# Speeding Up FastICA by Mixture Random Pruning

Sabrina Gaito and Giuliano Grossi

Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
Via Comelico 39, I-20135 Milano, Italy
grossi@dsi.unimi.it

**Abstract.** We study and derive a method to speed up kurtosis-based FastICA in presence of information redundancy, i.e., for large samples. It consists in randomly decimating the data set as more as possible while preserving the quality of the reconstructed signals. By performing an analysis of the kurtosis estimator, we find the maximum reduction rate which guarantees a narrow confidence interval of such estimator with high confidence level. Such a rate depends on a parameter $\beta$ easily computed a priori combining together the fourth and the eighth norms of the observations.

Extensive simulations have been done on different sets of real world signals. They show that actually the sample size reduction is very high, preserves the quality of the decomposition and impressively speeds up FastICA. On the other hand, the simulations also show that, decimating data more than the rate fixed by $\beta$, the decomposition ability of FastICA is compromised, thus validating the reliability of the parameter $\beta$. We are confident that our method will follow to better approach real time applications.

## 1 Introduction

Independent Component Analysis (ICA) ([1,2,3,4]) is a method to identify a set of unknown and generally non-Gaussian source signals whose mixtures are observed, under the only assumption that they are mutually independent. ICA has become more and more popular and, thanks to the few assumptions needed and its feasibility, it is applied in many areas such as blind source separation (BSS) which we are interested in [5].

More in general, ICA aim is to describe a very large set of data in terms of variables better capturing the essential structure of the problem. In many cases, due to the huge amount of data, it is crucial to make ICA analysis as fast as possible. From this point of view, one of the most popular algorithm is the well-known FastICA [6], which is based on the optimization of some nonlinear contrast functions [7] characterizing the non-Gaussianity of the components. Because of its widespread uses, in this paper we refer only to the kurtosis-based FastICA [6].

M.E. Davies et al. (Eds.): ICA 2007, LNCS 4666, pp. 185–192, 2007.

Our aim is to speed up FastICA by a suitable pruning of the linear mixtures that preserves the output quality. Essentially, the method proposed consists in randomly select a subset of data of size $d'$ less than the original size $d$ whose sample kurtosis is not too far from the right one. More in details, we perform an analysis of the kurtosis estimator on the sub-sample with the purpose to find the minimum reduction ratio $\rho = \frac{d'}{d}$ which guarantees a narrow confidence interval with high confidence level.

In particular, we identify a data-dependent parameter, called $\beta$, which combines both fourth and eighth norms of the observations, from which the reduction rate depends on.

The main step in our method is to compute $\beta$ on the mixed signals and obtain the actual reduction ratio $\rho = \frac{\beta}{\delta\epsilon^2}$, where $\epsilon$ and $\delta$ are the fixed confidence interval parameters of the sub-sample kurtosis. Then we randomly decimate the sample and we apply FastICA to the reduced dataset.

To assess the reliability of $\beta$ many simulations have been done on different sets of both real world and artificial signals. The experiments show that, accordingly to the $\beta$, a consistent ratio of reduction can be normally applied when the sample size is considerable, achieving a great benefit in terms of computation time. Furthermore, since $\beta$ (and consequently $\rho$) decreases also with respect to the number of signals $n$, the simulations show that the computation time is weakly affected by $n$. Moreover, the experiments give also prominence that when forcing the reduction ratio over the bounds derived by our analysis, the reconstruction error of FastICA grows noticeably.

Section 2 describes the pruning methodology. The effect of the data reduction will be analyzed in term of analysis of the kurtosis estimator in Section 3. In the same section the statistical meaning of the parameter $\beta$ is explained. In Section 4 we apply the method on a large set of real signals extracted from audio signals showing the performance of the proposed method.

## 2 Random Pruning

The model we assume for ICA is instantaneous and the mixture is linear and noiseless:

$$\mathbf{X} = \mathbf{AS},$$

where the $n \times d$ matrices $\mathbf{X}$ and $\mathbf{S}$ are respectively the observed mixtures and the mutually independent unknown signals, while $\mathbf{A}$ is a full rank $n \times n$ mixing matrix. Thus, $n$ is the number of mixed non-Gaussian signals and $d$ is their length. Therefore, for each $i \in [1\mathinner{..}n]$ the $i$-th row $\boldsymbol{x}_i$ of $\mathbf{X}$ represents a i.i.d. sample of size $d$ of the random variable $x_i$ representing the $i$-th mixture.

The goal of ICA is to estimate the demixing matrix $\hat{\mathbf{W}} \approx \mathbf{A}^{-1}$ in order to reconstruct the original sources signals

$$\hat{\mathbf{S}} = \hat{\mathbf{W}}\mathbf{X}.$$

Kurtosis-based FastICA is a very simple fixed-point algorithm with satisfactory performance, but it is time consuming for large scale real signals because its computational complexity is $\mathcal{O}(nd^3)$ [6].

In order to spare running time, before running FastICA we operate a random pruning on the mixtures procedure reducing the data by decimating the sample up to the minimum size allowed by $\beta$.

Denoting with $\|\boldsymbol{x}_i\|_p$ the usual $p$-norm, the overall procedure, with the preprocessing pruning preliminary phase, can be summarized in the following steps:

```
Pruning preprocessing
    1. $\beta(\boldsymbol{x}_i) = \frac{\|\boldsymbol{x}_i\|_8^8}{\|\boldsymbol{x}_i\|_4^8} \quad \forall i \in [1\,..\,n]$
    2. $\beta = \max_{\boldsymbol{x}_i} \beta(\boldsymbol{x}_i)$
    3. $d' = \frac{1}{\delta\varepsilon^2}(d\beta - 1) \approx \frac{d\beta}{\delta\varepsilon^2}$
    4. random draw $I_{d'} \subseteq [1\,..\,d]$ of size $d'$
    5. $\forall i \in [1\,..\,n]\ \forall j \in I_{d'}\ \ y_{ij} = x_{ij}$ so that $\boldsymbol{y}_i = (y_{ij_1}, \dots, y_{ij_{d'}})$
FastICA
    1. Perform FastICA on the matrix $\mathbf{Y}$ (whose $i$-th row is $\boldsymbol{y}_i$) instead
       of $\mathbf{X}$, obtaining $\hat{\mathbf{W}}$ by maximizing the sequence kurt $\left[\boldsymbol{w}_i^T \mathbf{Y}\right]$, where
       $\boldsymbol{w}_i^T$ is the $i$-th row of $\hat{\mathbf{W}}$
    2. Reconstruct the signals $\hat{\mathbf{S}} = \hat{\mathbf{W}}\mathbf{X}$.
```

Note that the decimation process throws away the same set of intermediate data points in all mixtures.

## 3 Theoretical Motivation

In this section we look for a lower bound for the reduction ratio $\rho$. The main step in FastICA where the sample size is relevant is when the kurtosis is being estimated on the data set.

Assuming, as usual, that each mixture $\boldsymbol{x}_i$ has zero mean and unitary variance, the kurtosis of each random variable $x_i$ reduces to its fourth moment $\mathsf{M}_4[x_i]$. Thus we analyze the effects coming from the use of a reduced data set in terms of confidence interval of the sample fourth moment.

The fourth moment estimate is generally performed on the whole sample $\boldsymbol{x}_i$ of size $d$ via the sample fourth moment $\hat{\mathsf{M}}_4^d[\boldsymbol{x}_i]$:

$$\hat{\mathsf{M}}_4^d[\boldsymbol{x}_i] = \frac{1}{d}\sum_{t=1}^{d} \boldsymbol{x}_{it}^4,$$

having the following mean and variance:

$$\mathsf{E}\left[\hat{\mathsf{M}}_4^d[\boldsymbol{x}_i]\right] = \mathsf{M}_4[x_i], \qquad \mathsf{var}\left[\hat{\mathsf{M}}_4^d[\boldsymbol{x}_i]\right] = \frac{1}{d}(\mathsf{M}_8[x_i] - (\mathsf{M}_4[x_i])^2).$$

Let us now estimate $\mathsf{M}_4[x_i]$ on the basis of the sub-sample $\boldsymbol{y}_i$.

Using the Chebyschev inequality we obtain the probability bounds:

$$\mathsf{Pr}\left\{\mathsf{M}_4[x_i](1-\varepsilon) \le \hat{\mathsf{M}}_4^{d'_i}[\boldsymbol{y}_i] \le \mathsf{M}_4[x_i](1+\varepsilon)\right\} \ge 1 - \frac{\mathsf{var}\left[\hat{\mathsf{M}}_4^{d'_i}\right]}{\varepsilon^2(\mathsf{M}_4[x_i])^2}$$
$$= 1 - \frac{\mathsf{M}_8[x_i] - (\mathsf{M}_4[x_i])^2}{d'\varepsilon^2(\mathsf{M}_4[x_i])^2}.$$

Setting the previous term equal to the confidence $1-\delta$, fixing the margin of error $\varepsilon$ and introducing the sample moments, we derive the minimum sample size $d'_i$ which respects the probability bound above:

$$d'_i = \frac{\hat{\mathsf{M}}_8^d[\boldsymbol{x}_i] - (\hat{\mathsf{M}}_4^d[\boldsymbol{x}_i])^2}{\delta\varepsilon^2(\hat{\mathsf{M}}_4[\boldsymbol{x}_i])^2}.$$

Expressing the sample moments in terms of norms:

$$\hat{\mathsf{M}}_4^{d'}[\boldsymbol{x}_i] = \frac{1}{d'}\|\boldsymbol{x}_i\|_4^4 \quad \text{and} \quad \hat{\mathsf{M}}_8^{d'}[\boldsymbol{x}_i] = \frac{1}{d'}\|\boldsymbol{x}_i\|_8^8,$$

we obtain:

$$d'_i = \frac{1}{\delta\varepsilon^2}\left(\frac{d\|\boldsymbol{x}_i\|_8^8}{\|\boldsymbol{x}_i\|_4^8} - 1\right).$$

It is evident that the minimum allowed sample size depends on the ratio of the two norms $\|\boldsymbol{x}_i\|_8^8$ and $\|\boldsymbol{x}_i\|_4^8$. Their statistical meaning is related to the variance of the estimator of the fourth moments estimated on the whole sample as:

$$\mathsf{var}\left[\hat{M}_4^d[\boldsymbol{x}_i]\right] = \frac{1}{d^2}(\|\boldsymbol{x}_i\|_8^8 - \frac{1}{d}\|\boldsymbol{x}_i\|_4^8)$$

Of course a low variance implies a good estimate and the possibility of highly reduce the sample size $d'_i$.

Since it holds that:

$$\frac{1}{d} \le \frac{\|\boldsymbol{x}_i\|_8^8}{\|\boldsymbol{x}_i\|_4^8} \le 1,$$

we note that the better ratio for the variance is $\|\boldsymbol{x}_i\|_8^8 = \frac{1}{d}\|\boldsymbol{x}_i\|_4^8$. On the other side, the variance of the estimator is highest when $\|\boldsymbol{x}_i\|_8^8 = \|\boldsymbol{x}_i\|_4^8$.

Introducing the parameter

$$\beta = \max_{\boldsymbol{x}_i} \frac{\|\boldsymbol{x}_i\|_8^8}{\|\boldsymbol{x}_i\|_4^8}$$

the minimum allowed sample size is:

$$d' = \frac{1}{\delta\varepsilon^2}(d\beta - 1) \approx \frac{d\beta}{\delta\varepsilon^2}$$

and the reduction ratio is:

$$\rho = \frac{d'}{d} = \frac{\beta}{\delta\varepsilon^2}.$$

## 4 Numerical Experiments

In this section we report the summary of extensive computer simulations obtained from the executions of FastICA on different set of sampled source signals: speech, musical and environmental sounds of various nature, mixed with randomly generated matrix. All the experiments have been carried out on Pentium P4 (2GHz, 1GB RAM) through software environment MATLAB 7.0.1.

The main purpose of the simulations is to apply the preprocessing pruning technique in order to appreciate the performance of FastICA both in terms of computation complexity and of quality of the reconstructed signals. Specifically, we are interested in validating the reliability of the parameter $\beta$ observing the performance decay. This attitude may find application in real time scenarios where high sampling rate can make prohibitive the use of the ICA technique.

All signals considered in the experiments are very big (order of magnitude $10^5$ and $10^6$) because for short sample size FastICA sometimes fails to converge or gets stuck at saddle points [8].

To measure the accuracy of the demixing matrix we use the performance index reported in [9], which represents a plausible measure of discrepancy between the product matrix $\mathbf{P} = (p_{ij})_{n\times n} = \mathbf{A}\hat{\mathbf{W}}$ and the identity matrix, defined as:

$$\mathrm{Err} = \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{|p_{ij}|}{\max_k |p_{ik}|} - 1\right) + \sum_{j=1}^{n}\left(\sum_{i=1}^{n}\frac{|p_{ij}|}{\max_k |p_{kj}|} - 1\right).$$

Due to the limit of space we present here only the most illustrative example, which examines signals of size $d = 10^6$. Table 1 shows the results on different groups of $n$ signals (with $2 \le n \le 10$).

**Table 1.** Average performance index and average computation time of FastICA on various groups of signals (from 2 to 10 with $d = 10^6$). Second column reports the reduction ratio $\rho < 1$, third and fourth columns report the performance index both with full and reduced sample size respectively. The last two columns report the computation times in both the cases. The numbers between brackets are the standard deviations calculated on the 30 trials.

| n | $\rho < 1$ | Err ($\rho = 1$) | Err ($\rho < 1$) | Time ($\rho = 1$) | Time ($\rho < 1$) |
|---|---|---|---|---|---|
| 2 | 0.03 (0.01) | 0.02 (0.05) | 0.03 (0.02) | 2.5 (0.9) | 0.1 (0.0) |
| 3 | 0.27 (0.01) | 0.04 (0.02) | 0.05 (0.02) | 4.5 (0.8) | 1.3 (0.6) |
| 4 | 0.25 (0.07) | 0.11 (0.11) | 0.11 (0.05) | 6.7 (0.8) | 1.7 (0.6) |
| 5 | 0.22 (0.07) | 0.18 (0.07) | 0.33 (0.63) | 9.4 (1.3) | 2.1 (0.7) |
| 6 | 0.19 (0.07) | 0.37 (0.15) | 0.46 (0.14) | 12.0 (1.7) | 2.4 (0.9) |
| 7 | 0.16 (0.06) | 0.62 (0.70) | 0.97 (0.97) | 14.7 (1.1) | 2.4 (1.0) |
| 8 | 0.16 (0.06) | 1.08 (0.75) | 1.44 (1.12) | 18.5 (2.1) | 2.9 (1.1) |
| 9 | 0.12 (0.04) | 1.23 (1.30) | 1.75 (2.70) | 26.5 (3.8) | 2.8 (0.9) |
| 10 | 0.11 (0.04) | 1.43 (0.29) | 1.91 (2.23) | 33.5 (3.4) | 2.8 (1.0) |

For each group we randomly generated 30 mixtures in order to observe, on average, both the time of convergence and the performance index of FastICA for the whole and the reduced samples respectively. All the experiments are obtained at confidence level 0.9 and margin of error 0.1.

Based on the simulations we can draw the following conclusions.

1. Sample size is highly reduced (up to one hundred times) while the quality of the decomposition is preserved, as highlighted by the performance index. Here, in particular, $\beta = \rho * 10^{-3}$ is sufficiently small, lying in the range between $10^{-5}$ and $10^{-4}$.
2. The discrepancy between the error given by the whole sample and that given by the pruned sample increases very slowly with $n$ (number of signals) as shown graphically in Fig. 1 (the lowest two errors corresponding to the third and fourth column of Table 1).
3. To assess the reliability of $\beta$, in the same figure we report the data obtained with a reduction ratio of one order of magnitude under that provided by analysis, i.e., with $\rho_{sub} = 10^{-1}\rho$ (highest error in the graphic). This experiment shows that the error grows noticeably.
4. As far as computation time is concerned, Fig. 2 (average times corresponding to the fifth and sixth column of Table 1) highlights the impressive gain of the computational cost. This gain depends on the fact that the computational cost is cubic with respect to sample size. Moreover, it can be noticed that in our pruning FastICA the computation time depends weakly on the number of signals because $\beta$ decreases with respect to $n$.

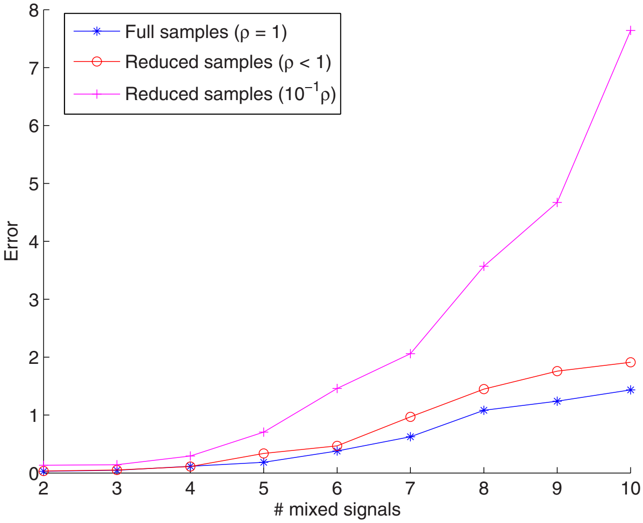

**Fig. 1.** Three average errors measured for various groups of signals ($d = 10^6$): the first is obtained with $\rho = 1$ (without reduction), the second decimated with $\rho = \beta * 10^3$ (where $\beta$ is computed in according to the previous analysis) and the third with $\rho_{sub} = \beta * 10^2$ (reducing $\beta$ of one order of magnitude)

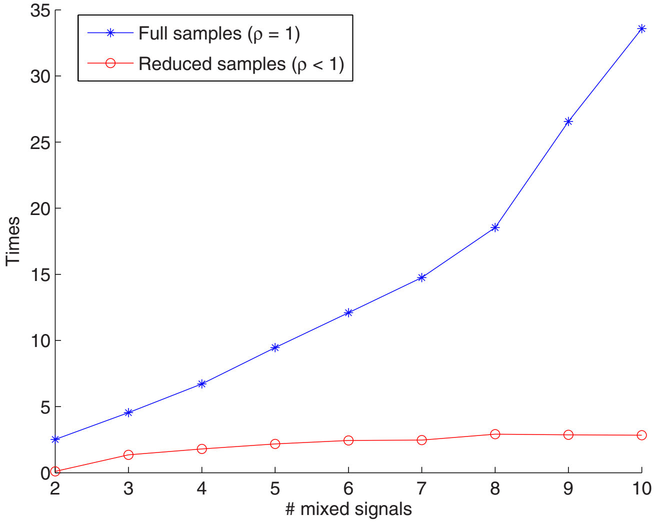

**Fig. 2.** Average times of FastICA on different groups of signals of full and reduced size: the first is obtained with $\rho = 1$ (without reduction), the second by decimation with $\rho = \beta * 10^3$

## 5 Conclusions

The contribution of this paper is the derivation of a signal-dependent parameter useful to randomly decimate high-dimensional mixtures in order to reduce the time in kurtosis-based FastICA executions. Such a parameter has been validated both in terms of rigorous high-order moments analysis and by means of computer simulations on real word signals. The results encourage to study the pruning technique deeper by exploring different sub-sampling methodologies and different contrast functions used in ICA. Finally, we are confident that our method can be used in real-time applications dealing with high sampling rate, where the online decimation permits to reasonably reduce the mixture size enabling FastICA to operate tightly.

## References

1. Comon, P.: Independent component analysis - a new concept? Signal Processing 36, 287–314 (1994)
2. Jutten, C., Herault, J.: Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. Signal Processing 24, 1–10 (1991)
3. Hyvrinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Chichester (2001)
4. Cichocki, A., Amari, S.: Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. John Wiley & Sons, Chichester (2002)

5. Cardoso, J.: Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem (1990)
6. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483–1492 (1997)
7. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10(3), 626–634 (1999)
8. Tichavsky, P., Koldovsky, Z., Oja, E.: Performance analysis of the fastica algorithm and cramr-rao bounds for linear independent component analysis. IEEE Transaction on Signal Processing 54(4), 1189–1202 (2006)
9. Amari, S., Cichocki, A.: Recurrent neural networks for blind separation of sources. In: Proceedings of International Symposium on Nonlinear Theory and Applications. vol. I, pp. 37–42 (1995)