SPIKES AND BUMPS: ARTEFACTS GENERATED BY INDEPENDENT COMPONENT ANALYSIS WITH INSUFFICIENT SAMPLE SIZE

Aapo Hyvärinen, Jaakko Särelä, and Ricardo Vigário

Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 2200, FIN-02015 Espoo, Finland
{Aapo.Hyvarinen, Jaakko.Sarela, Ricardo.Vigario}@hut.fi
http://www.cis.hut.fi/projects/ica/

ABSTRACT

We point out that if independent component analysis or blind source separation is performed in high dimensions with an insufficient sample size, this may lead to generation of artefactual source signals due to overlearning (or overfitting). Such artefactual source signals are practically zero almost everywhere, expect at the point of a single spike or bump. The existence of strong time-correlations in the data increases the probability of the occurence of the artefacts. These results are essentially independent of the particular algorithm used for ICA.

1. INTRODUCTION

Indendent component analysis (ICA) [4, 8] is a statistical model where the observed data is expressed as a linear transformation of source signals (or independent components) that are nongaussian and mutually independent. We may express the model as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{1}$$

where $\mathbf{x}(t) = (x_1(t), x_2(t), ..., x_m(t))$ is the vector of observed random variables, $\mathbf{s}(t) = (s_1(t), s_2(t), ..., s_n(t))$ is the vector of the independent components, and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. Exact conditions for the identifiability of the model were given in [4].

Several methods for estimation of the ICA model have been proposed in the literature [1, 2, 3, 4, 7, 8]. The performance of the algorithms is usually analyzed in terms of consistency, or classical finite-sample properties like asymptotic MSE [3, 6] and robustness [6]. The purpose of this paper is to point out that in the case of insufficient sample sizes, all the ordinary ICA methods tend to produce results that are characterized by estimates of the source signals (independent components) that have a single spike or bump, and are practically zero everywhere else. This is because the criteria in the ICA algorithms can be interpreted as measures of nongaussianity, and in the space of source signals of unit variance (and possibly with some constaints on frequency content as well), nongaussianity is usually maximized by such spike/bump signals. Thus this is a form of overlearning or overfitting typical of ICA methods. Such overlearning can sometimes be reduced by appropriate dimension reduction.

2. SPIKES AS SPARSITY MAXIMIZING SIGNALS

The phenomenon under consideration become easily comprehensible if we consider the extreme case where the sample size N equals the dimension of the data m, and these are both equal to the number of independent components n. Let us collect the realizations $\mathbf{x}(t)$ of $\mathbf{x}$ as the columns of the matrix $\mathbf{X}$, and denote by $\mathbf{S}$ the corresponding matrix of the realizations of $\mathbf{s}(t)$. Then (1) is of the form

$$\mathbf{X} = \mathbf{A}\mathbf{S}. \tag{2}$$

Note that all the matrices in (2) are square. This means that by changing the values of $\mathbf{A}$, we can give any values whatsoever to the elements of $\mathbf{S}$. This is a case of serious overlearning not unlike the classical case of regression with equal numbers of data points and parameters. Thus it is clear that the estimate of $\mathbf{S}$ that is obtained by ICA estimation depends little on the observed data. For example, assume that we constrain the estimates of the source signals to be uncorrelated and of unit variance [7], and assume that the densities of the source signals are known to be supergaussian (i.e. positively kurtotic [7]). Then the (constrained) ML estimation

of $\mathbf{A}$ consists of finding a $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_n)^T = \hat{\mathbf{A}}^{-1}$ that maximizes a measure of the supergaussianities (or sparsities) of the estimates of the source signals. For example, with Laplace distributions we obtain:

$$\mathbf{W} = \arg\max_{\mathbf{W}} \sum_t \sum_i -|\mathbf{w}_i^T \mathbf{x}(t)|. \quad (3)$$

where $\mathbf{W}$ is constrained to give uncorrelated source signals of unit variance. It is easy to prove that this is minimized by a $\mathbf{W}$ that gives as $\mathbf{S}$ a permutation (and sign change) matrix, i.e. source signals that are zero at all points except one, and those points are not overlapping. In unconstrained maximum likelihood estimation, the constraint is replaced by a penalty term of the form $\log|\det \mathbf{W}|$, which means that the maximizing $\mathbf{W}$ is slightly different, but still characterized by spiky source signals.

Thus we have shown that ICA estimation with an insufficient sample size leads to a form of overlearning that gives artefactual source signals. Such source signals are characterized by large *spikes*.

An important fact shown in the experiments section is that a similar phenomenon is much more likely to occur if the source signals are not i.i.d. in time, but have strong time-dependencies. In such cases the sample size needed to get rid of overlearning is much larger, and the source signals are better characterized by *bumps*, i.e. low-pass filtered versions of spikes. An intuitive way of explaining this phenomenon is to consider such signal as being constant on N/k blocks of k consecutive sample points. This means that the data can be considered as having really only N/k sample points; each sample point has simply been repeated k times. Thus, in the case of overlearning, the estimation procedure gives 'spikes' that have a width of k time points, i.e. bumps.

In some cases, overlearning can be reduced by appropriate dimension reduction by, for example, principal component analysis (PCA). If projections that contain noise (i.e. unnecessary information) are omitted as a preprocessing step, the data length/dimension ratio is improved, and artefactual estimates may be avoided.

3. EXPERIMENTAL RESULTS

Two sets of experiments were designed in order to illustrate the results presented above. In the first set, shown in Fig. 1, we used artifially generated signals to illustrate the basic phenomenon, as well as the effects of the choice of compression rate and filtering. The second set of experiments is presented in Figs. 2 through 4, dealing with real life medical applications.

All the experiments were made using MATLAB code, using either the fixed-point algorithm [7, 5] as implemented in the FastICA package, or the gradient descent algorithm for maximum likelihood (or infomax) estimation [1, 2, 3], as implemented in the package by Tony Bell. Both are available on the World Wide Web. The results were qualitatively similar for both packages.

3.1. Artificial data

Three positively kurtotic signals, with 500 sample points each, were used in these simulations, and are depicted in Fig. 1 *a*). 500 noisy mixtures were produced, where normally distributed i.i.d. noise was added to each weighted mixture separately. The variance of the added noise was 1/10000 of the variance of the signals.

As an example of a perfect ICA decomposition, Fig. 1 *b*) shows the result of applying the fixed-point and gradient descent algorithms to the mixed signals. In both approaches, the preprocessing (whitening) stage included a compression of the data into the first 3 principal components. It is evident that both algorithms are able to extract all the initial signals.

When the whitening is made with very small dimension reduction (we took up to 400 whitened vectors), we see the appearance of Dirac-like solutions, which is an extreme case of kurtosis maximization (Fig. 1 *c*)). The algorithm used in FastICA was of a deflationary type, from which we plot the first 5 components extracted. As for the gradient descent, which was of a symmetric type, we show 5 representative solutions to the 400 extracted.

Figure 1 *d*) presents an intermediate stage of compression (from the original 500 mixtures we took 50 whitened vectors). It is clear that most of the desired independent components are revealed by both methods, even though each resulting vector is noisier than the ones showed in *b*).

For the final example, in Fig. 1 *e*), we low-pass filtered the mixed signals, prior to the independent component analysis, using a 10 tap-delay MA filter. Taking the same amount of compression as in *d*), we can see that we loose all the original sources: the decompositions show a bumpy structure corresponding to the low-pass filtering of the Dirac-delta outputs presented in *c*). Through low-pass filtering, we have reduced the information contained in the data, and so the estimation is rendered impossible even with this, not very weak, compression rate.

3.2. EEG and MEG data

In earlier work, we have shown that FastICA is well suited for artefact removal from electro- and magne-

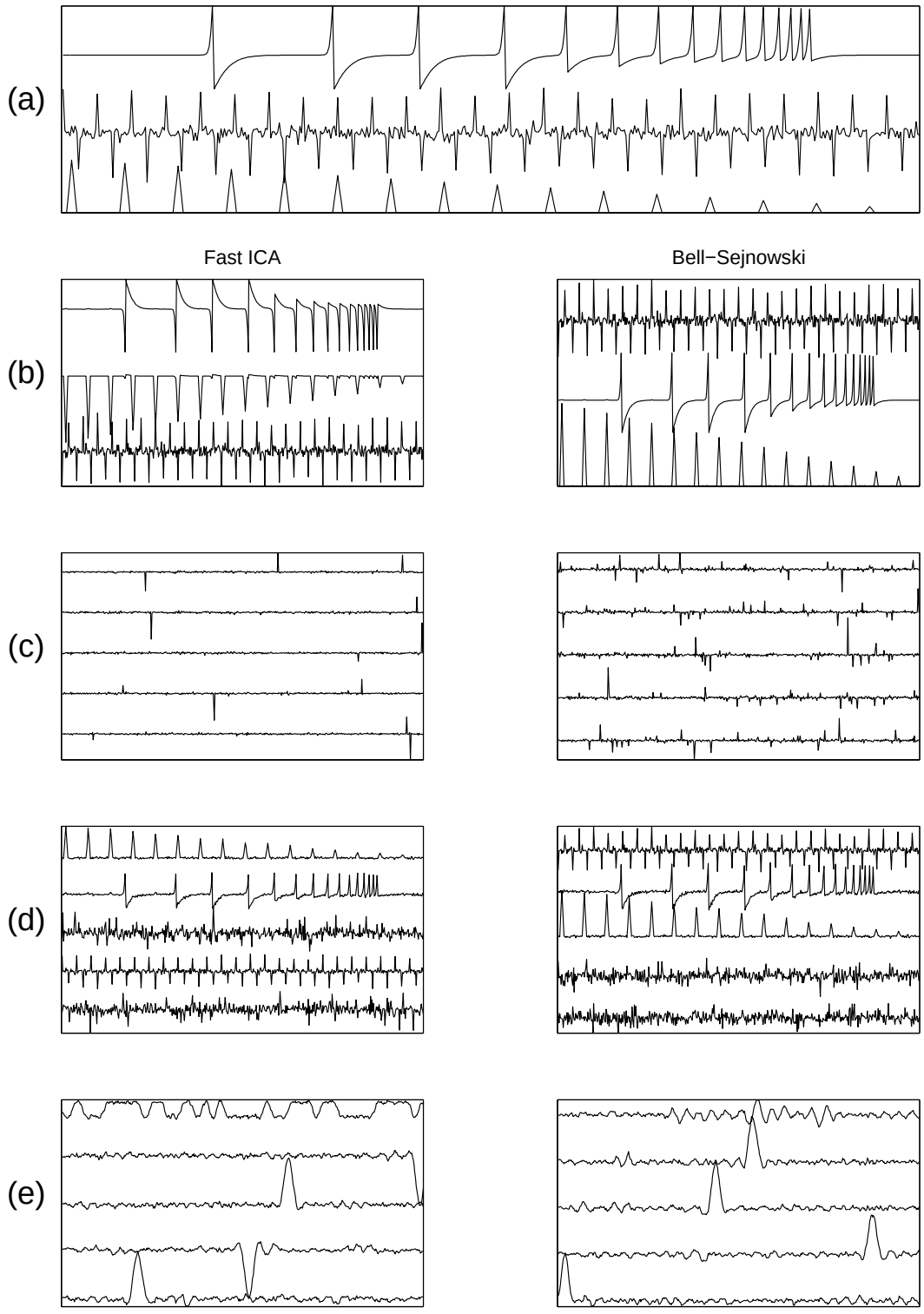

Figure 1: *Illustration of the importance of the choice of compression rate and filtering in artificially generated data, using a fixed point algorithm [7, 5] (the FastICA MATLAB package) and a gradient descent algorithm [1, 2, 3] (MATLAB code by Tony Bell). a) Original positively kurtotic signals . b) ICA decomposition in which the preprocessing includes a compression to the first 3 principal components. c) Poor, i.e. too low compression rate situation. d) Decomposition using an intermediate compression rate (50 components retained). e) Same results as in d) but using low-pass filtered mixtures*

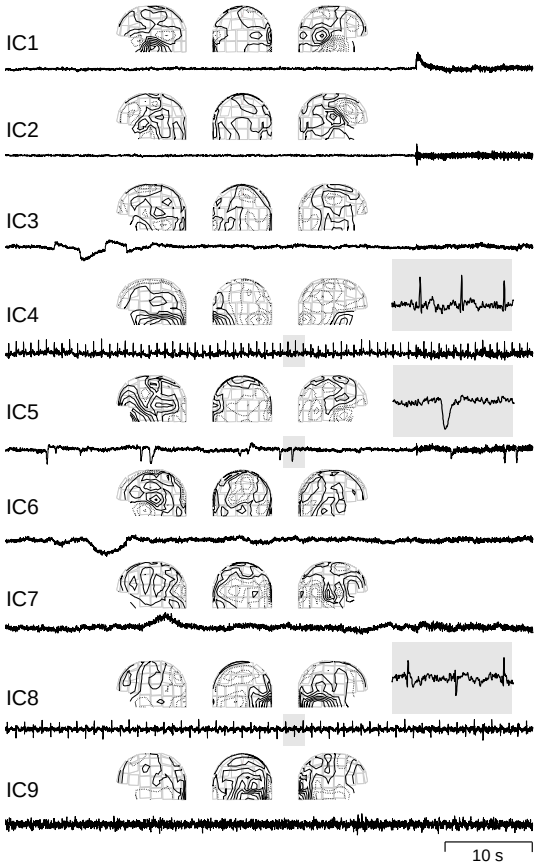

Figure 2: *Nine independent components found from the MEG data. For each component the left, back and right views of the field patterns generated by these components are shown — full line stands for magnetic flux coming out from the head, and dotted line the flux inwards (from [10]).*

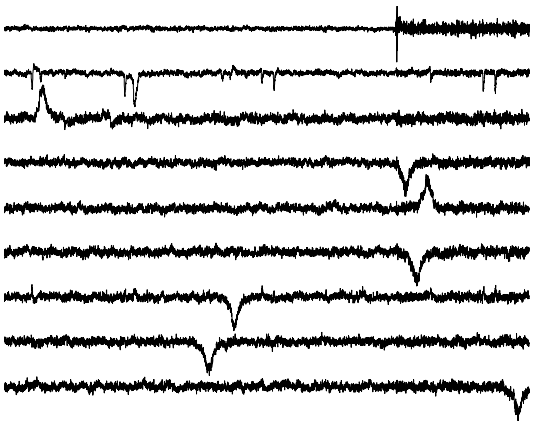

Figure 3: *Same study as in Fig. 2, in which the whitening stage was made without compression.*

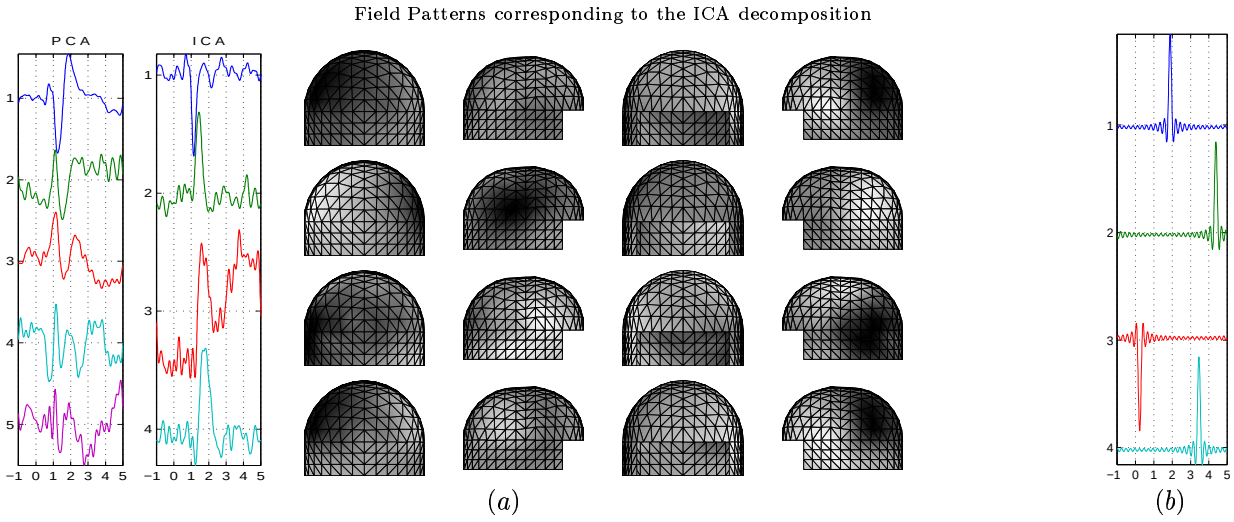

Figure 4: *Independent component decomposition of auditory evoked fields using a reasonable compression rate (a) [11], and no compression at all (b).*

toencephalographic recordings (EEG and MEG, respectively) [9, 10]. In addition, we have presented a study on ICA wave decomposition of auditory evoked fields [11]. In this section we will see how the present study affects the results reported on those papers.

During the extraction of artefacts from MEG data, reported in [10], we have used a reasonable compression rate, obtaining the results reproduced in Fig. 2. The field patterns are the regressions of each component on the original data, and can be seen as the columns of the estimated $\hat{\mathbf{A}}$ matrix. These patterns help interpreting and localizing the sources of the independent signals (e.g. IC1 and IC2 clearly represent activity of two different sets of muscles, in the right temporal area). A closer look into the field patterns of the bumpy signals IC6 and IC7 confirms the artefactual (overlearned) structure of the estimates, since the corresponding field patters are not physiologically meaningful. As a matter of fact, we can increase the number of solutions of that type by reducing the compression rate at the whitening stage. Figure 3 is an example where no compression was performed. Even though we can still see a couple of artefacts, most of the solutions obtained (120 out of 122) are meaningless bumps.

Finally, a clear illustration of the dangers of a poor choice of compression rate is depicted in Fig. 4. Figure 4 *a*) shows the decomposition of auditory evoked fields into independent components, as presented in [11]. In these components we can see the contra- and ipsilateral responses from the brain, to a train of auditory stimuli. When there is no compression in the data, the resulting decomposition is shown in Fig. 4 *b*). Note that the picture shown correspond to the extreme case of no compression. As seen in the previous example, the coexistence of *good* solutions with bumps is possible in an intermediate compression condition. Then, it may be difficult to distinguish between the independent components corresponding to meaningful solutions and the artefactual estimates shown in Fig. 4 *b*).

4. CONCLUSION

We showed a typical effect of overlearning (overfitting) by ICA algorithms. This consists of producing estimates of the source signals that are zero everywhere except for a single spike or bump. Reducing the dimension of the data by PCA is one way of reducing such overlearning. Of course, the best way to avoid overlearning would be to use a larger data set. Overlearning is especially probable to produce in the case of strongly time-dependent signals. We showed the relevancy of the results to separation of EEG and MEG signals, where artefactual bumps may be erroneously interpreted as meaningful signals.

5. REFERENCES

[1] S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, Cambridge, MA, 1996.

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

[3] J.-F. Cardoso and B. Hvam Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

[4] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[5] A. Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3917–3920, Munich, Germany, 1997.

[6] A. Hyvärinen. One-unit contrast functions for independent component analysis: A statistical analysis. In *Neural Networks for Signal Processing VII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, pages 388–397, Amelia Island, Florida, 1997.

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

[8] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[9] R. Vigário. Extraction of ocular artifacts from EEG using independent component analysis. *Electroenceph. clin. Neurophysiol.*, 103(3):395–404, 1997.

[10] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems 10*, pages 229–235. MIT Press, 1998.

[11] R. Vigário, J. Särelä, and E. Oja. Independent component analysis in wave decomposition of auditory evoked fields. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 287–292, Skövde, Sweden, 1998.