International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:3, 2008

Random Projections for Dimensionality Reduction in ICA

Sabrina Gaito, Andrea Greppi, and Giuliano Grossi

***Abstract*—** In this paper we present a technique to speed up ICA based on the idea of reducing the dimensionality of the data set preserving the quality of the results. In particular we refer to FastICA algorithm which uses the Kurtosis as statistical property to be maximized. By performing a particular Johnson-Lindenstrauss like projection of the data set, we find the minimum *dimensionality reduction rate* ρ, defined as the ratio between the size k of the reduced space and the original one d, which guarantees a narrow confidence interval of such estimator with high confidence level. The derived dimensionality reduction rate depends on a *system control parameter* β easily computed *a priori* on the basis of the observations only. Extensive simulations have been done on different sets of real world signals. They show that actually the dimensionality reduction is very high, it preserves the quality of the decomposition and impressively speeds up FastICA. On the other hand, a set of signals, on which the estimated reduction rate is greater than 1, exhibits bad decomposition results if reduced, thus validating the reliability of the parameter β. We are confident that our method will lead to a better approach to real time applications.

***Keywords*—** Independent Component Analysis, FastICA algorithm, Higher-order statistics, Johnson-Lindenstrauss lemma.

I. Introduction

Independent Component Analysis (ICA) ([1], [2]) is a method to identify a set of unknown and generally nongaussian source signals whose mixtures are observed, under the only assumption that they are 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) and feature extraction [3].

More in general, ICA field consists in describing a very large set of data, like those involved in applications such as the speech recognition or the imaging feature extraction, in terms of variables that better capture the essential structure of the problem. Due to the huge amount of data, it is crucial to make ICA analysis as fast as possible. Many attempts have been done to find more and more efficient algorithms [3].

Our aim is to speed up ICA from a different point of view. We show that the high-dimensional data set can be embedded into a lower dimensional space with a limited loss in the results quality. In particular, our dimensionality reduction preserves the Kurtosis of the original data which is the statistical property maximized by the FastICA algorithm [4], thus speeding up the overall computation.

We study how the Kurtosis is affected by our technique with a probability approach. To this end we consider the step of FastICA in which the Kurtosis is estimated looking for the minimum projection space size which guarantees a narrow probability bound to its estimator with a high confidence.

In particular, we identify a specific parameter of the system which is given by the ratio between the eighth norm and the fourth one to the square (the standard l_4 and l_8 norms defined in euclidean space) of the observations, from which the *dimensionality reduction rate* ρ, defined as the ratio between the reduced sample size k and the original one d, depends on. Therefore, the reduction rate of the mixed signals can be established a priori on the basis of the observations only. The statistical meaning of such a *system control parameter*, which we call β, is also investigated.

Section 2 reports a very short description of ICA formalism. The proposed Johnson-Lindenstrauss like projection is showed in Section 3 while the corresponding confidence intervals obtained both with Chebyschev and Hoeffding inequalities are derived in Section 4. In Section 5 we apply the method on a large set of real data extracted from audio signals showing the performance of the proposed method. Finally, Section 6 is devoted to the conclusions.

II. ICA Formulation

Let $\vec{s}(t) = [s_1(t), \ldots, s_n(t)]^T$ be a vector of source signals at time t that are mutually statistically independent, with zero-mean and at most one is gaussian. A vector of their linear mixtures $\vec{x}(t) = [x_1(t), \ldots, x_n(t)]^T$ is observed at time t with $t = 1, \ldots, d$.

The classical statistical approach consists of considering each signal $x_i(t)$ as a set of d realizations of the random variables x_i. Thus, for each i the set $\vec{x}_i = \{x_i(1), \ldots, x_i(d)\}$ represents a sample of size d of x_i.

Here we refer to the standard linear data model used in ICA and BSS ([1], [2]):

$$\vec{x}(t) = A\,\vec{s}(t),$$

where A is an unknown $n \times n$ scalar matrix. ICA decomposes $\vec{x}(t)$ by estimating the matrix A which makes the sources as independent as possible. That is, it finds the separating matrix W in such a way that

$$\hat{\vec{s}}(t) = W\,\vec{x}(t),$$

is an estimate of the independent components $\vec{s}(t)$.

ICA's basic idea is to exploit nongaussianity. According to the central limit theorem, the mixtures x_i of the independent sources s_i are closer to gaussian than the sources themselves. Thus ICA looks for the local maximum of nongaussianity of the signals x_i under the constraint that the variance is

Authors are with the Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, 20135 Milano, Italy (e-mail: {gaito,grossi}@dsi.unimi.it).

constant. Among the many algorithms proposed for ICA we refer to the class of algorithms which uses the Kurtosis as measure of nongaussianity. Furthermore, we are interested in the applications involving very large set of data well analyzed by the FastICA algorithm ([4]).

The only step in FastICA where the sample size is relevant is when the Kurtosis is being estimated on the data set. Next two sections show a suitable projection and its consequences on the estimate of the Kurtosis in confidence intervals terms.

III. Random Projections Preserving Independence and Kurtosis

In a seminal paper [5], Johnson and Lindenstrauss assert that any set of n points in d-dimensional metric space can be embedded into k-dimensional Euclidean space – where k is logarithmic in n and independent from d – so that all pairwise distances are maintained within an arbitrary small factor.

Lemma 3.1: **(JL-lemma)** Given $\varepsilon > 0$ and an integer n, let k be a positive integer such that $k = \mathcal{O}(\log n/\varepsilon^2)$. For every set P of n points in $\mathbf{R}^d$ there exists a random mapping $f : \mathbf{R}^d \to \mathbf{R}^k$ such that, for all $\vec{u}, \vec{v} \in P$,

$$(1-\varepsilon)\|\vec{u}-\vec{v}\|^2 \leq \|f(\vec{u})-f(\vec{v})\|^2 \leq (1+\varepsilon)\|\vec{u}-\vec{v}\|^2 \quad . \qquad (1)$$

Over the years, the probabilistic method has allowed for the original proof of JL-lemma to be greatly simplified and sharpened, while at the same time giving conceptually simple randomized algorithms for constructing the embeddings. The key idea is to use extremely simple probability distributions (e.g, gaussian) to perform random projections in the spirit of JL-lemma. The mainstay of all these results consists in the fact that squared length of a projected point $f(\vec{x})$ is sharply concentrated about the squared length of $\vec{x}$, since $\mathsf{E}\left[\|f(\vec{x})\|^2\right] = \|\vec{x}\|^2$. Effectively, being R a random matrix with independent entries, the squared inner product $(\vec{x} \cdot \vec{r}_j)^2$ of a point $\vec{x}$ with each column $\vec{r}_j$ of R $(1 \leq j \leq k)$ act as an estimator of $\|\vec{x}\|^2$, and $\|f(\vec{x})\|^2$ the sum of such estimators.

Unfortunately, it may be observed that doing such a projections even if the elements of $\vec{x}$ are mutually independent the same does not hold for the elements of $f(\vec{x})$, because each one is a linear combination of $\vec{x}$ themselves with a column of R (inner product).

Another limit these kinds of embeddings are afflicted is that the property showed for the l_2 norm does not apply to other norms as l_4 norm or bigger. In fact, it is easy to show that $\mathsf{E}\left[\|f(\vec{x})\|_4^4\right] \neq \|x\|_4^4$, making them ineffective for applications in which preserving higher order moments or cumulants is very important.

We propose a class of very easy *random projection* suitable in dimensionality reduction while preserving both the independence between the vector components and high order cumulants. For these projections we show weaker concentration results than those of Johnson and Lindenstrauss because they depend on suitable parameters computed on the instance at hand. The main advantage, in many cases is that computing such a system control parameter and successively ICA on reduced data is much less expensive than calculating ICA on the original data.

To insert the ICA model in the Johnson-Lindenstrauss framework we consider each signal $\vec{x}_i$ as a determined point in $\mathbf{R}^d$, thus obtaining n points in $\mathbf{R}^d$, represented by a $n \times d$ matrix X whose i^{th} is $\vec{x}_i$.

It is now natural to ask whether the high-dimensional point set $X \subseteq \mathbf{R}^{n\times d}$ could be embedded into a lower dimensional space $Y \subseteq \mathbf{R}^{n\times k}$ without suffering great distortion. As stated in the previous section, it is not possible to proceed with projections like those expressed in the JL-lemma because the independence is not preserved. Hence, we construct an embedding by picking a subset of coordinates of the original space by mean of Bernoulli trials.

Definition 3.1: Let $\vec{r}$ be a vector of d i.i.d. Bernoulli random variables of parameter ρ and $I_{\vec{r}} = \{t : r_i = 1\}$ the set of indexes of the non zero components of $\vec{r}$. An *independence preserving random map* $f : \mathbf{R}^d \to \mathbf{R}^k$ gives the vector

$$f(\vec{x}) = (x_i(t_1), \ldots, x_i(t_k)),$$

where its coordinates $t_j \in I_{\vec{r}}$ for all $j \in [1, \ldots, k]$ and $k = \sum_{i=1}^d r_i = |I_r|$, obtained by the product between each component of $\vec{r}$ with the correspondent component of $\vec{x}_i$

Observe that the dimension k of the space of projection has expected value $\mathsf{E}[k] = \rho d$, allowing to control this dimension by ρ.

In this setting, it is possible to show that all moments are preserved by this random map. In particular, this is true for a classical measure of nongaussianity used in ICA as the sample kurtosis, defined as

$$\begin{aligned}\mathsf{kurt}\left[f(\vec{x}_i)\right] &= \frac{1}{k}\sum_{t=1}^{k} f(x_i(t))^4 - 3\left(\frac{1}{k}\sum_{t=1}^{k} f(x_i(t))^2)\right)^2 \\ &= \frac{1}{k}\|f(\vec{x}_i)\|_4^4 - \frac{3}{k^2}\|f(\vec{x}_i)\|_2^4\end{aligned}$$

where $\|\vec{x}\|_p$ denotes the norm l_p of the vector $\vec{x}$.

Theorem 3.1: Let f be the random map defined above, then for each signal $\vec{x}_i$ it holds:

$$\mathsf{E}\left[\mathsf{kurt}\left[f(\vec{x}_i)\right]\right] = \mathsf{kurt}\left[\vec{x}_i\right].$$

Proof: Using the fact that every variable r_i is idempotent it is easy to show that

$$\mathsf{E}\left[r_i^p\right] = \mathsf{E}\left[r_i\right] = \rho \qquad \text{and} \qquad \mathsf{var}\left[r_i^p\right] = \mathsf{var}\left[r_i\right] = \rho(1-\rho).$$

We obtain that

$$\begin{aligned}\mathsf{E}\left[\|f(\vec{x}_i)\|_p^p\right] &= \mathsf{E}\left[\sum_{t=1}^{d}(x_i(t)r_t)^p\right] = \sum_{t=1}^{d} x_i^p(t)\mathsf{E}\left[r_t^p\right] \\ &= \rho\sum_{i=1}^{d} x_i(t)^p = \rho\|\vec{x}_i\|_p^p.\end{aligned}$$

Analogously, since the r_t are mutually independent, the variance is

$$\begin{aligned}\mathsf{var}\left[\|f(\vec{x}_i)\|_p^p\right] &= \mathsf{var}\left[\sum_{t=1}^{d}(x_i(t)r_t)^p\right] = \sum_{t=1}^{d} x_i(t)^{2p}\mathsf{var}\left[r_t^p\right] \\ &= \rho(1-\rho)\|\vec{x}_i\|_{2p}^{2p}\,.\end{aligned}$$

Hence, from the definition of kurtosis on the sample $f(\vec{x}_i)$ we have that:

$$\begin{aligned} \mathsf{E}\left[\mathsf{kurt}\left[f(\vec{x}_i)\right]\right] &= \mathsf{E}\left[\frac{1}{k}\|f(\vec{x}_i)\|_4^4 - \frac{3}{k^2}\|f(\vec{x}_i)\|_2^4\right] \\ &= \frac{1}{\rho d}\mathsf{E}\left[\|f(\vec{x}_i)\|_4^4\right] - \frac{3}{(\rho d)^2}\mathsf{E}\left[\|f(\vec{x}_i)\|_2^4\right] \\ &= \frac{1}{d}\|\vec{x}_i\|_4^4 - \frac{3}{d^2}\|\vec{x}_i\|_2^4 \\ &= \mathsf{kurt}\left[\vec{x}_i\right] \ . \end{aligned}$$

■

Thus the proposed random map preserves all the moments in mean, its variability will be study as probability bounds in next section.

IV. Probability Bounds

Since the kurtosis is a linear combination of the second and the fourth moment, we give a bound for both the moments by means of two classical probability inequalities, that is the Chebyschev[1] and the Hoeffding[2] ones. The first uses the variance and for our purpose gives more tight bound for a low number of points or signals ($n < 10$ circa). Vice versa the Hoeffding inequality is better for a greater number of signals because its negative exponential behavior.

In general, by applying Chebyschev inequality to the moment of order p we obtain the probability bounds:

$$\begin{aligned} \mathsf{Pr}\left\{\left|\|f(\vec{x}_i)\|_p^p - \rho\|\vec{x}_i\|_p^p\right| \geq \varepsilon\rho\|\vec{x}_i\|_p^p\right\} &\leq \frac{\mathsf{var}\left[\|f(\vec{x}_i)\|_p^p\right]}{\varepsilon^2\rho^2\|\vec{x}_i\|_p^{2p}} \\ &\leq \frac{1-\rho}{\rho\varepsilon^2}\frac{\|\vec{x}_i\|_{2p}^{2p}}{\|\vec{x}_i\|_p^{2p}} \ . \end{aligned}$$

By using instead the Hoeffding inequality to the moment of order p we obtain:

$$\mathsf{Pr}\left\{\left|\|f(\vec{x}_i)\|_p^p - \rho\|\vec{x}_i\|_p^p\right| \geq \varepsilon\rho\|\vec{x}_i\|_p^p\right\} \leq -2e^{-2\varepsilon^2\rho^2\frac{\|\vec{x}_i\|_{2p}^{2p}}{\|\vec{x}_i\|_p^{2p}}} .$$

Thanks to the usual centering and prewhitening procedures each signal $\vec{x}_i$ has zero mean and unitary variance. Thus the relevant moment is the fourth one for which we have the probability bounds.

$$\mathsf{Pr}\left\{\left|\|f(\vec{x}_i)\|_4^4 - \rho\|\vec{x}_i\|_4^4\right| \geq \varepsilon\rho\|\vec{x}_i\|_4^4\right\} \leq \frac{1}{n^2}$$

when we link the confidence level to the number of signals as $\frac{1}{n^2}$.

Finally we derive the dimensionality reduction rate ρ by Chebyschev:

$$\frac{1-\rho}{\rho\varepsilon^2}\beta_{x_i} \leq \frac{1}{n^2} \qquad \Rightarrow \qquad \rho \geq \frac{\beta n^2}{\varepsilon + \beta_{x_i} n^2},$$

and by Hoeffding

$$-2e^{-\frac{2\varepsilon^2\rho^2}{\beta}} \leq \frac{1}{n^2} \qquad \Rightarrow \qquad \rho \geq \frac{1}{\varepsilon}\sqrt{\beta_{x_i} \ln 2n}.$$

where $\beta_{x_i} = \frac{\|\vec{x}_i\|_8^8}{\|\vec{x}_i\|_4^8}$ is the system control parameter. Its statistical meaning is related to the variance of the estimator of the fourth moment computed on the whole sample as:

$$\mathsf{var}\left[\mathsf{kurt}\left[x_i\right]\right] = \frac{1}{d^2}(\|x_i\|_8^8 - \frac{1}{d}\|x_i\|_4^8).$$

Of course a low variance implies a good estimate and the possibility of highly reducing the data set. Since it holds that:

$$\frac{1}{d} \leq \frac{\|x_i\|_8^8}{\|x_i\|_4^8} \leq 1,$$

we note that the best ratio for the variance occurs when

$$\|x_i\|_8^8 \approx \frac{1}{d}\|x_i\|_4^8.$$

On the other side, the variance of the estimate is high when

$$\|x_i\|_8^8 \approx \|x_i\|_4^8.$$

Thus if β_{x_i} is small, say near $\frac{1}{d}$ it means that our original sample is very suitable to be well reduced. When it is high, say near 1, it is not.

V. Simulation Results

In this section we report the summary of extensive computer simulations obtained from the executions of FastICA on different sets of sampled source signals: speech, musical and environmental sounds of various nature, mixed with randomly generated matrix. The goal is to give experimental evidence and validity to the intuition that the information underlying sampling on large number is often redundant. All the experiments have been carried out through software environment MATLAB 7.0.1.

The purpose of the first experiment is to show the degree of reduction (up to 100 times) without decreasing the reconstructing ability of FastICA too much. To give a quantitative measure which reflects the real impact of the dimension reduction we use two kinds of performance index: a particular *Signal to Noise Ratio* (SNR or relative error) on the reconstructed signals and an index (absolute error), called *performance index*, which refers to the accuracy of the reconstructed mixing matrix for a given sample size [6].

The SNR for a given signal x_i and a reconstructed signal y_i is defined as

$$\mathrm{SNR}_{x_i} = \frac{\mathsf{E}\left[(s_i - y_i)^2\right]}{\mathsf{E}\left[s_i^2\right]},$$

while the overall SNR index, for the set of signals $\{x_1, \ldots, x_n\}$ is obtained by averaging SNR_{x_i}, that is,

$$\mathrm{SNR} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{SNR}_{x_i}.$$

Let $W \approx A^{-1}$ denote the matrix carried out by FastICA, then a plausible measure of distance is represented by the

[1] The Chebyschev inequality for a given random variable X indicates that $\mathsf{Pr}\left\{|X - \mathsf{E}\left[X\right]| \geq \lambda\right\} \leq \frac{\mathsf{var}[X]}{\lambda^2}$.

[2] The Hoeffding inequality for a given set of independent observation $X_1, \ldots, X_n$ such that $a_i \leq X_i \leq b_i$ and $S = \sum_i X_i$, indicates that $\mathsf{Pr}\left\{|S - \mathsf{E}\left[S\right]| \geq \lambda\right\} \leq 2e^{-\frac{2\lambda^2}{\sum_i (b_i - a_i)^2}}$.

performance index given by the discrepancy between the product $P = AW$ and the identity matrix, defined as:

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

Table 1 shows the tests on different groups of n high-dimensional signals (with $2 \leq n \leq 35]$), each of length $d = 10^6$. All the values are obtained at confidence level of 0.9 and accuracy 0.1.

TABLE I
PERFORMANCE INDEX OF FASTICA ON VARIOUS GROUPS OF SIGNALS (FROM 2 TO 35): SECOND COLUMN REPORTS THE VALUES OF β, THE THIRD COLUMN REPORTS THE VARIOUS REDUCTION RATE $\rho < 1$ (DEPENDING ON β), WHILE THE LAST TWO COLUMNS REPORTS THE PERFORMANCE INDEX BOTH WITH FULL AND REDUCED SAMPLE SIZE.

n	$\beta \times 10^3$	$\rho < 1$	Error ($\rho = 1$)	Error ($\rho < 1$)
2	0.037	0.007	0.01	0.01
3	0.328	0.098	0.06	0.12
4	0.183	0.073	0.10	0.10
5	0.296	0.148	0.22	0.46
10	0.100	0.105	1.36	1.63
15	0.047	0.046	9.47	21.39
20	0.039	0.038	9.33	17.70
25	0.042	0.041	13.98	26.02
30	0.031	0.030	36.22	60.38
35	0.064	0.064	50.73	89.13

In the table the errors referred by the performance index are reported. The second column shows the value of parameter β, the third column shows the value of ρ obtained as a function of β and the last two columns report the error, in terms of performance index we obtain with the whole sample ($\rho = 1$) and with the sample of reduced size ($\rho < 1$). We can observe that

1) when β is sufficiently small (ranging from 10^{-5} to 10^{-4}), it allows to reduce the sample size up to one hundred times, while preserving the confidence and the accuracy.
2) the error increases very slowly with n and however it guarantees no explosion of the distortion when measured in terms of signal to noise ratio. On the contrary, as shown in Figure 1, the SNR index seems not affected by this sensible reduction of the dimension.
3) On the other hand, the computation time of FastICA hardly depends on the size of the sample, as shown in Figure 2.

To test the reliability of β we use a set of signals whose beta values are very big, i.e., near to 1, which represents, as explained in the previous section, its better value in terms of variance. As a consequence the rate of reduction ρ is greater than one and then no reduction is admitted. We force however the reduction to $\rho = 0.01$ and calculate the performance error and report the data in Table 2. To show the goodness of β as system control parameter, it can be noted that, since the variance is very high, the error grows significantly both for the whole sample and for the reduced sample, giving a very poor performance of FastICA in each case.

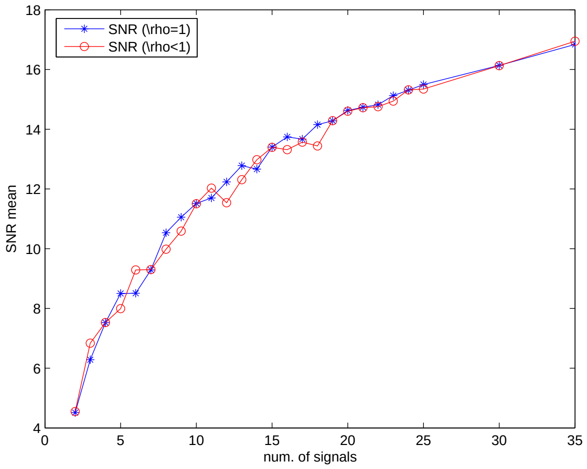

Fig. 1 The SNR for a group of signals of full and reduced size

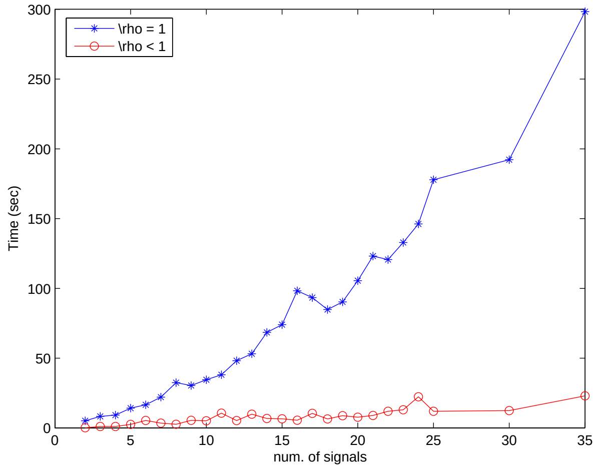

Fig. 2 Simulation times (in seconds) of FastICA

As a consequence, we have that the ratio SNR grows as shown in Figure 3.

VI. CONCLUSIONS

Concluding, we can assert that it is possible to execute the FastICA algorithm on a set of data reduced by projection in a lower dimensional space yet preserving the higher-order statistics. The rate of reduction ρ depends particularly on the value of the parameter β which is calculated on particular instance of signals and whose meaning has been exploited. Simulations confirm the reliability of the method and show a

TABLE II

PERFORMANCE INDEX OF FASTICA ON VARIOUS GROUPS OF SIGNALS (FROM 2 TO 10) WITH HIGH VALUES OF THE β PARAMETER: SECOND COLUMN REPORTS THE VALUES OF β, WHILE THE LAST TWO COLUMNS REPORTS THE PERFORMANCE INDEX BOTH WITH FULL AND REDUCED SAMPLE SIZE.

n	β	Error ($\rho = 1$)	Error ($\rho = 0.01$)
2	0.009	2.22	2.20
3	0.419	5.06	6.38
4	0.446	10.79	10.29
5	0.446	17.99	17.48
6	0.373	24.31	26.32
7	0.029	34.26	35.03
8	0.045	44.78	48.12
9	0.022	56.27	60.13
10	0.030	71.22	72.63

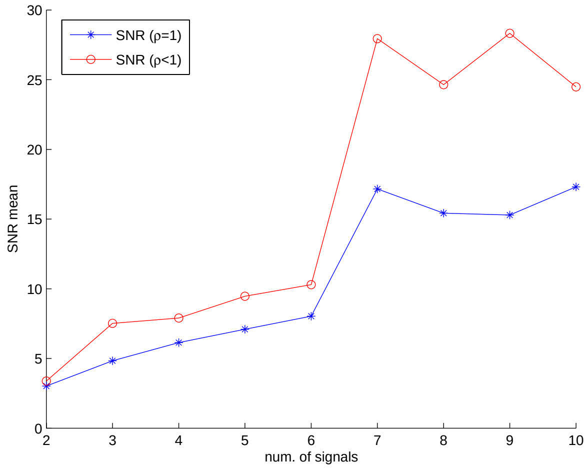

Fig. 3 The SNR for a group of signals of very high values of the beta parameter

so relevant speeding up of FastICA that it makes our technique suitable to real time applications.

VII. ACKNOWLEDGMENT

We would like to thank Alberto Bertoni and Simone Bassis for some helpful comments and suggestions.

REFERENCES

[1] P. Comon, “Independent component analysis - a new concept?” *Signal Processing*, vol. 36, pp. 287–314, 1994.

[2] C. Jutten and J. Herault, “Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture,” *Signal Processing*, vol. 24, pp. 1–10, 1991.

[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, J. W. . Sons, Ed. John Wiley & Sons, 2002.

[4] A. Hyvärinen and E. Oja, “A fast fixed-point algorithm for independent component analysis,” *Neural Computation*, vol. 9, pp. 1483–1492, 1997.

[5] W. B. Johnson and J. Lindenstrauss, “Extension of Lipschitz mappings into a Hilbert spaces,” *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[6] S. Amari and A. Cichocki, “Recurrent neural networks for blind separation of sources,” in *Proceedings of International Symposium on Nonlinear Theory and Applications*, vol. I, 1995, pp. 37–42.