ROBUST INDEPENDENT COMPONENT ANALYSIS

Sajjad H. Baloch†, Hamid Krim† and Marc G. Genton‡
†ECE Dept., North Carolina State University, Raleigh, NC.
‡Dept. of Statistics, Texas A&M University, College Station, TX.

ABSTRACT

Independent Component Analysis (ICA) attempts to separate independent components present in the mixture signals. Several criteria have been suggested for ICA in the past, including kurtosis and negentropy. Kurtosis suffers from a drawback of being outlier sensitive. As a remedy, we propose Robust ICA (RICA), which employs *appropriate* robust estimators. In this paper, we compare the robustness properties of RICA with kurtosis- and negentropy-based ICA. Since robust estimators are insensitive to outliers in contrast to maximum likelihood estimates (MLE), we demonstrate that in the presence of outliers, RICA works better than kurtosis- and negentropy-based ICA.

1. INTRODUCTION

A lot of work has been done on ICA since its introduction in 1980s by Hérault, *et. al.*[7]. Cardoso [2] used it for blind source separation (BSS) using higher order cumulant tensors and eventually proposed JADE algorithm [3]. In [1], Bell, *et. al.*, presented an information maximizing algorithm for separating independent components (IC). Later Hyvärinen and Oja proposed an algorithm [10] based on maximum kurtosis. These algorithms employ MLE estimators which are sensitive to outliers. In this paper, we propose using robust estimators combined with maximum kurtosis algorithm to make it robust to outliers.

This paper is organized as follows: We start with a brief introduction to ICA given next followed by robust estimators in Section 3. We present the proposed method in Section 4 and substantiate the technique with an application to BSS in Section 5. Finally, we provide concluding remarks in Section 6.

2. INDEPENDENT COMPONENT ANALYSIS

ICA finds ICs, $\mathbf{s}$, given n mixture signals, $\mathbf{x} = [x_1, \ldots, x_n]^T$:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

with no prior information about $\mathbf{s}$ except that they are statistically independent. The goal is to estimate the mixing matrix $\mathbf{A}$ and then by inverting it, we get:

$$\mathbf{s} = \mathbf{W}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}. \tag{2}$$

In this paper, we will employ kurtosis based ICA, which is due to Delfosse and Loubaton [4], that tries to maximize the non-Gaussianity of the component estimates. More specifically, it utilizes the central limit theorem according to which *a sum of any two independent random variables will be more Gaussian that the individual variables*. Since

$$y = \sum_i w_i x_i = \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{A}\mathbf{s} = \mathbf{z}^T\mathbf{s} \tag{3}$$

is implicitly a linear combination of the ICs $\{s_i\}$, it is more Gaussian than any of them and is least Gaussian when it equals one of s_i's. In other words, kurtosis-based ICA searches for a solution that maximizes non-Gaussianity of y. Since kurtosis of a Gaussian random variable is unique, it is used as a measure of non-Gaussianity.

Kurtosis of a random variable Y is defined as:

$$\kappa[Y] := E\left[Y^4\right] - 3\left(E\left[Y^2\right]\right)^2. \tag{4}$$

Note that:

$$\kappa[Y] \begin{cases} > 0 & \text{if } Y \text{ is super-Gaussian} \\ = 0 & \text{if } Y \text{ is Gaussian} \\ < 0 & \text{if } Y \text{ is sub-Gaussian} \end{cases} \tag{5}$$

Hence, finding ICs is equivalent to the following optimization problem:

$$\max_{\mathbf{w}} |\kappa[y]| = \max_{\mathbf{w}} \left| E\left[(\mathbf{w}^T\mathbf{x})^4\right] - 3\left(E\left[(\mathbf{w}^T\mathbf{x})^2\right]\right)^2 \right|. \tag{6}$$

In order to maximize the objective function $\mathcal{J} = |\kappa[y]|$ given by Eq. 6, gradient ascent method may be used. In case, mixtures $\mathbf{x}$ are pre-whitened to $\mathbf{z}$, the gradient of $\mathcal{J}$ is:

$$\mathcal{G}_{\mathcal{J}}(\mathbf{w}) = 12\text{sign}\left(\kappa[\mathbf{w}^T\mathbf{z}]\right)\left(E[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3] - \mathbf{w} \parallel \mathbf{w} \parallel^2\right). \tag{7}$$

As a simplification, the last term in Eq. 7 may be neglected for whitened mixtures, if $\mathbf{w}$ is normalized to unit norm after each iteration. Hence, the update at each iteration is:

$$\begin{aligned} \triangle\mathbf{w} &\propto \text{sign}\left(\kappa[\mathbf{w}^T\mathbf{z}]\right) E[\mathbf{z}(\mathbf{w}^T\mathbf{z})^3], \\ \mathbf{w} &= \mathbf{w}/ \parallel \mathbf{w} \parallel . \end{aligned} \tag{8}$$

The algorithm given above finds only one of the weight vectors. For entire weight matrix, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]^T$, we have to repeat it for each IC. In order to avoid convergence to the same value, outputs $\mathbf{w}_1^T\mathbf{x}, \ldots, \mathbf{w}_n^T\mathbf{x}$ must be decorrelated after every iteration.

3. ROBUST ESTIMATORS

Given N samples, $x_1, \ldots, x_N$, MLE estimates of mean, variance and kurtosis are:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{9}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2, \tag{10}$$

$$\hat{\kappa} = \frac{N^2[(N+1)m_4 - 3(N-1)(\hat{\sigma}^2)^2]}{(N-1)(N-2)(N-3)}, \tag{11}$$

0-7803-9461-6/05/$20.00 ©2005 IEEE 61

where m_4 is the fourth order sample moment:

$$m_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^4 \tag{12}$$

These estimators, especially those for higher order statistics, are not robust to a small fraction of outlying data. We, therefore, suggest using robust estimators instead of sample estimates. We will discuss two types of estimators: order statistics based estimators (OSE) and maximum likelihood-like estimators (MLLE).

3.1. Robust Estimate of Mean

In contrast to sample mean, insensitivity of median to outliers suggests that any estimator based on median will yield robustness.

1) Maximum Likelihood-like Estimate: Recall the MLE estimator of the mean μ of N independent samples drawn according to a conditional distribution $p(x|\mu)$. Here, we will discuss only the case where μ appears in $p(x|\mu)$ in the form $x - \mu$. The log likelihood function is:

$$\log \mathcal{L}(\mu) = -\sum_{i=1}^{N}\log p(x_i - \mu) = -\sum_{i=1}^{N}\rho(x_i - \mu), \tag{13}$$

where ρ is called the *loss function*. Minimizing the log likelihood function yields:

$$\sum_{i=1}^{N}\psi(x_i - \mu) = 0, \tag{14}$$

where $\psi(x) = \frac{\partial\rho(x)}{\partial x}$ is called the *influence function*. If we choose $\rho(x) = x$ (Gaussian), we get sample mean given by Eq. 9, which is not robust to outliers. On the other hand, if we choose $\rho(x) = \text{signum}(x)$, we get median. Note that the argument, $x - \mu$, is the distance between parameter μ and observation x_i. For the latter case, the distance, $x - \mu$, is independent of the relative size of the observations x_i, and is, therefore, more robust than the former. In essence, instead of assuming a Gaussian distribution, influence function may be chosen a priori to ensure robustness and efficiency over a wide range of distributions

2) Trimmed Mean: It is intuitive that the effect of the outlying values can be minimized by removing a certain quantile of the sample data. Hence, trimmed mean is defined as an average of inner fraction of observations:

$$\hat{\mu}_\alpha := \frac{1}{N-2k}\sum_{i=k}^{N-k+1} x_i, \tag{15}$$

where $k = \alpha N$ rounded to an integer.

3.2. Robust Estimate of Variance

A robust MLLE estimate of variance is a solution to Eq. 14 with the following influence function [9]:

$$\psi(x) = \begin{cases} x & \text{if } |x| \le \alpha \\ \alpha\text{sgn}(x) & \text{otherwise.} \end{cases} \tag{16}$$

Since variance of X is the spread of X about the mean, for which median $\tilde{X}$ may be taken as a robust estimate, the deviation of X from $\tilde{X}$ yields a robust estimate of the standard deviation known as *median absolute deviation* (MAD). Hence:

$$MAD = d\left(\text{median}\{|x_i - \tilde{X}| : i = 1, \ldots, N\}\right), \tag{17}$$

where d is a correction factor for consistency; for Gaussian distribution $d = 1.482$.

Rousseeuw *et. al.*[14] proposed a highly robust estimator $Q_N(\mathbf{x})$ based on the same idea:

$$Q_N(\mathbf{x}) = d\{|x_i - x_j| : i < j, i, j = 1, \ldots, N\}_{(k)}, \tag{18}$$

where $k = \lfloor({}^N\text{C}_2) + 2)/4\rfloor + 1$ and $\lfloor . \rfloor$ denotes the integer part. d is a correction factor, which is 2.2191 for Gaussian distribution. The beauty of this estimator is that it does not depend on the location parameter.

3.3. Robust Estimate of Kurtosis

Without loss of generality assume that observations are zero-mean. We transform the observations according to the map:

$$f(x) : x \mapsto x^2, \tag{19}$$

and subsequently employ a robust variance estimator using transformed observations $\{f(x_i)\}$ to get a robust estimate of kurtosis.

Another robust estimate of kurtosis is [12]:

$$\hat{\kappa} = \frac{(E_7 - E_5) + (E_3 - E_1)}{E_6 - E_2} - 1.23, \tag{20}$$

where E_i is the i^{th} octile, for $i = 1, \ldots, 7$.

3.4. Robust Estimate of Covariance

A robust estimator of variance can be utilized for estimating covariance [9], [6]:

$$Cov(X, Y) = \frac{\alpha\beta}{4}[Var(X/\alpha + Y/\beta) - Var(X/\alpha - Y/\beta)], \tag{21}$$

where $\alpha = \sigma_X$ and $\beta = \sigma_Y$ is recommended [5].

Rousseeuw *et. al.*[13] presented an estimate of covariance, known as MCD estimate, based on selecting a subset $\{x_{i_1}, \ldots, x_{i_h}\}$ of size h, where $1 \le h \le n$, that minimizes the generalized variance, i.e., the determinant of the covariance matrix is computed from the subset. The location estimator is then defined as:

$$\hat{\mu}_n = \frac{1}{h}\sum_{j=1}^{h} x_{i_j}, \tag{22}$$

and the scatter estimator as:

$$\hat{\Sigma}_n = c_p\frac{1}{h}\sum_{j=1}^{h}(x_{i_j} - \hat{\mu}_n)(x_{i_j} - \hat{\mu}_n)^T, \tag{23}$$

where c_p is a consistency factor and $h \approx 0.75n$.

4. ROBUST ICA

We propose employing robust estimators with kurtosis based ICA. As a preprocessing, we whiten the data and hence, estimate covariance matrix using an "appropriate" estimator given in Section 3.4. Update at each iteration, Eq. 8, requires estimates of mean, variance, covariance and kurtosis, which are given in Section 3.

5. EXPERIMENTAL RESULTS

We have tested outlier sensitivity of RICA using several sets of mixtures of real as well as synthetic signals from WAVELAB. RICA has proved to be much robust to outliers as compared to traditional ICA using kurtosis and negentropy.

In this section, we present some simulation results for a BSS problem and will compare robustness properties of ICA and RICA. We take `sine`, `tweet` and `greasy` as our test signals (Fig. 1), with their respective histograms in Fig. 2. This choice is quite diverse on two accounts:

1) `Tweet` and `greasy` signals are frequency rich while `sine` signal is composed of a single frequency.
2) `Tweet` and `greasy` are supergaussian while `sine` is subgaussian ($\kappa_{\mathtt{t}} = 2.9, \kappa_{\mathtt{g}} = 5.1, \kappa_{\mathtt{s}} = -1.5$).

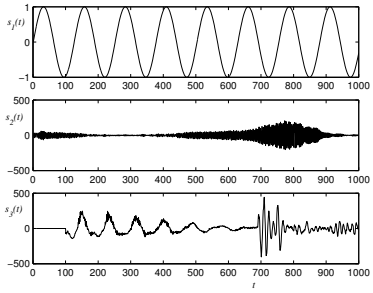
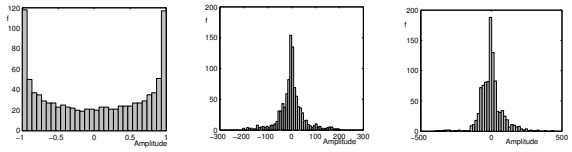


Fig. 1. *Independent Components: (a)* `Sine`*; (b)* `Tweet`*; (c)* `Greasy`.

Mixtures, shown in Fig. 3, were generated with a random mixing matrix. Their histograms, given in Fig. 4, indicate that they are more Gaussian than any of the ICs.



(a) (b) (c)

Fig. 2. *Histograms of independent components: Subgaussian for (a) Sine; Supergaussian for (b) Tweet; (c) Greasy.*

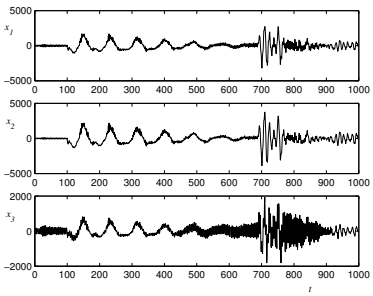
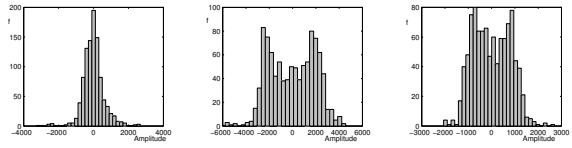


Fig. 3. *Mixture signals: (a) $x_1(t)$; (b) $x_2(t)$; (c) $x_3(t)$.*

To study the effect of outliers, ten random outliers were added to the mixture signals to generate signals shown in Fig. 5 with corresponding histograms in Fig. 6. Application of



(a) (b) (c)

Fig. 4. *Histograms for mixture signals: (a) $x_1(t)$; (b) $x_2(t)$; (c) $x_3(t)$.*

ordinary ICA with kurtosis yields results given in Fig. 8 after replacing samples outside 0.5 percentiles on either side with the neighboring means. Clearly, the algorithm fails to find any meaningful constituent signal.

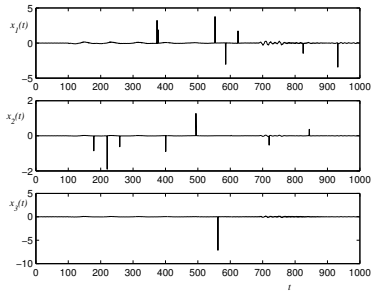
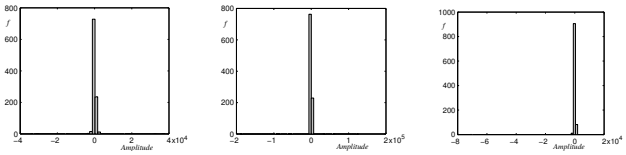


Fig. 5. *Mixtures with outliers: (a) $x_1(t)$; (b) $x_2(t)$; (c) $x_3(t)$.*



(a) (b) (c)

Fig. 6. *Histograms of mixtures with outliers: (a) $x_1(t)$; (b) $x_2(t)$; (c) $x_3(t)$.*

On the other hand, implementation of RICA involves three crucial steps:

1) whitening
2) evaluation of the sign of kurtosis of $\mathbf{w}^T\mathbf{z}$ in Eq. 6
3) covariance of $\mathbf{z}$ and $(\mathbf{w}^T\mathbf{z})^3$ in Eq. 6

which require robust estimation of covariance and kurtosis. We explored various robust estimators given in the previous sections. For kurtosis, we utilized robust estimate given in Eq. 20. Combination of Eq. 19 and Eq. 18 may not be used because it gives the magnitude of kurtosis, while we are interested in its sign only. For whitening, using Eq. 21 with Eq. 18 may result in negative definite covariance matrix, which in turn results in a whitening matrix that maps the mixture signals to the complex plane. We, therefore, recommend MCD estimator for covariance estimation. Learning process has not been found to be affected by the choice of robust estimators for mean and variance.

The results obtained from RICA are illustrated in Fig. 9, after setting the outlying values to the neighboring means. Clearly, the results are far better than those for ordinary ICA. Only the `sine` signal has some distortion, the amount of which is within acceptable limits.

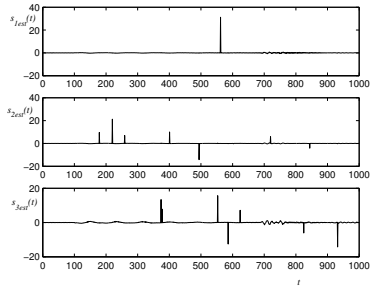


Fig. 7. *ICs estimated using kurtosis-based ICA: (a) $\hat{s}_1(t)$; (b) $\hat{s}_2(t)$; (c) $\hat{s}_3(t)$.*

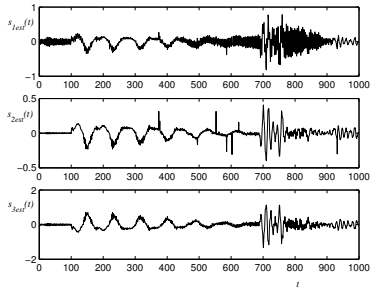


Fig. 8. *ICs estimated using kurtosis-based ICA and finally removing outliers: (a) $\hat{s}_1(t)$; (b) $\hat{s}_2(t)$; (c) $\hat{s}_3(t)$.*

5.1. Comparison with Negentropy-based ICA

In this section, we compare the results with those obtained for negentropy based ICA (Fig. 10). Obviously, the learned signals are not good estimates of the independent components. It must be pointed out that these results were obtained with half the number of outliers as that in the previous case. RICA, therefore, performs far better than ICA through negentropy and has nice robustness properties.

6. CONCLUSIONS

In this paper, we used robust estimators for ICA instead of employing MLE estimators to propose RICA. We explored several robust estimators for various statistics. Any of the estimators given in Section 3.1 and 3.2 can be used for the estimation of mean and variance. For covariance, if OSE estimators are coupled with Eq. 21, covariance matrix may become

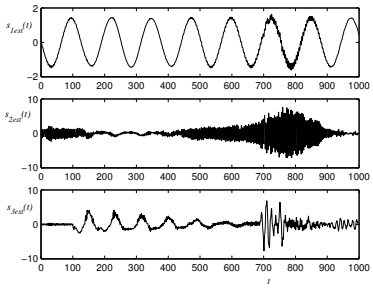


Fig. 9. *ICs estimated using RICA: (a)* `Sine`*; (b)* `Tweet`*; (c)* `greasy`*.*

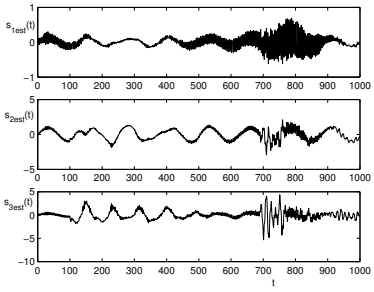


Fig. 10. *ICs estimated with negentropy-based ICA in the presence of 5 outliers: (a)* `Tweet`*; (b)* `Sine`*; (c)* `Greasy`*.*

negative definite in some cases. Such an estimator should, therefore, be avoided; we recommend using MCD estimator. To estimate the sign of kurtosis, we suggest using Eq. 20.

A comparison of RICA with kurtosis-based ICA and negentropy-based ICA reveals that RICA is more robust against outliers, while the rest fail to learn independent components in the presence of outliers. For instance, negentropy-based ICA fails even with half the number of outliers. In the absence of outliers, RICA is not inferior to the other two methods, although RICA is computationally slower. As a future work, we are working on evaluating breakdown point of RICA.

REFERENCES

[1] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution", *Neural Computation*, 7, pp. 1129-1159, 1995.

[2] J. F. Cardoso, "Blind identification of independent signals", *Proc. Workshop on Higher-Order Spectral Analysis*, 1989.

[3] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals", *IEE Proceedings-F*, 140(6), pp. 362-370, 1993.

[4] N. Delfosse, P. Loubaton, "Adaptive separation of independent sources: A deflation approach", *Signal Processing*, 45, pp. 59-83, 1995.

[5] R. Gnanadesikan and J. R. Kettenring, "Robust estimates, residuals, and outlier detection with multiresponse data", *Biometrics*, 28, pp. 81-124, 1972.

[6] R. Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd ed., Wiley, New York, 1997.

[7] J. Hérault and B. Ans, "Circuits neuronaux à synapses modifiables: décodage de messages composites par apprentissage non supervisé", *C.-R. de l' Académie des Sciences*, 299 (III-13), pp. 525-528, 1984.

[8] J. Hérault, C. Jutten and B. Ans, "Détection de grandeurs primitives dans un message composite par une architechture de neuromimétique en apprentissage non supervisé", *Actes du Xème colloque GRETSI*, pp. 1017-1022, 1985.

[9] P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.

[10] A. Hyvärinen and E. Oja, "A fast fixed point algorithm for independent component analysis", *Neural Computation*, 9(7), pp. 1483-1492, 1997.

[11] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, Wiley, 2001.

[12] J. J. A. Moors, "A quantile alternative for kurtosis", *The Statistician*, 37, pp. 25-32, 1988.

[13] P. J. Rousseeuw, "Least median of squares regression", *J. Amer. Statist. Assoc.*, 79, pp. 871-880, 1984.

[14] P. J. Rousseeuw and C. Croux, "Explicit scale estimators with high breakdown point", L_1 *Statistical Analyses and Related Methods*, Dodge, Y. (Ed.), pp. 77-92, Elsevier, Amsterdam, 1992.