Journal of Machine Learning Research 4 (2003) 1447-1469 Submitted 10/02; Published 12/03

Overlearning in Marginal Distribution-Based ICA: Analysis and Solutions

Jaakko Särelä JAAKKO.SARELA@HUT.FI
Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Espoo, FINLAND

Ricardo Vigário RICARDO.VIGARIO@HUT.FI
Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Espoo, FINLAND
and
Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489 Berlin, GERMANY

Editors: Te-Won Lee, Jean-François Cardoso, Erkki Oja and Shun-ichi Amari

Abstract

The present paper is written as a word of caution, with users of independent component analysis (ICA) in mind, to overlearning phenomena that are often observed.
We consider two types of overlearning, typical to high-order statistics based ICA. These algorithms can be seen to maximise the negentropy of the source estimates. The first kind of overlearning results in the generation of spike-like signals, if there are not enough samples in the data or there is a considerable amount of noise present. It is argued that, if the data has power spectrum characterised by $1/f$ curve, we face a more severe problem, which cannot be solved inside the strict ICA model. This overlearning is better characterised by bumps instead of spikes. Both overlearning types are demonstrated in the case of artificial signals as well as magnetoencephalograms (MEG). Several methods are suggested to circumvent both types, either by making the estimation of the ICA model more robust or by including further modelling of the data.
Keywords: Independent component analysis, blind source separation, overlearning, overfitting, spikes, bumps, high-order ICA.

1. Introduction

Independent component analysis (ICA) (Jutten, 1987, Cardoso, 1989, Jutten and Herault, 1991, Comon, 1994, Hyvärinen et al., 2001) is a statistical signal processing technique that models a set of observations, $\mathbf{x}$, with an instantaneous linear mixing of independent latent variables, $\mathbf{s}$:

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

In this model, $\mathbf{n}$ is additive noise. Note that here we consider the mixing $\mathbf{A}$ to be time invariant and the latent components to be statistically independent. Furthermore, it is usual to assume, implicitly or explicitly, the observations to be independent in time.

The goal of ICA is to recover the latent components from the observations. If noise is negligible, this can be achieved by the determination of an inverse linear mapping from $\mathbf{x}$ to $\mathbf{s}$, say

$\hat{\mathbf{s}} = \mathbf{B}\mathbf{x}$. $\mathbf{B}$ is often found through the optimisation of some information theoretic criterion, which privileges estimates that are statistically independent. The mixing matrix can then be estimated as $\hat{\mathbf{A}} = \hat{\mathbf{B}}^{+}$. Most algorithms, directly or indirectly, minimise the *mutual information*, I, between the component estimates. It can be shown (Hyvärinen et al., 2001) that minimising I corresponds to the maximization of the *negentropy*, a measure of non-Gaussianity of the components.

The exact maximisation of the negentropy is difficult and computationally demanding because a correct estimation of the source densities is required. Most of the existing ICA algorithms can be viewed as approximating negentropy through simple measures, such as high-order cumulants (Cardoso, 1989, Hyvärinen, 1998, Hyvärinen et al., 2001). In Bayesian ICA (Lappalainen, 1999, Miskin and MacKay, 2001, Attias, 2001, Choudrey and Roberts, 2001, Chan et al., 2001), however, the source distributions are modelled explicitly, using a mixture of Gaussians model.

ICA can be used either to solve the so-called blind source separation (BSS) problem, or as a feature extraction technique. In BSS, the main focus is the determination of the underlying independent sources. This is the main goal when attempting to identify artifacts and signals of interest from biomedical signals, such as magnetoencephalograms (MEG) (Vigário et al., 1997, Jahn and Cichocki, 1998, Vigário et al., 2000, Tang et al., 2002) and electroencephalograms (EEG) (Makeig et al., 1996, Vigário, 1997, Jung et al., 2000). It has also been used to blindly separate audio signals (Torkkola, 1999), or to demix multispectral images (Parra et al., 2000, Funaro et al., 2001). This list is by no means an exhaustive one, for further applications see Hyvärinen et al. (2001).

The other central application for ICA is feature extraction, where it provides a set of bases, which can be used to represent the observed data. So far, some of the main applications of this feature extraction strategy include the study of natural image statistics (Hurri et al., 1996, Bell and Sejnowski, 1997) and the development of computational models for human vision (Hoyer and Hyvärinen, 2000, Hyvärinen et al., 2001), although it has been as well used for denoising (Hyvärinen, 1999). Once more, see Hyvärinen et al. (2001) for more detailed information.

We have shown (Hyvärinen et al., 1999) that, in the presence of insufficient samples, most ICA algorithms produce very similar types of overlearning. These consist of source estimates that have a single spike or bump and are practically zero everywhere else, regardless of the observations $\mathbf{x}$.

The present paper is written as a word of caution to users of any standard high-order ICA method. It revises the characteristics of the overlearning phenomena, and suggests some useful solutions. With this in mind, it is organised as follows: In Section 2 we review the ICA methods used in this paper. Section 3 describes the data sets used to illustrate the problems of spikes and bumps. In Section 4 we characterise the overlearning problems in more detail. Section 5 consists of several attempts to solve the problems in high-order ICA algorithms.

We recommend the further reading (Särelä and Vigário, 2003), for a complementary analysis of the present overlearning problem. There, we address the overlearning problem through the dual perspective of the accurate estimate of the ICA model and the adequacy of such a model for the data. That reference is strongly grounded on ensemble learning and Bayesian theory.

2. High-order ICA

In many popular algorithms for ICA, the unmixing matrix $\mathbf{B}$ is optimised by maximising a suitable *contrast function*, which measures the deviation of the source estimates from a Gaussian variable with same mean and variance. One family of such contrast functions is the high-order[1] cumulants:

1. Hence the name high-order ICA.

skewness, kurtosis etc. These are zero for Gaussian variables. Other possibilities include different approximations of negentropy.

In this paper, the FastICA algorithm (Hyvärinen, 1999) was chosen, without loss of generality, as a representative of the high-order ICA algorithms. Absolute value of kurtosis is used as the contrast function, due to its convenient analytical properties. However, in Section 5.3 the more robust contrast function are tested, to see whether they can aid in solving the overlearning problem in hand. All the results can easily be extended to other high-order ICA algorithms, as demonstrated elsewhere (Hyvärinen et al., 1999).

To estimate *one* of the independent components, FastICA uses a contrast function of the form $J_G(\mathbf{w}) = [E\{G(\mathbf{w}^T\mathbf{v})\} - E\{G(\nu)\}]^2$, where $\mathbf{v}$ stands for the whitened data and ν is a Gaussian variable with the same low-order statistics as the source estimates. This contrast function is maximised using an iterative Newton method. On one of the updates, the iteration is:

$$\mathbf{w}^+ = E\{\mathbf{v}g(\mathbf{w}^T\mathbf{v})\} - E\{g'(\mathbf{w}^T\mathbf{v})\}\mathbf{w}$$
$$\mathbf{w}^* = \mathbf{w}^+/\|\mathbf{w}^+\|,$$

where g is the derivative of G and $\mathbf{v}$ is the prewhitened data. To find more than one independent component, this contrast function can be maximised several times. The independent components are the local maxima of this contrast function. Hyvärinen (1999) suggested three different functions G:

$$G_1(s) = \log\cosh(s) \tag{2}$$
$$G_2(s) = -\exp(-s^2/2) \tag{3}$$
$$G_3(s) = \frac{1}{4}u^4.$$

There it was said that G_1 is a good general purpose function and G_2 is justified if robustness is very important. For sources of fixed variance, the maximum of G_3 coincides with the maximum of kurtosis, the fourth cumulant. All these contrast functions can be viewed as approximations of negentropy, though it is argued by Hyvärinen (1998) that cumulant based approximations, such as G_3, often give poor estimates, being mainly sensitive to the tails of the distributions, and thus sensitive to outliers as well.

Related gradient based technique is the so-called Infomax (Bell and Sejnowski, 1995), which is based on the maximization of the total entropy of the network outputs. Aside from gradient based algorithms there are also approximative algebraic solutions to ICA. The most widely known ICA algorithm of this type is JADE (Cardoso and Souloumiac, 1993), which is based on joint diagonalisation of fourth order cumulant tensors.

3. Data Sets

Artificially generated data, as well as MEG recordings, will be used to illustrate and characterise the spike and bump effects of overlearning. The artificial data is particularly important, as we are capable of following the complete generative process, and therefore assess the quality of the results. The MEG recordings are used due to their signal characteristics, as well as to show how the overlearning problems may affect the analysis of real data.

In particular, MEG was chosen because, with plausible assumptions, the relations between the electromagnetic sources inside the head, and the magnetic fields measured by the MEG apparatus were shown to be *linear* and *instantaneous* (Hämäläinen et al., 1993, Vigário et al., 2000).

On the other hand, the signal to noise ratio in MEG is often very poor, and the power spectrum of the noise can be characterised by a $1/f$-curve. Such a structure is typical to many complex natural phenomena (see Bak et al., 1988, for an extensive study on the subject), justifying the use of MEG as a representative of real world phenomena.

Finally, in spite of the stated above, MEG was chosen here for it has been successfully analysed by standard ICA techniques (see Vigário et al., 2000, Jahn and Cichocki, 1998, Tang et al., 2002).

The first data set, henceforth called *artificial data set*, consists of 500 samples of 500 linear mixtures of three artificially generated signals. The kurtoses of these three signals are respectively 8.65, -1.50, 12.16. The signals, together with three representative mixtures, are shown in Figure 1a-b. A negligible amount of noise is added to the mixtures to make the covariance matrix full rank. If more samples are needed, the original 500 are repeated, but the noise process is kept independent.

The second data set, henceforth called *MEG data set*, consists of magnetoencephalograms with deliberately induced artefacts. An extensive description of this data set, as well as a report on successful extraction of the artefacts using FastICA, can be found in work by Vigário et al. (1997). Twelve representative channels, from a total of 122, as well as some of the extracted components, are shown in Figure 1c-d.

In several portions of the paper, in particular when assessing the performance of the various attempts to solve the overlearning phenomena, a small set of targets may be required. If this is no problem in the case of the artificially generated data, the same is not true for the *MEG data set*. Hence, the first 2000 samples of a chosen subset of components in Figure 1d are used, and shown in Figure 1e. These components contain, respectively, the cardiac, the eye blinks and the digital watch artefacts. For simplicity of notation, we will call these the *MEG focus signals*.

To characterise the noise in the *artificial data set*, a *Gaussian i.i.d. data set* was generated. It consists of several channels of independent Gaussian noise. The noise is independent both in time and across channels.

Finally, to simulate the background MEG data, a *random walk data set* was generated. It consists of 12500 samples and 50 channels of independent random walk processes, with additive Gaussian noise. This should mimic the $1/f$ power spectra, typical of the background MEG. In contrast to the *Gaussian i.i.d. data set*, this is not independent over time, but only across channels. A representative set of channels is depicted in Figure 1f.

4. Problem Setting

4.1 Spikes Maximise Super-Gaussianity

The overlearning problem of spikes becomes easily comprehensible, when we consider a case where the sample size T equals the dimension of the data m. By collecting the realisations of $\mathbf{x}$ in (1) into a matrix and assuming the noise negligible, we get the following equation:

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

where all the matrices are square. Now, by changing the values of $\mathbf{A}$, we can give any values for $\mathbf{S}$. This phenomenon is similar to the classical overlearning in the case of linear regression with

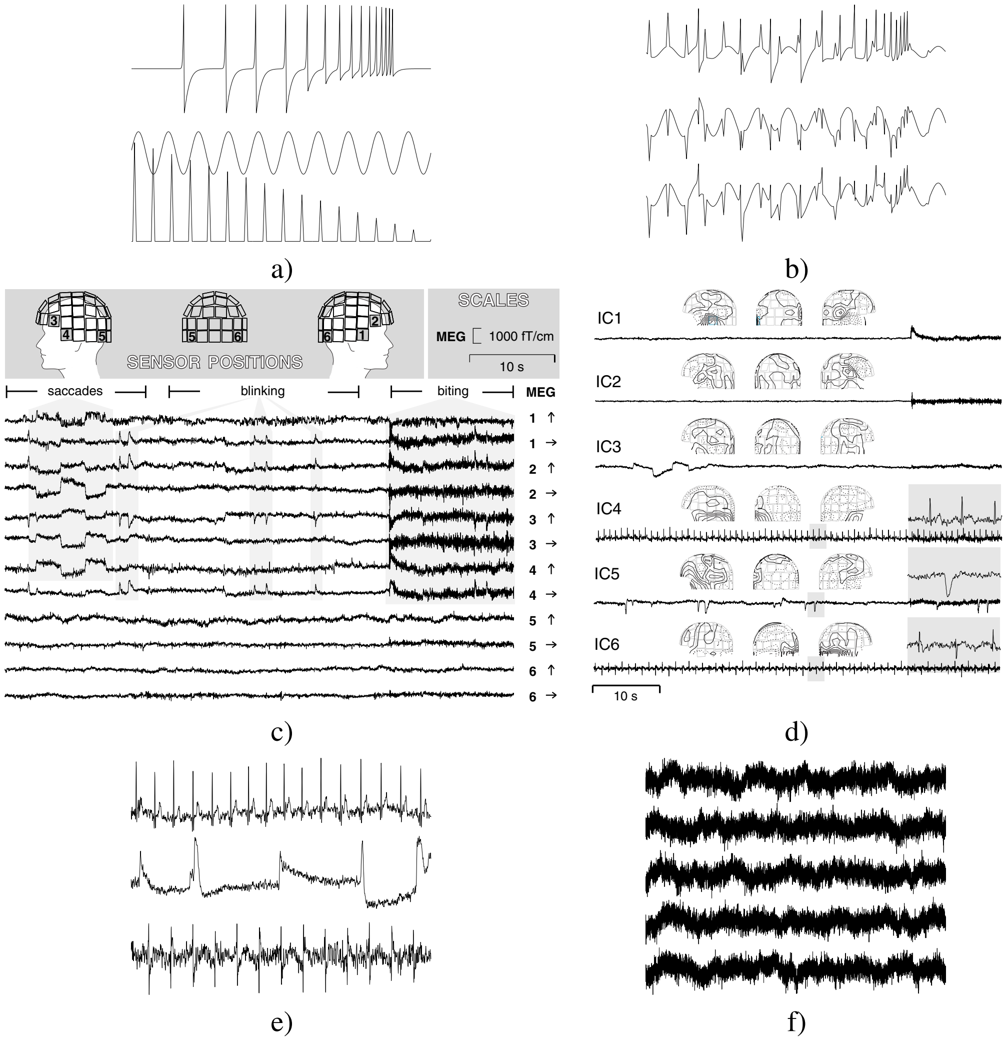


Figure 1: *a) Three artificial signals with zero mean and unit variance. b) Three representative mixtures, built from this data set. c) A subset of the* MEG data set *(Vigário et al., 2000). d) Some of the independent components found using FastICA (Vigário et al., 2000). e) MEG focus signals. f) A representative set of signals in the* random walk data set.

equal number of observations and parameters. In the present case then, whatever our principle for choosing $\mathbf{S}$ is, the result depends little on the observed data $\mathbf{X}$, but is totally determined by our estimation criterion.

Let us, as an example, consider the FastICA algorithm, using the absolute value of kurtosis as its contrast function (Hyvärinen, 1999). The following propositions are proven in the Appendix A:

Proposition 1 *Denote by* $\mathbf{s} = (s(1), ..., s(T))^T$ *a* T*-dimensional (non-random) vector that could be interpreted here as a sample of a scalar random variable. Denote by* H *the set of such vectors (samples)* $\mathbf{s}$ *that have zero mean and unit variance,* i.e. $\frac{1}{T}\sum_t s(t) = 0$ *and* $\frac{1}{T}\sum_t s(t)^2 = 1$.

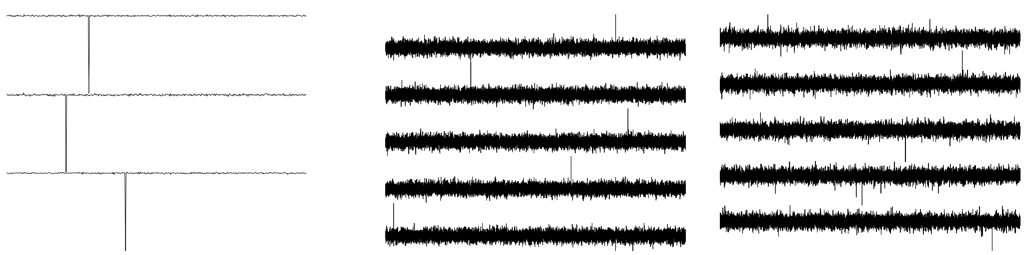
Then the kurtosis of $\mathbf{s}$*, defined as* $\frac{1}{T}\sum_t s(t)^4 - 3$*, is maximised in* H *by vectors of the form*

$$\mathbf{s}^* = \pm\sqrt{T}\mathbf{e}_i + o(\sqrt{T}), \tag{4}$$

where $\mathbf{e}_i$ *denotes a vector whose ith component equals 1 and other components are zero, and* $o(\sqrt{T})$ *denotes terms that are needed to have zero mean and unit variance and which are insignificant for large* T*. In other words, maximum kurtosis is attained by a spike-like signal, with just one significantly non-zero component. The distribution of the signal has maximally heavy non-symmetric tail maximally far away from zero, while the rest of the probability mass is concentrated on a single value very close to zero.*

Proposition 2 *Under the conditions of Proposition 1, also the absolute value of kurtosis is maximised by vectors of the form (4), for all practical sample sizes,* T*.*

From the above it follows that the absolute value of kurtosis is maximised by spike-like signals of the form (4). In the extreme case where we have as many samples, T, as dimension in the data, m, our estimate for the source matrix $\mathbf{S}$ is determined solely by the optimisation criteria. Hence, the estimate is roughly a permutation matrix, multiplied by $\sqrt{T}$. To illustrate this, the *artificial data set* is used, with $T = m = 500$. A sample of the results of applying FastICA to that data can be seen in Figure 2a. As expected, spikes are the result of such extreme case of overlearning. Note that the kurtoses of such components, respectively 480, 470 and 480, are far higher than those of the original sources.



a) b) c

Figure 2: *a) Three independent components, estimated by FastICA, from the* artificial data set. *b) Artificially generated spikes from the* Gaussian i.i.d. data set. *c) FastICA estimates for the* Gaussian i.i.d. data set.

4.2 Are we Saved if $T > m$?

It does not sound very dangerous to suffer from overlearning, when there are only as many samples as there are dimensions. Does the problem vanish, if we have more? In parameter estimation (*e.g.* linear regression), a rule of thumb states that you need, at least, a number of samples equal to ten times the number of free parameters. In the case of ICA, we need to estimate the unmixing matrix $\mathbf{B}$. Assuming prewhitened data, the number of free parameters is roughly $n^2/2$, where n is the number of sources.[2] Thus we should have $T > 5 \times n^2$ samples.

2. If $m > n$, the spare dimensions can be removed during whitening.

In the case of random variables (or infinite samples of random variables), it has been proven that, under rather flexible assumptions, *central limit theorem* (CLT) guarantees that a sum of n independent variables is more Gaussian than the most non-Gaussian original variable (see Papoulis, 1991). Thus, it is guaranteed that *e.g.* the kurtosis (or the absolute value of kurtosis, if you wish) of any linear projection $s_i = \mathbf{w}_i^T \mathbf{x}$ is at most equal to the kurtosis of the most kurtotic independent source. In some sense, this result precludes the emergence of the type of overlearning mentioned earlier. Yet, because we do not have infinite realizations of the measurements, it is conceivable to attain higher values of kurtosis than those of the most kurtotic source, as will be shown next.

Consider the *Gaussian i.i.d. data set*, with its 12500 samples of 50 dimensional "measurements". Since every channel is Gaussian and independent, every projection, according to the CLT, should be Gaussian. However, it is easy to produce a spike by enhancing one time instance and dampening all others: *e.g.* we simply use one time instance as our unmixing vector, *i.e.* $\hat{s}_i = \mathbf{x}(t_0)^T \mathbf{x}$, where $\mathbf{x}(t_0)$ corresponds to the realization of $\mathbf{x}$ at t_0. Some spikes generated in this way are shown in Figure 2b. Their positive kurtoses, ranging between 0.53 and 0.70, make them appear significantly non-Gaussian. FastICA produces similar spikes with comparable kurtoses, which can be seen from Figure 2c.

Hence, in the case of finite data sets, the maximization of measures such as the absolute value of the kurtosis, may result in the generation of spiky components, if these have greater kurtoses than the actual independent components.

4.3 Bumps Emerge, when Low Frequencies Dominate

Consider the *random walk data set*. Again, the marginal distributions of all channels are Gaussian and thus every projection should, in theory, be Gaussian. Nonetheless, because we are still in presence of finite data samples, we can try to generate spikes using the same strategy as before. Because nearby samples in time are strongly correlated, the result is not any more a single spike, but rather a bump. Figure 3a shows five such bumps. Spikes are still visible, but now they emerge in the middle of a small bump. The kurtoses of those signals are clearly non-zero (0.74, 0.96, 0.84, 0.65 and 0.92).

A stronger "non-Gaussian" effect can be produced by forcing the unmixing vector to be a weighted average around some time point. Then, the bump is

$$\begin{aligned} s(t) &= \mathbf{b}^T \mathbf{x}(t), \qquad \text{where} \\ \mathbf{b} &= \sum_{t=0}^{L} w(t)\, \mathbf{x}(t_0 - L/2 + t), \end{aligned} \tag{5}$$

where $w(t)$ is the windowing function, normalised to $\sum w(t) = 1$. $L+1$ is the width of the window, which is, at best, also the width of the bump. Some bumps, using a triangular window of length $L = 1001$, are presented in Figure 3b. Here the spike is completely absorbed by the bump and the kurtoses are even greater (1.15, 0.98, 1.48, 0.69 and 1.51). Again, FastICA produces similar results (Figure 3c).

Note that the kurtoses of the generated smooth bumps are already quite close to 2, the theoretical maximum for the absolute value of the kurtosis of sub-Gaussian signals. This limitation may render practically impossible for ICA to find sub-Gaussian sources, such as some rhythmic activities in MEG recordings.

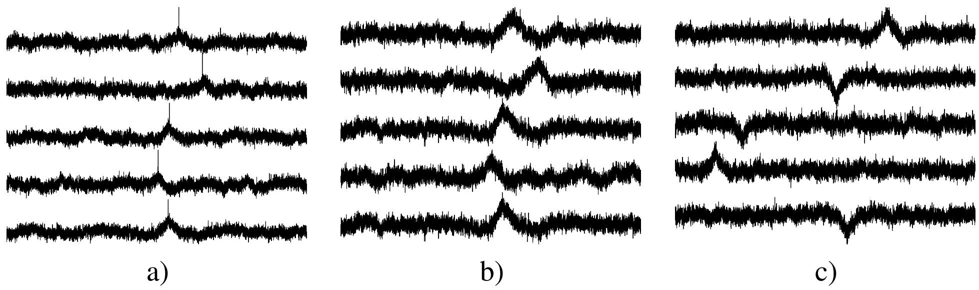


Figure 3: *a) Artificially generated bumps from the* random walk data set. *b) Artificially generated smooth bumps. c) Components estimated by FastICA.*

4.4 Bumps are the Preferred Overlearning in MEG

Due to its $1/f$ nature, MEG data tends to show a bump-like overlearn, much like the one in the *random walk data set*, rather than the spike type observed for the *Gaussian i.i.d. data set*. This fact can be observed in the last five components shown in Figure 4.

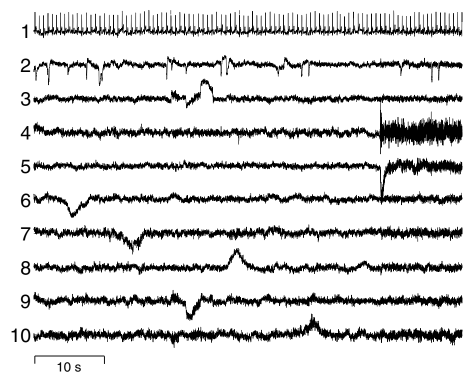


Figure 4: *Some ICA components, including both true underlying sources (top) and bump-like overlearnings (bottom).*

Because of the considerable amount of data (over 12000 samples, for a total of 127 channel measurement), as well as the high values of kurtoses of several meaningful components, these seem to be allowed to co-exist with the overlearns (see the first five components in Figure 4). The kurtoses of the first five components range from 10.30 to 44.46, whereas the last five range from 9.37 to 14.10.

5. Attempts to Solve the Problems in High-order ICA

Overlearning may stem from an improper estimation of the ICA model. In that case, a solution may be attempted by either increasing the number of samples per free parameter estimated, or by adopting a more robust approximation for the negentropy. Sections 5.1 and 5.2 will deal in some detail with the former strategy, whereas Section 5.3 will address the latter one.

If the ICA model is insufficient or inadequate for the data, a solution may lie in additional modeling of the data. In Section 5.4 we present two such strategies, where local temporal relations are removed, and standard ICA is applied to the remaining innovation processes.

Once more we recommend the reading of Särelä and Vigário (2003) for further information on these two forms of overlearning, as well as for a Bayesian approach to their solutions.

5.1 Acquiring more Samples

Asymptotically, the kurtoses of the spikes and bumps tend to zero, when the number of samples increases. How fast does that happen? Consider, once again, the generation of spikes in the case of the *Gaussian i.i.d. data set*. The data has no intrinsic high-order statistics (see Section 3), but it is prone to spike-like overlearning (Section 4.2). With very large number of samples, T, it can be expected that the source estimate, $\mathbf{x}(t_0)^T\mathbf{x}(t)_{t\neq t0}$, except for the spike, still has zero mean and unit variance. Then, the kurtosis of the component can be easily approximated (see proof in Appendix A) by:

Proposition 3 *Let $\mathbf{x}(t)$ be Gaussian i.i.d. signal. Let $s(t)$ be a spike generated from $\mathbf{x}(t)$ by $s(t) = \mathbf{x}_0^T\mathbf{x}(t)$, where $\mathbf{x}_0 = \frac{\mathbf{x}(t_0)}{||\mathbf{x}(t_0)||}$ is the unmixing vector and t_0 is the time instance at which the spike appear. Then the kurtosis of the spike $s(t)$ is approximately*

$$kurt(s(t)) \approx \frac{||\mathbf{x}(t_0)||^4}{T} \propto \frac{1}{T}, \tag{6}$$

In some cases, gathering more data can be easy, which could solve the problem, provided that the kurtoses of the expected components are not very close to zero. An illustration of this fact is given in Figure 5, where the *artificial data set* is analysed, with fixed dimensions, $m = 100$, and varying sample sizes. Figure 5a shows the correlation between the original signals and their closest estimates.[3] The number of samples varied from 500 to 30000. The two super-Gaussian signals, with their considerably high kurtoses, are correctly estimated with as few as 5000 samples, whereas a considerably larger amount of samples is required to reliably recover the sinusoidal signal, with its sub-Gaussian distribution.

A similar approach to the one summarised by (6) can be derived to approximate the relation between the kurtosis of a bump and the size of its samples, T. With the assumptions of a large T, and zero mean, unit variance and zero kurtosis for $s(t)$ when t outside the range $[t_0 - L/2, t_0 + L/2]$, we get:

$$\mathrm{kurt}(s(t)) \approx \sum_{t=t_0-L/2}^{t_0+L/2} \left(s(t)^4 - 6s(t)^2\right)/T, \tag{7}$$

where L is the width of the bump. Note that the assumptions made here are not as justified as in the case of spikes, since the width of the bump, L, may not be negligible when compared with T. Still, for a sufficiently large T, the error committed should be small. Thus we may be tempted to conclude that the kurtosis of a bump is proportional to $1/T$, as in the spikes. However, experimental evidence shows that the width of a bump grows roughly proportional to $\log T$. Hence the kurtosis of a bump is better described by the relation $\mathrm{kurt}(s(t)) \propto \log T/T$, which decreases much slower than $\mathrm{kurt}(s(t)) \propto 1/T$. Furthermore the kurtoses of bumps tend to be bigger than the kurtoses of spikes,

3. The graph shows an average over 100 different runs, in order to obtain a somewhat representative curve.

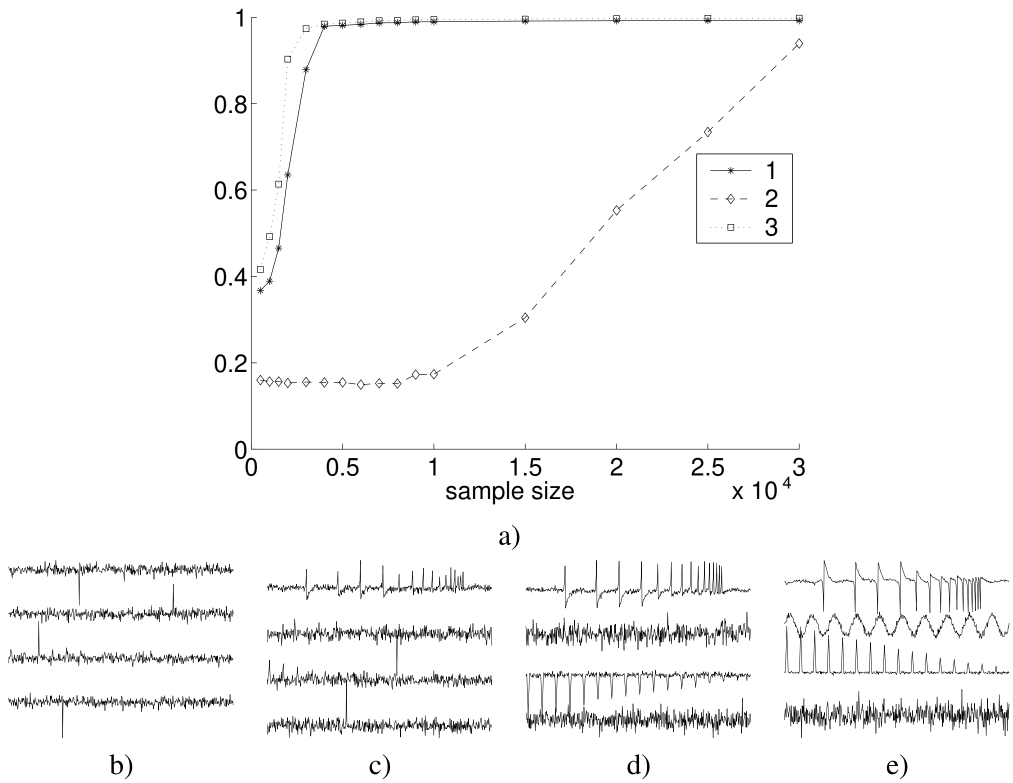
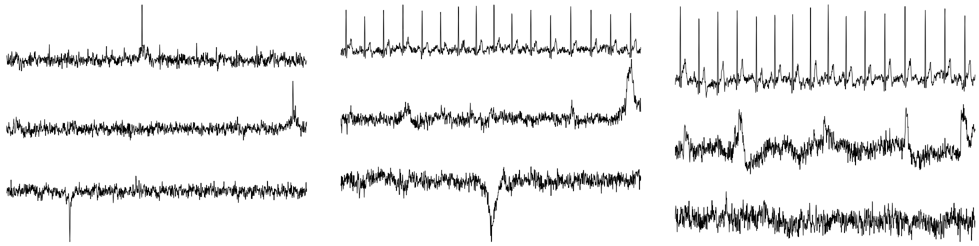


Figure 5: *a) Correlations between the original signals and the corresponding estimates, as a function of the sample size. ICA results for b) 500 points c) 1000 points d) 2000 points e) 9000 points. To show the possible spikes, only the segments of the 500 samples having the largest absolute value of kurtosis are shown.*

for a given sample size T. Thus the attempt to circumvent the bump-like overlearning phenomenon requires a much greater increase in the sample size than its spike counterpart.

An illustration of the effect of acquiring more samples to overcome a bump-like overlearning behaviour is shown in Figure 6, where the *MEG data set* is analysed. The number of samples varied from 2000 to 17000. Only the three estimated components that best matched the *MEG focus signals* are shown. ICA was able to separate the first focus component with as few as 4000 samples, but neither the second nor the third focus components were clearly identified, even with the highest sample sizes. Yet, some evidence of eye blinks is visible on the second component of Figure 6b. This component showed as well a bump that is not visible from the samples plotted. For comparison reasons, only the first 2000 samples are shown, for all configuration sizes. The third component consisted of a single bump, also not visible in the chosen region. Note as well the change in the width of the bumps when using 2000 or 4000 samples.

In conclusion, we may consider the acquisition of more samples to solve a spike-type of overlearning problem. Yet, this strategy does not seem to be an efficient one when dealing with a bump-type one. Furthermore, in many cases the gathering of more data may be expensive and/or



a) b) c)

Figure 6: *ICA analysis of the* MEG data set*, using a) 2000 points b) 4000 points c) 17000 points. Only the first 2000 samples are plotted.*

time consuming (*e.g.* increase the, often already long, time of acquisition of medical data). Hence there is need for more efficient strategies to solve these problems.

5.2 Reducing Dimension

Consider once more the spikes generated from the *Gaussian i.i.d. data set*. Their kurtoses do not depend only on the amount of samples, but are also highly dependable on the dimension of $\mathbf{x}(t)$. Using a similar strategy as in Proposition 3, we can approximate the relation between the kurtosis of a spike and the dimension of the data, by (the proof is in the Appendix A):

Proposition 4 *Let $s(t)$ be a spike as in Proposition 3. The kurtosis, in function of the number of dimensions m is*

$$kurt(s(t)) \approx \frac{||\mathbf{x}(t_0)||^4 - 6\,||\mathbf{x}(t_0)||^2}{T} = \frac{m^2 - 6m}{T} \propto m^2,$$

for large values of m.

Therefore, dimension reduction is a more efficient way to reduce the kurtosis of spikes than the increase in amount of samples.

Again, we can try the same reasoning to the study of bumps. We start from (7) and assume that a bump is generated using (5). In addition, it has become clear, through experiments, that the dimension m does not affect the width of the bump, hence

$$\mathrm{kurt}(s(t)) \propto m^2.$$

Overlearning phenomena do not occur due to a poor absolute amount of samples, but rather a poor ratio between this value and the dimension of the data, $\mathbf{x}(t)$. Hence, decreasing the number of dimensions seems a much more efficient way to avoid spikes and bumps than acquiring more samples. It is also easier to leave behind something we already have than to gather some new information.

How should we do this reduction optimally? If $m > n$, *i.e.* we have more sensors than sources, the solution is immediate. If the noiseless ICA model is the correct one, $m - n$ eigenvalues of the covariance matrix are zero, thus a projection to the non-zero eigenvectors loses no information. On the other hand, if the noisy ICA model is the correct one, *i.e.* there is some amount of independent sensor noise, we can still do a dimension reduction effectively using principal component analysis (PCA), which guarantees optimal reconstruction in a mean squared error sense. If the sensor noise has weak energy, its contribution will be mainly concentrated in the components having the smallest variances.

However, it can happen that the reduction of dimension n is not enough, even if performed in an optimal manner, *e.g.* if the number of observed channels, equal to the number of underlying sources, is too big for the number of samples. Could PCA still be used? In that case it is no longer possible to estimate accurately all the independent sources using traditional ICA algorithms. If the number of samples is sufficiently big, it can be hoped that the resulting components are some meaningful sums of a few of the original independent sources.

On the other hand, it is no more clear that the components corresponding to the largest eigenvalues should be the ones to be kept. This leads to a combinatorial problem. Some criterion could be made to choose the right ones. For example the directions with highest kurtoses could be taken.

Dimension reduction was tested, as a way to avoid spikes, using the *artificial data set*. PCA was used to perform different reduction rates. Figure 7a shows the correlations between the original sources and their closest estimates, as a function of the dimension of the compressed space. Note that only when keeping less than 30 principal components can we expect reliably to find all the three original sources. This result is also visible from Figures 7b-e. Once again, the sinusoidal component is the one that requires the strongest compression rates to emerge. Furthermore, the co-existence of reasonable estimates and spike-like solutions are visible for sub-optimal compressions.

The dimension reduction strategy was as well used to deal with the bump-like overlearning in the *MEG data set*. The results for different reduction rates are shown in Figure 8. With no dimension reduction, all components are essentially bumps,[4] and no good estimate of the *MEG focus signals* could be found. A reduction to a 122-dimensional space brings forth an almost perfect cardiac artefact. The eye blinks become very clear when keeping 77 or less principal components. Still, no compression rate enabled a consistent accurate estimate for the digital watch. We also see that a too radical compression (under 25) weakens the extraction capability: the *MEG focus signals* may not be well represented in the highest principal components.

We conclude that dimension reduction is an efficient way to avoid spikes, if we have more sensors than sources. If this is not the case, dimension reduction cannot be used in principle. In practice, we may use it and hope that the components found, sums of the underlying independent sources, are still somehow meaningful. However, avoiding bumps using dimension reduction seems more problematic, *e.g.* no digital watch could be found. This would suggest that the problem of bumps may arise also from the inadequacy of the model and not only from the fact the model is not estimated properly (see Särelä and Vigário, 2003, for further discussion on these matters).

5.3 More Robust Estimates for the Negentropy

So far we have used the absolute value of the kurtosis as the contrast function in FastICA. Hyvärinen (1998), however, argues that measures based on cumulants, give often poor estimates for the negentropy, being mainly sensitive to the tails of the distributions. There, better approximations for negentropy are proposed.

4. The bumps are not shown in the figure since, for easy comparison, only the first 2000 points of the components are shown.

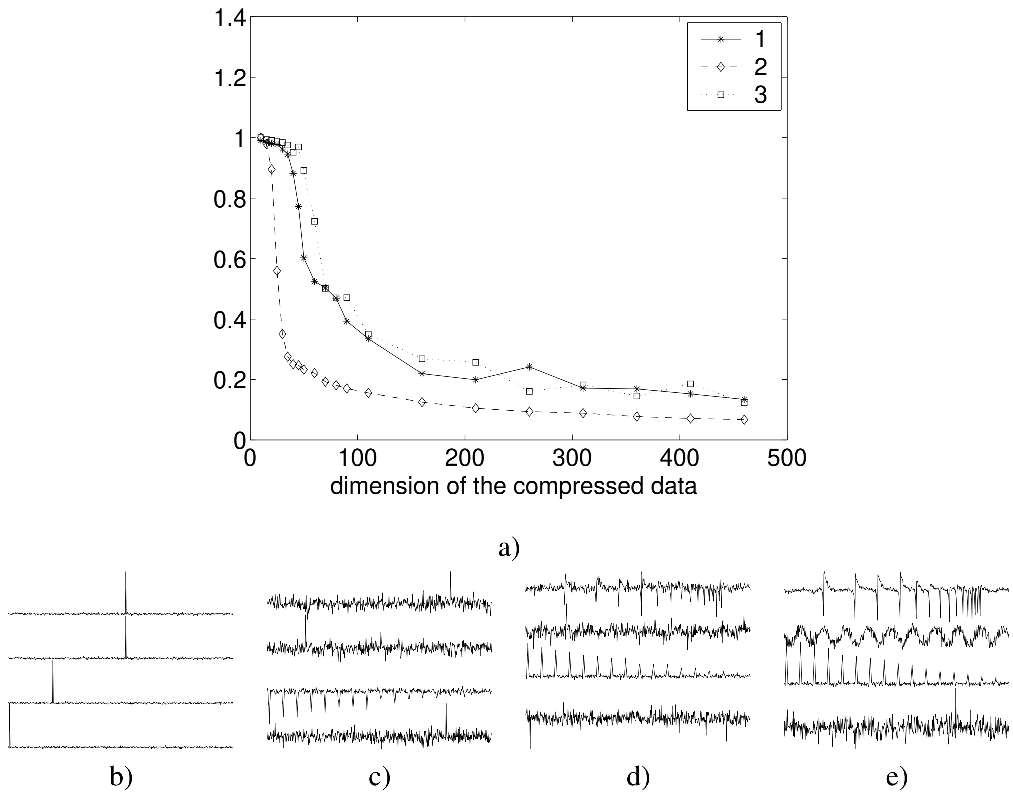


Figure 7: *a) Correlations between the original signals and the corresponding estimates, as a function of the compressed dimensions. ICA results of b) 460 dimensions c) 60 dimensions d) 45 dimensions e) 30 dimensions.*

The experiments on dimension reduction, reported in the previous section for the *artificial data set* and the *MEG data set*, were now repeated using G_1 (2) and G_2 (3) as contrast functions. The resulting correlation plots are shown in Figure 9.

No significant differences seem to exist, when comparing the results of the robust contrast functions to the kurtosis based one. If, in the *artificial data set*, these functions produce reliable estimates for the underlying sources, requiring only mild compression rates, this does not seem to be the case for the *MEG data set*.

5.4 Solving ICA for the Innovation Process

It may be possible to divide, linearly, the observations $\mathbf{x}(t)$ into two terms, $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$, in such a way that only $\mathbf{x}_1(t)$ is proned to a bump type of overlearning. Then, it should be possible to estimate the original mixing matrix $\mathbf{A}$ from the relation $\mathbf{x}_2(t) = \mathbf{A}\mathbf{s}_2(t)$. Similarly, $\mathbf{s}_2(t)$ corresponds now to the portion of the underlying sources, independent and non-Gaussian, which are associated with $\mathbf{x}_2(t)$. This should solve the problem of bumps.

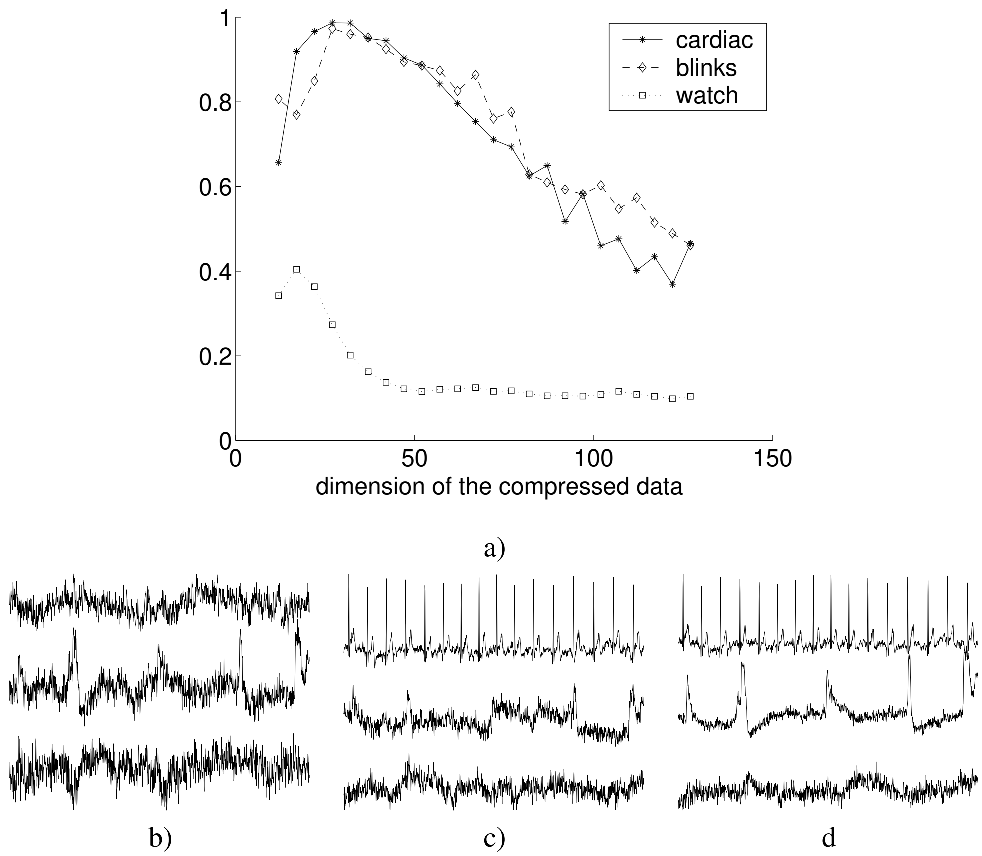


Figure 8: *a) Correlation between the* MEG focus signals *and their best estimates. ICA results for b) 127 dimensions c) 122 dimensions d) 77 dimensions.*

In the following sections we introduce two possible ways to implement this idea: the first one is based on a simple high-pass filtering pre-processing, whereas the second, more data driven, models the low frequencies by means of an AR process.

5.4.1 FILTERING THE LOW FREQUENCIES

Bumps are mainly dominated by low frequencies. Hence, dividing $\mathbf{x}(t)$ into a low frequency content $\mathbf{x}_1(t)$, prone to such overlearnings, will free $\mathbf{x}_2(t)$ to guide the search for the underlying independent components. Such a division may be achieved by simply high-pass filtering the data, prior to the ICA decomposition.

To find the correct cut-off frequency for the *MEG data set*, the average spectrum of the true artefacts is compared to the spectrum of the 7th source estimate in Figure 4, which contains a clear bump. Potentially disturbing effects, due to noise, were reduced by setting the signal to zero, for all sample points other than the bump itself. The average spectrum and the spectrum of the cleaned bump are shown in Figure 10a with solid and dashed lines respectively. Most of the bump spectrum is below 1 Hz, which was selected to be the cut-off frequency.

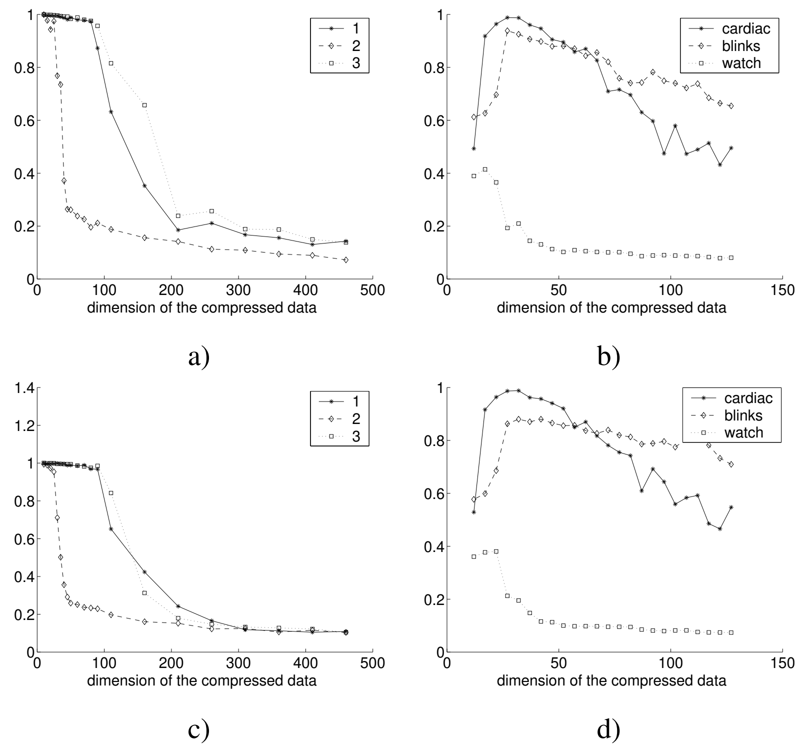


Figure 9: *Correlations between the target sources and their best estimates, using a) the* artificial data set *and* G_1 *b) the* MEG data set *and* G_1 *c) the* artificial data set *and* G_2 *d) the* MEG data set *and* G_2.

After high-pass filtered, the *MEG data set* was further analysed by FastICA. The independent components that best matched the *MEG focus signals* are shown in Figure 10b. For the first time, so far, we were capable of recovering all the *MEG focus signals* accurately.

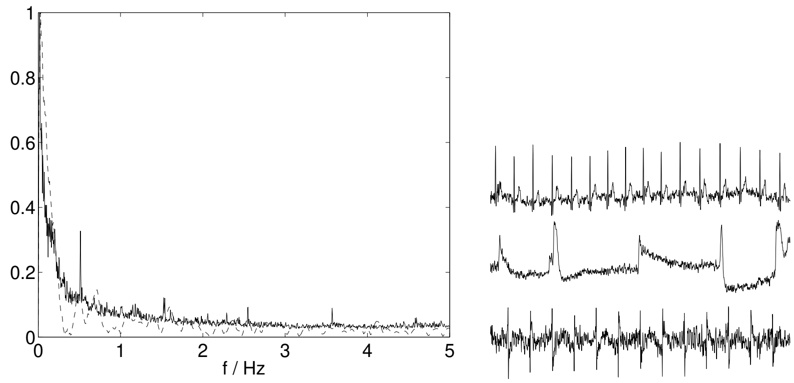
One problem in performing a simple high-pass filtering of the data is the determination of its cut-off frequency. In experiments, it was observed that the width of the bumps increases with the number of samples. Thus a constant cut-off frequency cannot be chosen for all MEG recordings. Data driven techniques should enable to find the best suited filtering parameters, in a more robust manner. One such method could be based on the singular spectrum analysis (SSA). For a recent review, see Yiou et al. (2000). Another, much simpler, will be explored in the next section.

5.4.2 AUTOMATIC FILTERING USING AR-PROCESSES

To automate the filtering, one may model the low frequencies using some simple model, *e.g.* an auto-regressive model:

$$\mathbf{x}(t) = \sum_{\tau=1}^{\mathcal{T}} \mathbf{c}_\tau^T \mathbf{x}(t-\tau) + \mathbf{x}_2(t),$$

where the sum is the AR-process and $\mathbf{x}_2(t)$ the innovation process.



a) b)

Figure 10: *a) Average spectrum of the artefacts in the MEG (solid line) and of a cleaned bump (dashed line). b) Results of applying FastICA to the high-pass filtered data.*

To test this technique with the *MEG data set*, we selected to model the low frequencies with a one-tap AR process: $\mathbf{x}(t) = c_1\mathbf{x}(t-1) + \mathbf{x}_2(t)$. For simplicity a common AR-coefficient was used for all channels. The AR-coefficient was estimated from the data resulting in $c_1 = 0.9$. This comes close to basic random walking or Brownian movement. After removing the AR-process from the data, FastICA was applied to the innovation process $\mathbf{x}_2(t)$. The components that best estimate the *MEG focus signals* are shown in Figure 11. All artefacts are now perfectly recovered, better so than with simple high-pass filtering.

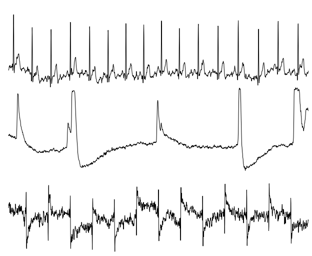


Figure 11: *Results with AR-process filtering.*

Hyvärinen (2001) presents a comparable method. There, the sources, rather than the observations, are modeled by a combination of an AR process and an innovation process. When the required assumptions for both methods are met, the methods are in fact identical.[5]

6. Discussion

In this paper we discussed the overlearning problem in standard independent component analysis (ICA) algorithms based on marginal distribution information (Hyvärinen et al., 1999). The resulting

5. Amari (2000) gives another related method.

overlearned components have a single spike or bump and are practically zero everywhere else. This is similar to the classical overlearning in *e.g.* linear regression.

The spike problem is typically due to a poor estimate of the ICA model, either because the contrast function is not appropriate or because it is inaccurately estimated. The solutions to this problem include the acquisition of more samples, or the reduction of dimensions. The latter one was shown to be a more efficient way to avoid the problem, provided that there are more sensors than sources.

The overlearning problem is especially severe and cannot be solved properly in practice by acquiring more samples nor by dimension reduction, when the data has strong time dependencies, such as a $1/f$ power spectrum. This spectrum characteristic is typical of magnetoencephalograms (MEG) as well as many other natural data. In MEG, this problem greatly reduces the potential of ICA in analysing the weak but complex brain signals, making the problem important to solve.

Attempts to avoid the overlearning bumps that occur in such situations can be done by imposing further modelling restrictions to the data. We presented several possible ways to achieve this: In the innovation process approach, the data is preprocessed so that the ICA model becomes an adequate model. This preprocessing can be done in various ways, of which high-pass filtering is maybe the simplest one. More data driven techniques such as singular spectrum analysis and modelling the low frequencies by an AR process were also discussed. This last approach achieved remarkable results for the analysis of the MEG data.

It has been argued that measures based on cumulants, such as the kurtosis, often fail in providing an adequate estimate of negentropy, which is a proper contrast function for ICA. More robust contrast functions have been developed. In this paper, we tested whether these help in solving the overlearning problems. They seem to be of some help, especially in the problem of spikes, but no decisive conclusions could be drawn.

Theoretical considerations on the origins of the overlearning problem can be found in related work (Särelä and Vigário, 2003). There, we analyse, in the Bayesian framework of ensemble learning, its dual nature. This duality stems from the adequacy of the ICA model for the date, and from its accurate estimation. The results therein are in par with the best solutions presented in this paper.

Not all BSS algorithms rely on high-order statistics. Another efficient family is based on canceling cross correlations of delayed copies of the data (see Belouchrani et al., 1997, Ziehe and Müller, 1998, for a fuller account). These methods, as any other optimisation strategy, have their own forms of overlearning. As they are not marginal distribution based ICA methods, they are not strictly in the scope of this paper. Therefore, we won't engage into detailed description of their overlearning. We simply state that the overlearning of this family results in various periodic solutions, mostly sinusoidal. Some discussion of overlearning in source separation algorithms can as well be found elsewhere (Valpola and Särelä, 2003).

Meinecke et al. (2002) and Müller et al. (2003) assess the reliability of an ICA decomposition by means of a resampling strategy. There, overlearned components show poor reliability. Although a valuable tool to detect overlearning effects, this strategy presents no solution to overcome them.

From the results in this study, we would suggest that, in presence of a spike-type overlearning, dimension reduction may be the simplest solution. If, on the other hand, we face a bump-type overlearning, the combination of an AR and your preferred ICA algorithmic approach may be better suited.

Acknowledgments

JS is funded by EU BLISS project (IST-1999-14190) and by the Academy of Finland. RV is funded by EU Marie Curie Fellowship (HPMF-CT-2000-00813). The authors wish to thank Dr. Harri Valpola for useful discussion and Dr. Aapo Hyvärinen for the proofs of Propositions 1 and 2 in Appendix A.

Appendix A. Proofs

Proposition 1 *Denote by* $\mathbf{s} = (s(1), ..., s(T))^T$ *a* T*-dimensional (non-random) vector that could be interpreted here as a sample of a scalar random variable. Denote by* H *the set of such vectors (samples)* $\mathbf{s}$ *that have zero mean and unit variance,* i.e. $\frac{1}{T}\sum_t s(t) = 0$ *and* $\frac{1}{T}\sum_t s(t)^2 = 1$.

Then the kurtosis of $\mathbf{s}$*, defined as* $\frac{1}{T}\sum_t s(t)^4 - 3$*, is maximised in* H *by vectors of the form*

$$\mathbf{s}^* = \pm\sqrt{T}\mathbf{e}_i + o(\sqrt{T}), \tag{8}$$

where $\mathbf{e}_i$ *denotes a vector whose ith component equals 1 and other components are zero, and* $o(\sqrt{T})$ *denotes terms that are needed to have zero mean and unit variance and which are insignificant for large* T*. In other words, maximum kurtosis is attained by a spike-like signal, with just one significantly non-zero component. The distribution of the signal has maximally heavy non-symmetric tail maximally far away from zero, while the rest of the probability mass is concentrated on a single value very close to zero.*

Proof: What we need to find are the maxima of $\sum_t s(t)^4$ under the constraints:

$$\frac{1}{T}\sum_t s(t)^2 = 1 \tag{9}$$

$$\frac{1}{T}\sum_t s(t) = 0 \tag{10}$$

The Kuhn-Tucker conditions show that at the maxima we must have

$$\lambda_1\mathbf{c} + \lambda_2\mathbf{s} = \mathbf{s}^3 \tag{11}$$

where $\mathbf{c}$ denotes a vector of all ones. By summing all the equations in (11) and using (10), we get

$$\lambda_1 = \frac{1}{T}\sum s(t)^3.$$

In a similar manner, we get λ_2 by multiplying the equations by $\mathbf{s}$, summing and using both (9) and (10):

$$\lambda_2 = \frac{1}{T}\sum s(t)^4$$

The constraint $\frac{1}{T}\sum_t s(t)^2 = 1$ implies that the $s(t)$ are at most of order $\sqrt{T}$. On the other hand, as will be seen below, the maximising $\mathbf{s}$ does contain components of order $\sqrt{T}$. This means that the first term on (11) is of a smaller order (at most $\sqrt{T}$) than the other terms (which are of order $T\sqrt{T}$). Thus the first term can be neglected as a first-order approximation, and we need to solve only

$$\lambda_2 \mathbf{s} = \mathbf{s}^3.$$

This condition is clearly fulfilled if and only if $\mathbf{s}$ is of the form:

$$\mathbf{s}^0 = \sqrt{\lambda_2} \sum_{i \in S} \pm \mathbf{e}_i$$

for some index set S. In other words, the components of $\mathbf{s}^0$ must be either zero, or equal to $\pm\sqrt{\lambda_2}$. On the other hand, due to $\frac{1}{T}\sum_t s(t)^2 = 1$, we have

$$\sum_t s^0(t)^2 = \lambda_2 \#S = T$$

where $\#S$ is the number of nonzero components in $\mathbf{s}^0$, and this implies

$$\lambda_2 = \frac{T}{\#S}$$

where $\#S$ must be nonzero. But this means that $\frac{1}{T}\sum s(t)^4 = \lambda_2 = (\#S)^{-1}$ is maximised when just one of the components of $\mathbf{s}^0$ is nonzero (where we neglected terms of lower order). This gives the solutions in (8). This proves the proposition.

Proposition 2 *Under the conditions of Proposition 1, also the absolute value of kurtosis is maximised by vectors of the form (8), for all practical sample sizes, T.*

Proof. It is easy to see that kurtosis is minimised by a vector $\mathbf{s}^0$, where half of the components are equal to 1 and other half of the components equal to -1. This vector has kurtosis equal to -2. On the other hand, the maximum (positive) kurtosis given by $\mathbf{s}^*$ in (8) equals $T-3$. Thus, for $T > 5$, we have $|T-3| > |-2|$, and even the absolute value of kurtosis is maximised by (8), still neglecting terms of lower order.

Proposition 3 *Let $\mathbf{x}(t)$ be Gaussian i.i.d. signal. Let $s(t)$ be a spike generated from $\mathbf{x}(t)$ by $s(t) = \mathbf{x}_0^T \mathbf{x}(t)$, where $\mathbf{x}_0 = \frac{\mathbf{x}(t_0)}{||\mathbf{x}(t_0)||}$ is the unmixing vector and t_0 is the time instance at which the spike appear. Then the kurtosis of the spike $s(t)$ is approximately*

$$kurt(s(t)) \approx \frac{||\mathbf{x}(t_0)||^4}{T} \propto \frac{1}{T},$$

Proof The kurtosis is defined $\mathrm{kurt}(s(t)) = E\left\{s(t)^4\right\} - 3E^2\left\{s(t)^2\right\}$. Here we have

$$\begin{aligned}
\mathrm{kurt}(s(t)) =& \frac{\sum_{t=1,t\neq t_0} s(t)^4 + s(t_0)^4}{T} - 3\left\{\frac{\sum_{t=1,t\neq t_0} s(t)^2 + s(t_0)^2}{T}\right\}^2 \\
=& \frac{\sum_{t=1,t\neq t_0} s(t)^4}{T} - 3\left\{\frac{\sum_{t=1,t\neq t_0} s(t)^2}{T}\right\}^2 \\
&+ \frac{s(t_0)^4}{T} - \frac{3}{T^2}\left(2\sum_{t=1,t\neq t_0} s(t)^2 s(t_0)^2 + s(t_0)^4\right).
\end{aligned} \tag{12}$$

Remembering that $\frac{1}{T}\sum_t s(t)^2 = 1$, we have $\sum_{t\neq t_0} s(t)^2 s(t_0)^2 = (T - s(t_0)^2)s(t_0)^2$. If T is large, $T \approx T-1$ and $1/T^2$ is negligible compared to $1/T$. Hence we have

$$\mathrm{kurt}(s(t)) \approx kurt(s(t)_{t\neq t_0}) + \frac{s(t_0)^4 - 6s(t_0)^2}{T} + o(\frac{1}{T^2}),$$

where $o(\frac{1}{T^2})$ is the last terms in (12) negligible for big values of T. The first term is roughly zero, because $\mathbf{x}(t)$ are independent in time and $s(t)_{t\neq t_0}$ thus roughly Gaussian.

Still remembering $s(t_0) = \frac{\mathbf{x}(t_0)}{||\mathbf{x}(t_0)||}^T \mathbf{x}(t_0)$ we finally get

$$\mathrm{kurt}(s(t)) \approx \frac{||\mathbf{x}(t_0)||^4 - 6\,||\mathbf{x}(t_0)||^2}{T} \propto \frac{1}{T}. \tag{13}$$

Proposition 4 *Let $s(t)$ be a spike as in Proposition 3. The kurtosis, in function of the number of dimensions m is*

$$kurt(s(t)) \approx \frac{||\mathbf{x}(t_0)||^4 - 6\,||\mathbf{x}(t_0)||^2}{T} = \frac{m^2 - 6m}{T} \propto m^2,$$

for large values of m.

Proof Starting from (13) we have

$$\begin{aligned}\mathrm{kurt}(s(t)) &\approx \frac{||\mathbf{x}(t_0)||^4}{T} = \frac{\left(\sum_{i=1}^m x_i(t_0)^2\right)^2 - 6\sum_{i=1}^m x_i(t_0)^2}{T} \\ &= \left[\left\{m\left(\mathrm{var}(x(t_0)) + E^2\{x\}\right)\right\}^2 - 6m\left\{\mathrm{var}(x(t_0)) + E^2\{x\}\right\}\right]/T\end{aligned}$$

Since $x \sim \mathcal{N}(0,1)$ we finally get

$$\mathrm{kurt}(s(t)) \approx \frac{m^2 - 6m}{T} \propto m^2,$$

for large values of m.

References

S.-I. Amari. Estimating functions of independent component analysis for temporally correlated signals. *Neural Computation*, 12:2083–2107, 2000.

H. Attias. ICA, graphical models and variational methods. In S. Roberts and R. Everson, editors, *Independent Component Analysis: Principles and Practice*, pages 95–112. Cambridge University Press, 2001.

P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality. *Physical review A*, 38(1):364–374, 1988.

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

A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–44, 1997.

J.-F. Cardoso. Source separation using higher order moments. In *Proc. ICASSP'89*, pages 2109–2112, 1989.

Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140:362 – 370, 1993.

Kwokleung Chan, Te-Won Lee, and Terrence Sejnowski. Variational learning of clusters of undercomplete nonsymmetric independent components. In *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, pages 492–497, San Diego, USA, 2001.

R.A. Choudrey and S.J. Roberts. Flexible Bayesian independent component analysis for blind source separation. In *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, pages 90–95, San Diego, USA, 2001.

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

FastICA. The FastICA MATLAB package. Available at `http://www.cis.hut.fi/projects/ica/fastica/`, 1998.

M. Funaro, E. Oja, and H. Valpola. Artefact detection in astrophysical image data using independent component analysis. In *Proc. of 3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, pages 43–48, San Diego, California, USA, December 9-12 2001.

M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65:413–497, 1993.

P. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191-210, 2000.

J. Hurri, A. Hyvärinen, J. Karhunen, and E. Oja. Image feature extraction using independent component analysis. In *Proc. NORSIG'96*, pages 475–478, Espoo, Finland, 1996.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

A. Hyvärinen. Complexity pursuit: Separating interesting components from time-series. *Neural Computation*, 13(4):883–898, 2001.

A. Hyvärinen, P. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1525 – 1558, 2001.

A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In M.I. Jordan, M.J. Kearns and S.A. Solla, eds., *Advances in Neural Information Processing 10*, pages 273–279. MIT Press, 1998.

A. Hyvärinen. Sparse code shrinkage: Denoising by maximum likelihood estimation. *Neural Computation*, 12(3):429 – 439, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley, 1st edition, 2001.

A. Hyvärinen, J. Särelä, and R. Vigário. Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, pages 425–429, Aussois, France, 1999.

O. Jahn and A. Cichocki. Identification and elimination of artifacts from MEG signals using efficient independent components analysis. In *Proc. of the 11th Int. Conf. on Biomagnetism (BIOMAG-98)*, Sendai, Japan, 1998.

T-P Jung, S. Makeig, T-W. Lee, M.J. McKeown, G. Brown, A. Bell, and T.J. Sejnowski. Independent component analysis of biomedical signals. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation*, pages 633 – 644, Helsinki, Finland, 2000.

C. Jutten. *Calcul neuromimétique et traitement du signal, analyse en composantes indépendentes*. PhD thesis, INPG, Univ. Grenoble, 1987. (in French).

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

H. Lappalainen. Ensemble learning for independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 7–12, Aussois, France, 1999.

Scott Makeig, Anthony Bell, Tzyy-Ping Jung, and Terrence Sejnowski. Independent component analysis of electroencephalographic data. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Neural Information Processing Systems 8*, pages 145–151, Cambridge MA, 1996. MIT Press.

F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. A resampling approach to estimate the stability of one- and multidimensional independent components. *IEEE Trans. Biom. Eng.*, 49 (12):1514 – 1525, 2002.

J. Miskin and D. MacKay. Ensemble Learning for blind source separation. In S. Roberts and R. Everson, editors, *Independent Component Analysis: Principles and Practice*, pages 209–233. Cambridge University Press, 2001.

K.-R. Müller, R. Vigário, F. Meinecke, and A. Ziehe. Blind source separation techniques for decomposing evoked brain signals. *International Journal of Bifurcation and Chaos*, 2003. in press.

A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd edition, 1991.

L. Parra, C. Spence, P. Sajda, A. Ziehe, and K.-R. Müller. Unmixing hyperspectral data. In T. Leen, T. Dietterich and V. Tresp, eds., *Advances in Neural Information Processing Systems 12*, pages 942 – 948. MIT Press, 2000.

Jaakko Särelä and Ricardo Vigário. A bayesian approach to overlearning in ICA: a comparison study. *Submitted for publication*, 2003.

A. Tang, B. Pearlmutter, N. Malaszenko, D. Phung, and B. Reeb. Independent components of magnetoencephalography: Localization. *Neural Computation*, 14:1827 – 1858, 2002.

Kari Torkkola. Blind separation for audio signals: are we there yet? In *Proc. Int. Workshop on Independent Component Analysis and Blind Separation of Signals (ICA'99)*, pages 239–244, Aussois, France, 1999.

Harri Valpola and Jaakko Särelä. Denoising source separation algorithms. *Submitted for publication*, 2003.

R. Vigário. Extraction of ocular artifacts from EEG using independent component analysis. *Electroencephalography and Clinical Neurophysiology*, 103(3):395–404, 1997.

R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, 47 (5):589–593, 2000.

R.N. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographics recordings. In M.I. Jordan, M.J. Kearns and S.A. Solla, eds., *Advances in Neural Information Processing 10*, pages 229–235. MIT Press, Cambridge MA, 1997.

Pascal Yiou, Didier Sornette, and Michael Ghil. Data-adaptive wavelets and multi-scale singular-spectrum analysis. *Physica D*, 142:254 – 290, 2000.

A. Ziehe and K.-R. Müller. TDSEP — an effective algorithm for blind separation using time structure. In *Proc. Int. Conf. on Neural Networks (ICANN'98)*, pages 675–680, Skövde, Sweden, 1998.