Random Pruning of Blockwise Stationary Mixtures for Online BSS

Alessandro Adamo and Giuliano Grossi

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

Abstract. We explore information redundancy of linearly mixed sources in order to accomplish the demixing task (BSS) by ICA techniques in real-time. Assuming piecewise stationarity of the sources, the idea is to prune uniformly and independently most of sample data while preserving the ability of Kurtosis-based algorithms to reconstruct the original sources using pruned mixtures instead of original ones. The mainstay of this method is to control the sub-mixtures size so that the Kurtosis is sharply concentrated about that of the entire mixtures with exponentially small error probabilities. Referring to the FastICA algorithm, it is shown that the dimensionality reduction proposed while assuring high quality of the source estimate yields to a significant reduction of the demixing time. In particular, it is experimentally shown that, in case of online applications, the pruning of blockwise stationary data is not only essential for guarantying the time-constraints keeping, but it is also effective.

1 Introduction

The goal of independent component analysis (ICA) is to describe very large sets of data in terms of latent variables better capturing the essential structure of the problem. One of the main application is instantaneous blind source separation (BSS) which arises in many areas such as speech recognition, sensor signal processing, feature extraction and medical science.

In many cases, due to the huge amount of sample data or real-time constraints, it is crucial to make ICA analysis as fast as possible. However, among the several algorithms proposed for solving this problem, because of elevated computation time, most of them can only work off-line. In this regard, one of the most popular algorithm is FastICA [5], which is based on the optimization of some nonlinear contrast functions [4] characterizing the non-Gaussianity of the components. Even though this algorithm is one of the fastest to converge, it has cubic time-complexity [5].

The purpose of this paper is to look at the performance of the FastICA algorithm when a controlled random pruning of the input mixtures is done, both on the entire mixture available and when they are segmented into fixed-size blocks. Naturally, it is likely to speed up FastICA preserving, at the same time, the demixture quality. The method proposed consists in randomly select a sufficiently small subset of sample data in such a way that its sample kurtosis is not too far from those of the entire observations. In other words, we perform an analysis of the kurtosis estimator on the subsample with the purpose to find the maximum reduction which guarantees a narrow confidence interval with high confidence level.

The aim of this study is also in line with other works recently developed to effectively exploit blockwise ICA mainly to face the nonstationarity problem, as done in [7, 8]. Even if the purpose of these two works was to develop separation algorithm asymptotically efficient in the Cramér Rao lower bound sense, they consider nonstationary signals (time varying distribution) that have however constant variance in time-intervals (piecewise stationarity).

As shown in the last section, related to tests based on real signals, this attitude to prune the most of data sample provide good results even in case of blockwise analysis, especially because the dimensionality reduction of the FastICA input data is, in some case, up to two orders of magnitude.

2 Basic ICA Model

In the ICA model we are dealing with, the latent random vector $\mathbf{s} = [s_1, \ldots, s_n]^T$ of source signals is assumed to be statistically independent, zero-mean and at most one Gaussian. A random vector of their instantaneous linear mixtures $\mathbf{x} = [x_1, \ldots, x_n]^T$ is a standard data model used in ICA and BSS [2, 6] given by:

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

where A is an unknown, nonsingular $n \times n$ scalar matrix, as well as unknown are the source distributions p_{s_i}, with $i = 1, \ldots, n$.

Given D i.i.d. realizations of $\mathbf{x}$ arranged as columns of the data matrix $X_D = [\mathbf{x}(1), \ldots, \mathbf{x}(D)]$, the goal of ICA is to estimate the matrix A which makes the sources as independent as possible. Thus, denoting with $W = A^{-1}$ such a matrix, ICA algorithms yield a separating matrix $\hat{W} \approx A^{-1}$ providing estimates of the latent components

$$\hat{\mathbf{s}} = \hat{W}\,\mathbf{x}. \tag{2}$$

A more suitable model for blockwise stationary components must consider probability density functions $p_{s_i}^{(k)}$ different for each block k. As a consequence, given L-length data blocks $X_L^{(k)}$, the estimate of latent components take the form

$$\hat{\mathbf{s}}^{(k)} = \hat{W}^{(k)}\,\mathbf{x}^{(k)}, \qquad \text{for } k = 1, 2, \ldots\ . \tag{3}$$

3 Probability Bounds for Random Pruned Mixtures

The method proposed in this section consists on discarding or pruning the most of the D samples $\mathbf{x}(1), \ldots, \mathbf{x}(D)$ by means of random subsampling, but guarantying a sharp concentration of the subsample Kurtosis about that of the entire mixture.

Let us formalize the idea in statistical sense for a given observation $\mathbf{x}_i(t)$ taken at times $t = \tau, 2\tau, \ldots, D\tau$, where τ is the sampling period.

Let y be a random variable assuming values over the set of observed data $\mathbf{x}_i = \{x_i(1), \ldots, x_i(D)\}$ with uniform distribution. Assuming zero mean data $\mathbf{x}_i$, the moment of order p of y and its kurtosis are defined respectively as

$$\mu_p = \mathsf{E}\left[y^p\right] = \frac{1}{D}\sum_{t=1}^{D}(x_i(t))^p \quad \text{and} \quad \mathsf{K}\left[y\right] = \mu_4 - 3\mu_2^2.$$

Fixed a dimension $d \leq D$, let $Y = \{y_1, \ldots, y_d\}$ be a set of iid random variables distributed as y. Let us now consider the following non linear function (sample kurtosis) of the variables in Y:

$$\phi(\mathbf{y}) = \frac{1}{d}\sum_{i=1}^{d} y_i^4 - 3\left(\frac{\sum_{i=1}^{d} y_i^2}{d}\right)^2. \tag{4}$$

It can be stated the following concentration result:

Theorem 1. *Given the finite zero mean sample data $\Omega = \{\omega_1, \ldots, \omega_n\}$ such that $|\omega_i| \leq 1$ and the set $Y = \{y_1, \ldots, y_d\}$ of iid uniform random variables assuming values on Ω. If $\phi(\mathbf{y})$ is defined as in (4), then it holds:*

$$\Pr\left\{|\phi(\mathbf{y}) - \mathsf{K}\left[y\right]| \geq \rho \mathsf{K}\left[y\right]\right\} \leq 4\exp\left\{-\frac{(\rho \mathsf{K}\left[y\right])^2 d}{2(6\sigma^2+1)^2}\right\}$$

where σ^2 is the variance of y_i.

Proof. Let $m_p = \frac{1}{d}\sum_{i=1}^{d} y_i^p$ denote the sample moment of order p of the variable set Y, $\omega_m = \min \Omega$ and $\omega_M = \max \Omega$ such that $|\omega_M - \omega_m| \leq 2$ because of the bound $|\omega_i| \leq 1$. Since it holds that $\Pr\{\omega_m \leq y_i \leq \omega_M\} = 1$, by applying Hoeffding inequality [3] we obtain:

$$\Pr\left\{|m_p - \mu_p| \geq \varepsilon\right\} \leq 2\exp\left\{\frac{-2\varepsilon^2 d^2}{\sum_{i=1}^{d}(\omega_M - \omega_m)^2}\right\} \leq 2\exp\left\{\frac{-\varepsilon^2 d}{2}\right\}. \tag{5}$$

Let's now consider the functional

$$\kappa(m_4, m_2)(\mathbf{y}) = \phi(\mathbf{y}) = m_4 - 3m_2^2. \tag{6}$$

In order to establish how $\kappa(m_4, m_2)$ concentrates around the kurtosis of the entire sample space $\mathsf{K}\left[y\right]$, i.e., the probability

$$\Pr\left\{|\kappa(m_4, m_2) - \mathsf{K}\left[y\right]| \geq \rho \mathsf{K}\left[y\right]\right\},$$

we use a linear approximation around the point (μ_4, μ_2), obtaining

$$\begin{aligned} \kappa(m_4, m_2) &\approx \kappa(\mu_4, \mu_2) + \bigtriangledown\kappa|_{(\mu_4,\mu_2)} \cdot [(m_4, m_2) - (\mu_4, \mu_2)] \\ &= \mu_4 - 3\mu_2^2 + m_4 - \mu_4 - 6\mu_2(m_2 - \mu_2). \end{aligned}$$

Therefore, substituting the event $|m_4 - \mu_4| + 6\mu_2|(m_2 - \mu_2)|$ to the event $|m_4 - \mu_4 - 6\mu_2(m_2 - \mu_2)|$, we derive the inequality:

$$\Pr\{|\kappa(m_4, m_2) - \mathsf{K}[\mathbf{y}]| \geq \rho\mathsf{K}[y]\} \leq \Pr\{|m_4 - \mu_4| + 6\mu_2|(m_2 - \mu_2)| \geq \rho\mathsf{K}[y]\}.$$

To balance the error between the two events we introduce two parameters α and β such that $\alpha + \beta = 1$ and

$$\begin{aligned} &\Pr\{|m_4 - \mu_4| + 6\mu_2|(m_2 - \mu_2)| \geq \rho\mathsf{K}[y]\} = \\ &\qquad \Pr\{|m_4 - \mu_4| \geq \alpha\rho\mathsf{K}[y]\} + \Pr\{6\mu_2|(m_2 - \mu_2)| \geq \beta\rho\mathsf{K}[y]\}. \end{aligned}$$

By using (5), after some algebras it results:

$$\alpha = \frac{1}{6\mu_2 + 1} \qquad \text{and} \qquad \beta = \frac{6\mu_2}{6\mu_2 + 1}.$$

Finally, since sample data in Ω are zero mean, i.e., $\sigma^2 = \mu_2$, by inequality (5) it holds

$$\Pr\{|m_4 - \mu_4| + 6\mu_2|(m_2 - \mu_2)| \geq \rho\mathsf{K}[y]\} \leq 4\exp\left\{-\frac{(\rho\mathsf{K}[y])^2 d}{2(6\sigma^2 + 1)^2}\right\}.$$

□

The main reason of Theorem 1 is to explore uniform pruning operation on the given samples $\mathbf{x}(1), \ldots, \mathbf{x}(D)$ without weakening too much the algorithm separation ability. This can be done by fixing an error $\varepsilon = \rho\mathsf{K}[y]$ and a precision δ such that

$$\Pr\{|\phi(\mathbf{y}) - \mathsf{K}[y]| \geq \varepsilon\} \qquad \text{and} \qquad \delta = 4\exp\left\{-\frac{\varepsilon^2 d}{2(6\sigma^2 + 1)^2}\right\},$$

in order to derive the minimum dimension d of the subsample respecting the error and confidence requested, i.e,

$$d = \frac{2(6\sigma^2 + 1)^2}{\varepsilon^2} \log\frac{4}{\delta}. \qquad (7)$$

Recalling the Kurtosis property of linear combination of a set of independent variables, in previous section it has been shown that

$$\phi(\mathbf{y}) \approx \mathsf{K}[x_i] = \sum_{j=1}^{n} a_{ij}^4 \mathsf{K}[s_j],$$

where $(a_{ij})_{1\times n}$ is the i-th column of the mixture matrix A appearing in eq. (1) and $\mathsf{K}[s_j]$ the Kurtosis of the j-th source. Therefore, having chosen of normalizing the sample data, i.e., rescaling each mixtures in the unit interval, it results that $\phi(\mathbf{y})$ is sharply concentrated around the Kurtosis of all mixtures.

4 Simulation Results

We have carried out a vast bulk of simulation experiments using data obtained from a variety of source signals, mainly derived by audio an speech real signals. The sources that we used included essentially unimodal sub and supergaussian distributions having zero mean, as well as some examples of distributions that are nearly Gaussian (interpreted as noise). We varied the number of mixed components, from 2 to 30 and studied the average behavior of the algorithm Pruning + FastICA (P-FastICA) to vary the confidence parameters which control the dimension d of the pruned mixtures.

To measure the accuracy of the demixing matrix we use a metric called *performance index* [1] defined as:

$$\mathcal{E} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \frac{|\rho_{ij}|}{\max_k |\rho_{ik}|} - 1 \right) + \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \frac{|\rho_{ij}|}{\max_k |\rho_{kj}|} - 1 \right),$$

where $(\rho_{ij})_{n \times n} = A\hat{W}$. Such a matrix should then be close to the identity (up to a permutation), and the variances of its nondiagonal elements $\mathrm{Var}[\rho_{ij}]$, for $i \neq j$, reflect the mean value of residual interference between the separated signals $\hat{W}X$.

To experimentally establish the worth of the pruning preprocessing for FastICA (but it could be extended to different kurtosis-based algorithms for ICA) we conduct two kinds of experiments, devoted respectively to offline and online applications whose results are discussed in the next two subsections. All experiments are done on an Intel Core 2 processor of 1.6 GHz, equipped with 2 GB of RAM and running MATLAB implementation of the algorithms.

4.1 Offline Experiments

In offline tests we directly compare the performances of FastICA and P-FastICA in terms of the performance index and execution times using mixtures of 10^6 samples. The huge amount of sample data considered is mainly due to two reasons: on one hand there is experimental evidence that FastICA suffers of instability and lack of convergence against small data, on the other hand the method proposed has a sense only when high frequency rates are involved in the mixing process.

The experiments carried out with the two algorithms have shown that the performance indexes are comparable in all executions while the times are smaller up to two orders of magnitude (Fig. 1 reports the obtained results). The parameters chosen for confidence interval ε and δ assume values on $[0.05, 0.1]$ and $[0.1, 0.2]$ respectively.

4.2 Online Experiments

In order to apply the pruning technique in real-time applications, we experimentally study how the separation operation is performed by P-FastICA on blocks

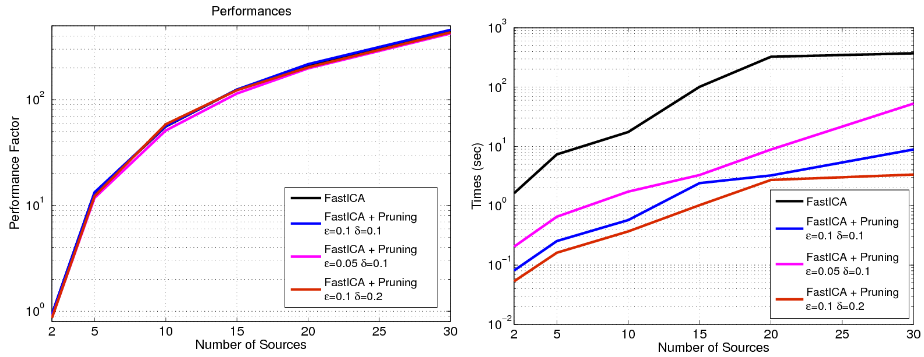

Fig. 1. Log-scale average performance index and computation times over 100 executions of FastICA and P-FastICA algorithms. The pruning parameters vary in the range $[0.05, 0.1]$ for ε and $[0.1, 0.2]$ for δ.

of data. The long (theoretically infinite) input mixtures must be segmented to fixed-size blocks of length imposed by the sampling ratio, here denoted by L. If there are n sensors, at every time unit we must arrange everything for processing n mixtures of L samples each one. There are two main operations to be performed: the first is the demixing task while the second is the block fitting to form an estimate of the original sources.

To spare execution time, the first task is preceded by pruning that allows to demix the blocks in a time unit fraction. The size of the subsample is indicated by parameter d given in (7), achieved by setting the confidence parameters explained in Section 2. The second task is critical because it is affected by the two well known indeterminacy problems of ICA: scaling factor and permutation ambiguity of the separated sources. For the type of signals we use, i.e., speech and audio signals, the scaling factor seems not to be a serious matter since the fitting of the output data blocks together do not let perceive trouble at all. As far as the permutation ambiguity is concerned, in Fig. 2 it is shown how the block-demixing and the reassembling process are tackled. All mixtures are synchronously break up every L samples and buffered on a block-matrix $X^{(k)}$ (depicted in light gray in the figure) of size $n \times L$, where index $k = 1, 2, \ldots$ denotes the block sequence. P-FastICA is then applied on a bigger block-matrix $Y^{(k)}$ (depicted in dark grey in the figure) of dimension $D > L$, obtained partially overlapping previous block of $D - L$ points. Overlapping is mainly motivated by the nonstationary of mixed sources across blocks showing that to correlate the samples of two adjacent blocks and initializing the optimization algorithm on the same point helps to locally rearrange in the right order the separated blocks. So, for every mixture-block k we find the estimate sources

$$\hat{S}^{(k)} = P^{(k)} \hat{W}^{(k)} X^{(k)},$$

where $\hat{W}^{(k)}$ is the demixing matrix carried out by P-FastICA with input matrix $Y^{(k)}$ and $P^{(k)}$ is a permutation matrix. The permutation allows to locally solve

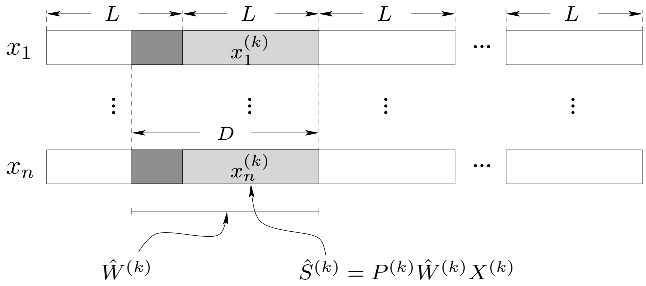

Fig. 2. Mixture segmentation in fixed-size blocks of length L and overlapping block of length $D > L$.

the reassembling operation between blocks already separated and to be joined. It is defined by element-wise thresholding the matrix product $(\rho_{ij}(k))_{n\times n} = A^{(k-1)}W^{(k)}$, where $A^{(k-1)}$ is the inverse of $\hat{W}^{(k-1)}$, with threshold 0.5:

$$p_{ij}^{(k)} = \begin{cases} 1, & \text{if } \rho_{ij}(k) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \qquad \text{for } k > 0,$$

where $W^{(0)} = I_n$.

Fig. 3 shows the average performance index and the computation times over 100 executions of FastICA and P-FastICA algorithms. The pruning parameters used here also vary in the range $[0.05, 0.1]$ for ε and $[0.1, 0.2]$ for δ. Observe that, even if our hw/sw architecture is not a real-time oriented apparatus, the performance index of the two algorithms are comparable while the computation time of the more pruned mixtures are under the unitary computation time.

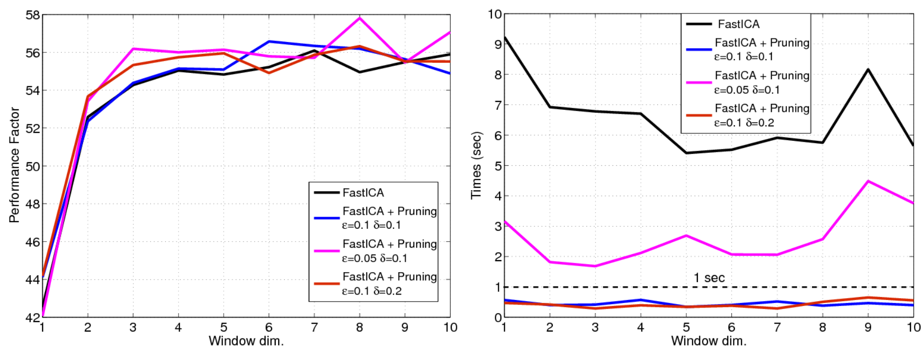

Fig. 3. Average performance index and computation times over 100 executions of block-wise FastICA and P-FastICA algorithms applied on 10 sources. The pruning parameters vary in the range $[0.05, 0.1]$ for ε and $[0.1, 0.2]$ for δ.

5 Conclusion

In this paper we explore the ability of a kurtosis-based algorithm like FastICA to carry out demixing task with a reduced number of observations with respect to those available. We exhibit a probability concentration result which allows to control a synchronized random pruning of the mixtures in order to hardly reduce the data dimension and consequently speed up the demixing process without too many losses in estimation quality.

In simulations, the algorithm is shown to be applicable in blind separation of a linear mixture of audio signals both offline and online.

References

1. 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)
2. Comon, P.: Independent component analysis - a new concept? Signal Processing 36, 287–314 (1994)
3. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58(301), 13–30 (1963)
4. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10(3), 626–634 (1999)
5. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483–1492 (1997)
6. Jutten, C., Herault, J.: Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. Signal Processing 24, 1–10 (1991)
7. Koldovský, Z., Málek, J., Tichavský, P., Deville, Y., Hosseini, S.: Blind separation of piecewise stationary non-gaussian sources. Signal Processing 89(12), 2570–2584 (2009)
8. Tichavsky, P., Yeredor, A., Koldovsky, Z.: A fast asymptotically efficient algorithm for blind separation of a linear mixture of block-wise stationary autoregressive processes. In: ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3133–3136. IEEE Computer Society (2009)