A SIMPLE AND ROBUST FASTICA ALGORITHM USING THE HUBER M-ESTIMATOR COST FUNCTION

Jih-Cheng Chao

Semiconductor Group
Texas Instruments
Dallas, Texas 75243 USA

Scott C. Douglas

Department of Electrical Engineering
Southern Methodist University
Dallas, Texas 75275 USA

ABSTRACT

In blind source separation and independent component analysis, it is desirable to select a separation criterion that results in a simple algorithm and achieves accurate and robust source estimates. In this paper, we propose to use the Huber M-estimator cost function as the contrast function within the FastICA algorithm of Hyvärinen and Oja. The algorithm obtained from this cost is particularly simple to implement. We establish key properties regarding the local stability of the algorithm for general non-Gaussian source distributions, and its separating capabilities are shown through analysis to be largely insensitive to the cost function's threshold parameter. Simulations comparing the performance of this algorithm to standard FastICA implementations are given.

1. INTRODUCTION

In blind source separation (BSS), one possesses a set of measured signal vectors

$$\mathbf{x}(k) \quad = \quad \mathbf{A}\mathbf{s}(k) + \boldsymbol{\nu}(k) \tag{1}$$

where $\mathbf{A}$ is an unknown $(m \times m)$ mixing matrix, $\mathbf{s}(k) = [s_1(k) \ \cdots \ s_m(k)]^T$ is a vector-valued signal of sources, and $\boldsymbol{\nu}(k)$ contains jointly Gaussian noise. In most treatments of the BSS task, the $\{s_i(k)\}$ are assumed to be statistically-independent, and $\mathbf{A}$ is full rank. The goal is to obtain a separating transform $\mathbf{B}$ such that

$$\mathbf{y}(k) \quad = \quad \mathbf{B}\mathbf{x}(k) \tag{2}$$

contains estimates of the source signals. In independent component analysis (ICA), the linear model in (1) may not hold, yet the goal is to produce signal features in $\mathbf{y}(k)$ that are as independent as possible. Numerous ICA and BSS algorithms have been developed [1]–[4]. Among these methods, the FastICA procedure in [4] has fast convergence, global convergence for kurtosis-based contrasts in the noise-free case, and no step size parameter.

In blind source separation algorithms, the choice of cost function or algorithm nonlinearities determine each procedure's average separation capabilities. For example, maximum likelihood natural gradient-based approaches fail if a stability condition on the algorithm nonlinearities and the source distributions is not satisfied [3]. Moreover, no fixed cost function within such algorithms will separate arbitrary source mixture types [5]. It has been shown, however, that the family of three-level quantizer nonlinearities

$$f_i(y) \quad = \quad \begin{cases} 0 & |y| < \phi_i \\ \text{sgn}(y) & |y| \geq \phi_i. \end{cases} \tag{3}$$

are universal, such that there always exists a value of ϕ_i depending on the ith output signal distribution for which the ML-based gradient algorithm is locally-stable at a separating solution [5].

In this paper, we explore the design of single-unit contrast functions for the FastICA algorithm. While the FastICA algorithm is known to be globally-convergent for a kurtosis-based contrast, other contrast choices provide local convergence only if they satisfy a certain stability condition with respect to the output signal distributions. While several cost function choices appear to work for different source types, we are unaware of any theoretical results concerning the local stability of such cost functions for *arbitrary* source distributions. We propose the use of the Huber M-estimator cost function from robust statistics [6] as a FastICA algorithm contrast. A simple version of the FastICA algorithm involving only multiplies, adds, and threshold operations is obtained from this cost. Moreover, analysis and simulation show that the algorithm's separation behavior is largely independent of the cost function's threshold parameter for many source distributions, making it a robust choice for unknown source p.d.f.'s. Simulations comparing the performance of the FastICA algorithm with two common cost function choices show the efficacy of the approach.

2. THE HUBER M-ESTIMATOR COST FUNCTION

In his work on robust estimation and statistics [6], Huber specified the M-estimate $\widehat{x}_N$ for a the location parameter of a set of scalar measurements $x(k)$, $1 \leq k \leq N$, as a solution to the following problem:

$$\widehat{x}_N \quad = \quad \arg\min_{\widehat{x}} \sum_{k=1}^{N} G(x(k) - \widehat{x}) \tag{4}$$

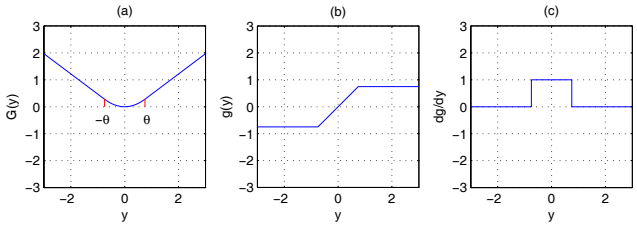

Fig. 1: The functions $G(y)$, $g(y)$, and $g'(y)$ for Huber's M-estimator cost function.

where $G(y)$ is a scalar cost function whose overall shape controls the robustness and accuracy of the estimate $\widehat{x}_N$ given imprecise knowledge about the characteristics of $x(k)$. Of particular interest is distributional robustness, in which the p.d.f. of $x(k)$ is unknown or deviates significantly from an assumed prior. Huber's work led to the specification of the so-called Huber M-estimator cost function

$$G(y) = \begin{cases} \dfrac{|y|^2}{2} & |y| < \theta \\ \theta|y| - \dfrac{\theta^2}{2} & |y| \geq \theta \end{cases} \quad (5)$$

where $\theta > 0$ is a threshold parameter designed to trade off the parameter estimation quality with the estimate's robustness to outliers and lack of prior distributional knowledge. The cost function in (5) combines the quadratic and absolute value functions in a convenient fashion. Minimizing (4) with the choice in (5) results in the implicit equation

$$\sum_{k=1}^{N} g(x(t) - \widehat{x}) = 0 \quad (6)$$

$$g(y) \equiv \frac{\partial G(y)}{\partial y} = \begin{cases} y & |y| < \theta \\ \theta \text{sgn}(y) & |y| \geq \theta \end{cases} \quad (7)$$

Fig. 1 shows the shapes of $G(y)$, $g(y)$, and for completeness, the derivative of $g(y)$ given by

$$\frac{dg(y)}{dy} = g'(y) = \begin{cases} 1 & |y| < \theta \\ 0 & |y| \geq \theta \end{cases} \quad (8)$$

3. USING THE HUBER M-ESTIMATOR COST WITHIN THE FASTICA ALGORITHM

The single-unit FastICA algorithm for extracting one non-Gaussian-distributed source from an m-dimensional linear mixture assumes that the source mixtures have been prewhitened by a linear transformation $\mathbf{P}$ where $\mathbf{v}(k) = \mathbf{P}\mathbf{x}(k)$ contains uncorrelated entries, such that the sample covariance of $\mathbf{v}(k)$ is the identity matrix. For the vector $\mathbf{w}_t = [w_{1t} \cdots w_{mt}]^T$, the FastICA update is

$$\widetilde{\mathbf{w}}_t = E\{g(y_t(k))\mathbf{v}(k)\} - E\{g'(y_t(k))\}\mathbf{w}_t \quad (9)$$

$$\mathbf{w}_{t+1} = \frac{\widetilde{\mathbf{w}}_t}{\sqrt{\widetilde{\mathbf{w}}_t^T \widetilde{\mathbf{w}}_t}}, \quad (10)$$

where $y_t(k) = \mathbf{w}_t^T \mathbf{v}(k)$ is the estimated source at time k and algorithm iteration t and the expectations in (9) are computed using N-sample averages.

In [4], the algorithm in (9)–(10) is formulated as the solution to the following optimization problem:

$$\text{maximize} \quad |E\{G(y_t(k))\} - E\{G(n)\}|^2 \quad (11)$$

$$\text{such that} \quad E\{y_t^2(k)\} = 1, \quad (12)$$

where n is a unit-variance Gaussian random variable. The criterion in (11) is described as the square of a simple estimate of the negentropy based on the scalar function $G(y)$, where $G(y)$ is "practically any non-quadratic function" [4]. The constraints on $G(y)$ depend on the statistics of the source extracted at the fixed point of the algorithm iteration, and the stability conditions are given in [4] as

$$[E\{s_i(k)g(s_i(k))\} - E\{s_i^2(k)\}E\{g'(s_i(k))\}] \times (E\{G(s_i(k))\} - E\{G(n)\}) > 0. \quad (13)$$

Several cost functions are suggested as possible choices for $G(y)$, including $G(y) = a^{-1}\log\cosh(ay)$ for $1 \leq a \leq 2$, $G(y) = -\exp(-y^2/2)$, and the kurtosis-based $G(y) = 0.25|y|^4$, although no verification of (13) for the first two choices of $G(y)$ and any well-known non-Gaussian distributions has been given.

Given the apparent freedom with which one can choose $G(y)$ within (11)–(12), we consider the Huber M-estimator cost in (5) as a possible choice for specifying the FastICA algorithm. We are also motivated by the similarity of (7) to (3) to allow this choice, particularly as $dg(y)/dy$ in (8) does not vanish nearly everywhere as does $df_i(y)/dy$. The resulting FastICA algorithm is particularly simple, as

$$E\left\{\frac{dg(y_t(k))}{dy}\right\} = \frac{1}{N}\sum_{k=1}^{N} H(\theta - |y_t(k)|), \quad (14)$$

where $H(u) = 1$ when $u \geq 0$ and is zero otherwise. It also is robust to data outliers, although this particular property is not the focus of this paper. Table 1 lists a short MATLAB script for implementing the multiple-unit version of this algorithm, in which the QR decomposition is used for signal deflation.

4. ON THE LOCAL STABILITY OF THE HUBER M-ESTIMATOR COST FOR FASTICA

Given the stability condition in (13), what can be said about the Huber M-estimator cost function when it is used in the FastICA algorithm? The following two theorems, proven in the Appendix, illustrate two properties about this cost.

Theorem 1: *Let $g(y)$ and $g'(y)$ have the forms in (7) and (8), respectively. Then, as long as $s_i(k)$ is non-Gaussian-distributed, there always exists a value of θ such that*

$$E\{s_i(k)g(s_i(k))\} - E\{s_i^2(k)\}E\{g'(s)\} \neq 0. \quad (15)$$

Table 1: FastICA algorithm with Huber M-estimator cost

```
---------------------------------------
[N,m] = size(x);
W = eye(m);
v = x/chol((x'*x)/N);
for i=1:iter
      y = v*W;
      t = (abs(y)<theta);
      g = y.*t + theta*sign(y).*(1-t);
      dg = sum(t);
      [W,R] = qr(v'*g - W*diag(dg));
end
---------------------------------------
```

Theorem 2: *Let $G(y)$ have the form in (5). Then, as long as $s_i(k)$ is non-Gaussian-distributed, there always exists a value of θ such that*

$$E\{G(s_i(k))\} - E\{G(n)\} \neq 0. \tag{16}$$

Taken together, these two theorems *do not* ensure (13). They suggest, however, that the design range for θ could be significant for many distributions. We substantiate this claim through the analysis below and by simulations in the next section. These results are also significant because, to our knowledge, few if any statements about the stability of a specific non-kurtosis-based cost function within FastICA have been given in the scientific literature. Moreover, it is unlikely that such results could be easily found given the complexity of the integrals for other $g(y)$ choices (*e.g.* $g(y) = \tanh(ay)$).

We have evaluated the range of θ values for which (13) is satisfied for four well-known zero-mean, unit-variance, non-Gaussian distributions: binary-$\{\pm 1\}$, uniform-$[-\sqrt{3}, \sqrt{3}]$, Laplacian, and four-level $\{-3/\sqrt{5}, -1/\sqrt{5}, 1/\sqrt{5}, 3/\sqrt{5}\}$. For the first three distributions, the Huber M-estimator cost produces an algorithm that is locally-stable for θ in the range $[0, s_{max})$, where s_{max} is the maximum possible value of $s_i(k)$ admitted by the source p.d.f. For the last distribution, the algorithm is predicted by (13) to be locally-stable for $\theta \in [0, 0.44721) \cup (0.5961, 1.342)$, which represents almost 90% of the range $[0, s_{max})$. In other words, any positive value of θ that places the nonlinear portion of $g(y)$ within the non-zero portion of the source p.d.f. often results in a locally-convergent algorithm. Again, this evaluation does not guarantee that FastICA with the Huber M-estimator cost will always work, but it suggests that one does not need to design specific values of θ to achieve separation.

In practice, one may not know what θ value to choose to obtain separation of a particular source mixture. A constant θ value creates the possibility of a poor match between the cost and any one source distribution. For this reason, we suggest that one *randomize* the value of θ over a range of positive values during coefficient adaptation. This randomization is not expected to significantly harm final separation performance given the linear shape of $g(y)$ over $|y| < \theta$; the main expected affect is a slight slowdown in convergence speed.

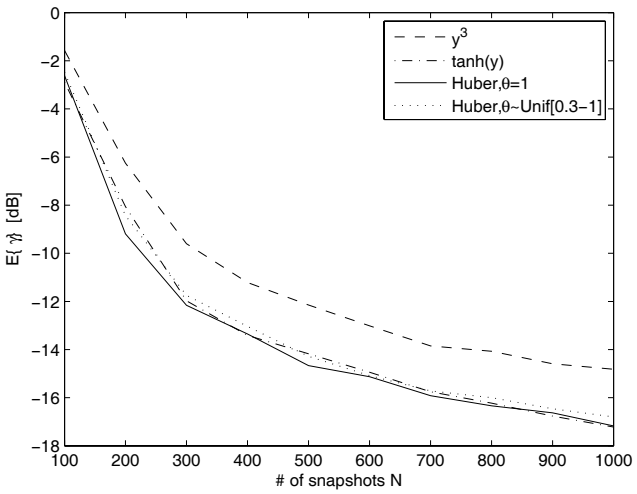

Fig. 2: $E\{\gamma\}$ vs. number of snapshots N for the various algorithms in the simulation example.

5. SIMULATIONS

We now explore the performance of the FastICA algorithm with Huber M-estimator cost via simulations. In these simulations, $m = 10$-source mixtures were generated consisting of three binary-$\{\pm 1\}$-, three uniform-$[-\sqrt{3}, \sqrt{3}]$-, two Laplacian-, and two four-level $\{-3/\sqrt{5},\ -1/\sqrt{5},\ 1/\sqrt{5},\ 3/\sqrt{5}\}$-distributed independent sources and a random mixing matrix. The multi-unit FastICA procedure was applied to this data for numbers of snapshots ranging from $N = 100$ to $N = 5000$ and for different θ values. The performance factor computed is the separation cost

$$\gamma = \frac{1}{2m}\left(\sum_{i=1}^{m}\sum_{l=1}^{m}\frac{|c_{il}|^2}{\max\limits_{1\le i\le m}|c_{il}|^2} + \frac{|c_{il}|^2}{\max\limits_{1\le l\le m}|c_{li}|^2}\right) - 1 \tag{17}$$

with $\mathbf{C} = \mathbf{WPA}$ as obtained at convergence of the algorithm. One hundred iterations were averaged to obtain each data point shown.

Fig. 2 compares the performance of FastICA with the Huber cost function and $\theta = 1$ and with the Huber cost function and a uniformly-randomized θ in the range $0.3 \le \theta \le 1$ at each iteration with two other versions of FastICA – one using $G(y) = 0.25y^4$ or $g(y) = y^3$, and another employing $G(y) = \log\cosh(y)$ or $g(y) = \tanh(y)$. As can be seen, the Huber cost function-based versions perform as well as or better than that based on the $\tanh(y)$ nonlinearity and they outperform the algorithm based on the kurtosis contrast. More significantly, our algorithm version with a randomized threshold parameter θ provides good separation performance across all sample sizes; performance deviations were less than ± 1dB from the algorithm with a fixed $\theta = 1$ value.

Fig. 3 illustrates the performance sensitivity of the FastICA algorithm with Huber M-estimator cost to the value of θ for these signal mixtures. As can be seen, the algorithm

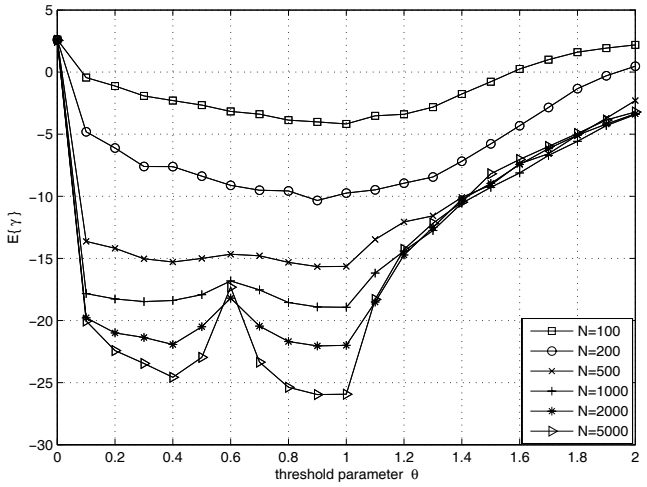

Fig. 3: $E\{\gamma\}$ vs. θ for the FastICA algorithm with Huber M-estimator cost in the simulation example.

performs well for values of θ satisfying $0.2 \leq \theta \leq 1$, and its performance degrades gracefully for higher θ values.

This Huber M-estimator cost was recently used to identify and separate uniformly-distributed sources from mixtures containing sparse impulsive sources for $m = 50$ and $N = 1175$ as well as for $m = 132$ and $N = 5000$ in an algorithm that won the Blind Source Separation portion of the 2005 Machine Learning for Signal Processing Workshop Data Analysis Competition [7]. The robustness of the Huber M-estimator cost was important to obtaining a winning solution.

6. CONCLUSIONS

In many blind source separation and independent component analysis algorithms, the cost function used to measure signal independence is a design parameter. In this paper, we have considered Huber's single-parameter M-estimator cost function for use within the well-known FastICA algorithm. The algorithm so obtained is computationally-simple, and the procedure works well for a wide range of threshold parameters θ. The reasons for the algorithm's robust behavior is indicated through statistical analysis.

7. APPENDIX

Proof of Theorem 1: Assume without loss of generality that $s_i(k) = s$ is unit-variance with an even-symmetric p.d.f., as the non-symmetric p.d.f. case can be handled through the symmetrizing transformation $p(s) \rightarrow 0.5(p(s) + p(-s))$. Considering the terms on the left-hand-side of (15) for the nonlinearities in (7) and (8), we have

$$E\{sg(s)\} = 2\left(\int_0^\theta s^2 p(s)ds + \theta\int_\theta^\infty sp(s)ds\right) \quad (18)$$

$$E\{g'(s)\} = 1 - 2\int_\theta^\infty p(s)ds. \quad (19)$$

Define $f_1(\theta) = E\{sg(s)\} - E\{g'(s)\}$. Then, we obtain

$$f_1(\theta) = 2\int_\theta^\infty (1 + \theta s - s^2)p(s)ds. \quad (20)$$

For Eq. (15) not to hold, $f_1(\theta) = 0$ for all possible values of θ. Suppose that $f_1(\theta) = c$ for some constant c. Then,

$$\frac{1}{2}\frac{\partial f_1(\theta)}{\partial\theta} = -(1 + \theta s - s^2)p(s)\big|_{s=\theta} + \int_0^\infty sp(s)ds \quad (21)$$

$$= 0 \quad (22)$$

which yields the relationship

$$p(\theta) = \int_\theta^\infty sp(s)ds. \quad (23)$$

As shown in [5], the only distribution $p(s)$ satisfying (23) is the unit-variance Gaussian distribution, for which it is straightforward to show that $f_1(\theta) = 0$. Thus, the theorem follows.

Proof of Theorem 2: Substituting (5) into the left-hand-side of (16) and simplifying yields the expression

$$E\{G(s) - G(n)\} = -\int_\theta^\infty (s-\theta)^2[p(s) - p_n(s)]ds \quad (24)$$

$$\equiv f_2(\theta) \quad (25)$$

for $s = s_i(k)$, where $p_n(s)$ is the unit-variance Gaussian distribution. For Eq. (16) not to hold, $f_2(\theta) = 0$ for all possible values of θ. Suppose that the slightly-more-general condition

$$f_2(\theta) = c_1\theta + c_2 \quad (26)$$

is true, where c_1 and c_2 are unknown constants. Then,

$$\frac{\partial f_2(\theta)}{\partial\theta} = 2\int_\theta^\infty (s-\theta)[p(s) - p_n(s)]ds = c_1 \quad (27)$$

$$\frac{\partial^2 f_2(\theta)}{\partial\theta^2} = -2\int_\theta^\infty [p(s) - p_n(s)]ds = 0. \quad (28)$$

For (26) to hold for all $\theta > 0$, we must have $p(s) = p_n(s)$, which results in $c_1 = 0$, $c_2 = 0$, and finally $f_2(\theta) = 0$. Thus, the theorem follows.

8. REFERENCES

[1] J.-F. Cardoso and A. Soloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proc. F.,* vol. 140, pp. 362-370, Dec. 1993.

[2] P. Comon, "Independent component analysis: A new concept?" *Signal Processing,* vol. 36, pp. 287-314, Apr. 1994.

[3] S. Amari, T. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks,* vol. 8, pp. 1345-1351, 1997.

[4] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks,* vol. 10, pp. 626-634, May 1999.

[5] H. Mathis and S.C. Douglas, "On the existence of universal nonlinearities for blind source separation," *IEEE Trans. Signal Processing,* vol. 50, pp. 1007-1016, May 2002.

[6] P. Huber, *Robust Statistics* (New York: Wiley, 1981).

[7] S.C. Douglas and J. Chao, "A blind source separation solution for the MLSP 2005 competition," invited talk, 2006 Machine Learning for Signal Processing Workshop, Mystic, CT, Sept. 2005.