed here separately.

A. Comparison of ML-ICA with theoretical bound of QPSK

In this scenario, maximum likelihood method estimation for ICA is simulated as discussed in section III-C. For comparison, the theoretical bound of BER for QPSK is taken into account.

Now it can be seen clearly from Figure 5 that the mixing curve with ML-ICA deviates from the theoretical bound curve at low SNR, but the deviation becomes smaller and smaller as the SNR increases and after 5dB the two curves i.e. theoretical bound and the ML-ICA are just 1dB adjacent to each other. Thus it is clear that in presence of noise and instantaneous mixing/demixing effect, the ML-ICA algorithm works with a considerably high performance specially at high SNRs. This is due to the perfect knowledge of pdf of sources which helps in estimating the separation solution. More precise is the knowledge of pdf and its parameters, more close we are to the QPSK theoretical bound.

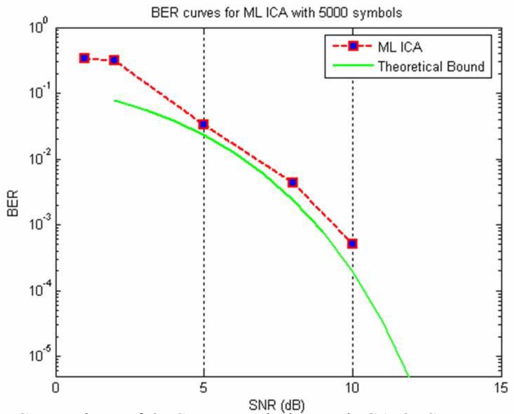

Fig. 5 Comparison of QPSK transmission and ICA QPSK transmission

B. Comparison of ML ICA for nonstationary channel

In this case, the maximum likelihood method is simulated with nonstationary mixing matrix. The mixing matrix remains the same throughout the simulation period with varying SNRs and the matrix is varied for each iteration. So five such curves are shown in Figure 6 with different mixing matrices. The number of QPSK symbols transmitted are same for each channel iteration.

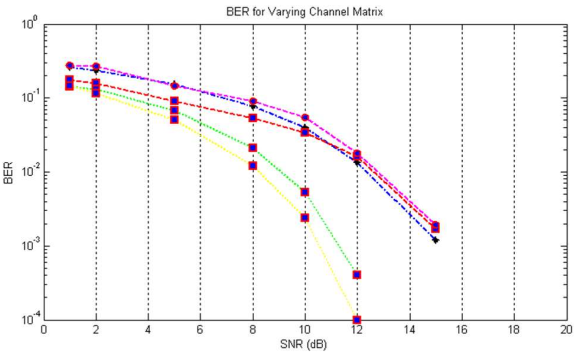

Fig. 6 Comparison of performance for different mixing matrices

It is observed that though there is not much difference or divergence between the curves for varying channels, however, there is also not a unanimous or unique curve each time. So this shows that the algorithm is sensitive to the channel mixing matrix. For a good conditioned matrix, with smaller condition number, the separation performance is better as compared to the one with an ill-conditioned matrix. Applying truncated singular value decomposition (SVD) to the channel demixing matrix and ignoring the smallest singular value can enhance the performance of overall system specially in the presence of noise.

C. Comparison of ML-ICA and HOS-ICA

The main crust of the paper is to determine the performance analysis of ML-ICA with HOS-ICA i.e. high order statistics ICA using kurtosis maximization algorithm. The channel matrix remains the same for both algorithms. The results over different SNRs for ML-ICA and Kurtosis Maximization are shown in Figure 7. We can see that the two curves are far apart and it clearly shows that ML algorithm works much better as compared to the Kurtosis maximization algorithm for noisy channel. There are two main observations for this scenario. One of the main reasons is the Gaussianity of noise sources. Since in Kurtosis maximization, we are actually moving away from Gaussianity, or in other words toward maximum non-Gaussianity, but here we are adding AWGN i.e. Gaussian noise sources to the transmitted data. Since from ICA restrictions we know that at most one source could be Gaussian [4], so this restriction seems to be violated here which affects the performance of HOS estimation. Another important reason is that HOS estimation requires a large amount of data to reliably compute the statistical parameters. As shown in Figure 8, different block of data have been transmitted to estimate components using kurtosis maximization algorithm and it can be seen that larger the amount of data transmitted, better is the performance because large amount of data estimates HOS more efficiently. The result is clear from Figure 8 that the algorithm performs well for larger block size.

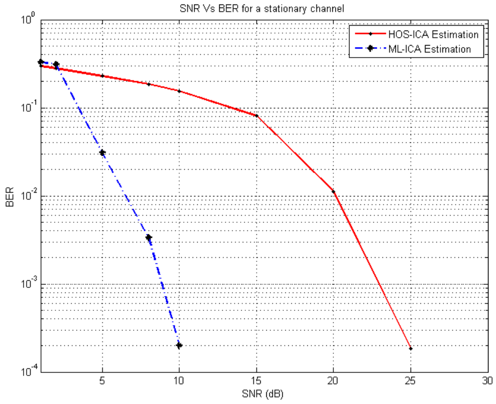

Fig. 7 Performance comparison of ML and Kurtosis maximization

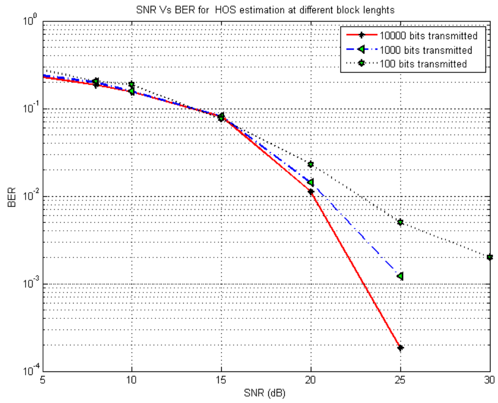

Fig. 8 Effect of block size on HOS estimation

V. Conclusions and Recommendations

After an insight into different simulation scenarios, it can been seen that ML algorithm for ICA works best if we know the pdf and other important parameters [6,7] of source signals. Kurtosis maximization algorithm works best for a noise free case and the performance is badly affected with the addition of noise. But there are certain open issues that need to be solved or looked upon for the general BSS problems in wireless communications. e.g. for QPSK transmission, the amplitude always remains constant and there is only an ambiguity of phase other than permutation. So important areas to be investigated is to use 16QAM or higher order constellations and observe the results with both magnitude and phase ambiguities. Another important issue is to find some quantitative or analytical expression for relating BER with condition number of the matrix, if the relationship exists. Moreover the simulated 2x2 MIMO system should be generalized to any n x n system and the performance be observed.

References

[1] S. Haykin, *Unsupervised Blind Source Separation Volume I: Blind Source Separation*, Prentice Hall, 1994.

[2] R. Aichner et al., “Time-domain blind source separation of non-stationary convolved signals ”, *in Proc. Int. Workshop on Neural Networks for Signal Processing*, Martigny, Switzerland, 2002.

[3] T. Nishikawa, H. Saruwatari, and K. Shikano, “Blind source separation based on multi-stage ICA combining frequency-domain ICA and timedomain ICA” in Proc. *IEEE ICASSP*, volume 1, pages 917-920, May 2002.

[4] A.Hyvarinen, Erkki Oja, *Independent Component Analysis*, John Willey and Sons. 2001

[5] Bell, A.J., and T.J.Sejnowski, “An information maximization approach to blind separation and blind deconvolution”, *in Neural Computation*, vol. 7, pp.1129-1159.

[6] S. Haykin, Blind Deconvolution, Prentice Hall, 1994.

[7] Torkkola, K., 1996, “Blind separation of convolved sources based on information maximization”, *in Porc. ICASSP*, Atlanta, GA, May 7-10, 1996.

[8] K. Torkkola, “Blind separation of radio signals in fading channels” *in Advances in Neural Information Processing System* 10, Denver, CO. Dec 1-6 1997.