Robust Parallel Fast-ICA Algorithms Using Batch and Adaptive MMSE Estimators

⋆Francisco Messina† and Bruno Cernuschi-Frías†‡

†Facultad de Ingeniería, Universidad de Buenos Aires, Argentina
‡IAM-CONICET, Buenos Aires, Argentina

Abstract. All the algorithms for ICA require high-order statistics to estimate the independent components. This is because second-order information is insufficient to assess that two random variables are independent of each other. It is known that the robustness of the high-order sample estimators is poor, meaning that a few outliers can change dramatically its value. In this paper, we generalize the alternative robust statistics for moments and cumulants introduced by Welling [1] presenting the MMSE-robust moments. Then we present a batch and adaptive versions of an algorithm for estimating the parameters that define the estimator. Finally, we modify two FastICA algorithms of ICA based on kurtosis and negentropy to apply the MMSE robust estimators and show some experiments with supergaussian sources to demonstrate the improvement.

1 Introduction

Simple statistics such as moments and cumulants have been used extensively to model data. The problem with the classical sample estimators for moments is that a few outliers can change the estimation completely, with the issue becoming more important as the order of the statistic increases. Since cumulants are functions of moments up to the same order, they also suffer from high sensitivity to outliers. For supergaussian distributions, that are peaky and heavy tailed, the problem is of particular importance.

Thus, it is important to be able to find more stable ways to estimate these statistics.

The paper is divided as follows. Section 2 presents a short review of the classical definitions of moments and cumulants for a random variable. Then we generalize these definitions to introduce what we call robust-MMSE moments and two algorithms are stated for the estimation of its parameters. Afterwards, in section 3 a review of the ICA model is presented with the conditions that need to be satisfied in order to make the estimation possible. Followed by that, in sections 4 and 5 the two usual criteria for nongaussianity maximization are presented. Finally, we present a few simple experiments in section 6 and the conclusions of the work in section 7.

⋆ This work was partially supported by the University of Buenos Aires, CONICET and ANPCyT.

2 Classical Moments and Robust-MMSE Moments Definitions

Recall that for a random variable X the moments μ_n are defined by $\mu_n = \mathbb{E}[X^n]$ and the cumulants κ_n by the Maclaurin series expansion of the second characteristic function $\psi_X(\omega)$ which is the logarithm of the first characteristic function $\phi_X(\omega)$ [2]

$$\psi_X(\omega) = log\left(\phi_X(\omega)\right) = \sum_{n=0}^{\infty} \frac{1}{n!}\kappa_n(i\omega)^n. \tag{1}$$

The idea of Welling [1] is to introduce an isotropic decay exponential factor which downweights outliers. This in turn implies preference for some location and scale so that it is necessary to assume the following fact: the random variable is zero-mean and unit-variance. In the ICA problem this is equivalent to assume that the data has already been sphered which is a typical preprocessing step.

We will generalize the exponential factor by replacing it by an arbitrary function that controls the robustness of the estimator. In addition, we add two parameters to minimize its MSE as follows

Definition 1. *The robust-MMSE moments are given by*

$$\mu_n^{(\alpha_n,\beta_n,g_n(.))} = \mathbb{E}[\alpha_n X^n g_n(X) + \beta_n], \tag{2}$$

where α_n and β_n are chosen as the parameters that minimize the mean square error between $\alpha_n X^n g_n(X) + \beta_n$ and X^n, while $g_n(.)$ controls its robustness.

Clearly, with this definition, there is no need to assume that X is zero-mean and unit variance because $g_n(.)$ is arbitrary. In this section, we relax this condition.

Thus, our generalization consists of taking some robust estimator and then apply to it an affine transformation that yields in theory minimum MSE.

Note that the classical moments can be recovered from this definition by $\mu_n^{(1,0,g_n(.)=1)} = \mu_n$ which is useful as a limit case that can still be used, for example, for random variables whose density is known to have finite support. Alternatively, for a random variable with a very heavy tailed distribution or in a situation where there are a considerable proportion of outliers we should use an appropiate decaying function in order to attenuate their effect. The advantage of this definition is that it allows to use different functions for different situations and different orders of the moment to be estimated.

Lets now obtain an algorithm to estimate α_n and β_n. In the following derivation we assume that the function $g_n(.)$ is fixed and known. Actually, as we mentioned, it controls the robustness of the estimator and should be chosen or estimated from the samples with some appropiate criterion. In section 6 we will present two alternatives. The MSE can be calculated as follows

$$MSE(\alpha_n, \beta_n) = \mathbb{E}[(\alpha_n X^n g_n(X) + \beta_n - X^n)^2]. \tag{3}$$

Then,

$$\begin{aligned} MSE(\alpha_n, \beta_n) &= \mathbb{E}[(X^n(\alpha_n g_n(X) - 1) + \beta_n)^2] \\ &= \alpha_n^2 \mathbb{E}\big[X^{2n} g_n^2(X)\big] + \beta_n^2 + 2\alpha_n \beta_n \mathbb{E}[X^n g_n(X)] \\ &- 2\alpha_n \mathbb{E}\big[X^{2n} g_n(X)\big] - 2\beta_n \mathbb{E}[X^n] + \mathbb{E}\big[X^{2n}\big]. \end{aligned} \tag{4}$$

Now, to minimize this function with respect to α_n and β_n we take the partial derivatives and set them equal to zero which yields,

$$\begin{aligned} \frac{\partial MSE(\alpha_n, \beta_n)}{\partial \alpha_n}\bigg|_{\alpha_n = \alpha_n^{opt}} &= 2\alpha_n \mathbb{E}\big[X^{2n} g_n^2(X)\big] + 2\beta_n \mathbb{E}[X^n g_n(X)] \\ &- 2\mathbb{E}\big[X^{2n} g_n(X)\big]\bigg|_{\alpha_n = \alpha_n^{opt}} = 0, \end{aligned} \tag{5}$$

$$\frac{\partial MSE(\alpha_n, \beta_n)}{\partial \beta_n}\bigg|_{\beta_n = \beta_n^{opt}} = 2\beta_n + 2\alpha_n \mathbb{E}[X^n g_n(X)] - 2\mathbb{E}[X^n]\bigg|_{\beta_n = \beta_n^{opt}} = 0. \tag{6}$$

From these equations we find

$$\alpha_n^{opt} = \frac{\mu_{2n}^{(1,0,g_n(.))} - \mu_n \mu_n^{(1,0,g_n(.))}}{\mu_{2n}^{(1,0,g_n^2(.))} - \left(\mu_n^{(1,0,g_n(.))}\right)^2}, \tag{7}$$

$$\beta_n^{opt} = \mu_n - \alpha_n^{opt} \mu_n^{(1,0,g_n(.))}. \tag{8}$$

Note that these equations show that the estimator of μ_n by the robust-MMSE moment $\mu_n^{\left(\alpha_n^{opt}, \beta_n^{opt}, g_n(.)\right)}$ is unbiased and, therefore, a minimum variance unbiased estimator.

These solutions depend on both the unknown moments μ_n and the expectations $\mathbb{E}[X^n g_n(X)]$, $\mathbb{E}\big[X^{2n} g_n(X)\big]$ and $\mathbb{E}\big[X^{2n} g_n^2(X)\big]$. We propose a simple approximation where the first quantity is replaced by the sample estimation of $\alpha_n^{(0)} \mathbb{E}[X^n g_n(X)] + \beta_n^{(0)}$, where $\alpha_n^{(0)}$ and $\beta_n^{(0)}$ are some initial parameters, and the others by its sample estimation. Note that since the function $g_n(.)$ is chosen to ensure robustness this approximation is reasonable.

Algorithm 1. *Batch estimation of the values of α_n^{opt} and β_n^{opt}.*

1. *Find the sample estimators $\hat{\mu}_n^{(1,0,g_n(.))}$, $\hat{\mu}_{2n}^{(1,0,g_n(.))}$ and $\hat{\mu}_{2n}^{(1,0,g_n^2(.))}$.*
2. *Choose some initial values for the parameters $\alpha_n^{(0)}$ and $\beta_n^{(0)}$ (e.g. $\alpha_n^{(0)} = 1$ and $\beta_n^{(0)} = 0$).*

3. *Estimate the parameters by*

$$\hat{\alpha}_n^{opt} = \frac{\hat{\mu}_{2n}^{(1,0,g_n(.))} - \hat{\mu}_n^{(\alpha_n^0,\beta_n^0,g_n(.))}\hat{\mu}_n^{(1,0,g_n(.))}}{\hat{\mu}_{2n}^{(1,0,g_n^2(.))} - \left(\hat{\mu}_n^{(1,0,g_n(.))}\right)^2}, \tag{9}$$

$$\hat{\beta}_n^{opt} = \hat{\mu}_n^{(\alpha_n^0,\beta_n^0,g_n(.))} - \hat{\alpha}_n^{opt}\hat{\mu}_n^{(1,0,g_n(.))}. \tag{10}$$

This is a one-step batch algorithm. One evident issue is that the initial values for the parameters have an important influence in the final solutions. One possible solution is to run in parallel many algorithms with different initial conditions. It is also possible, and more interesting, to elaborate an online version of the algorithm as follows.

Algorithm 2. *Adaptive estimation of the values of* α_n^{opt} *and* β_n^{opt}.

1. *Choose any values for the initial estimates of the parameters* $\hat{\alpha}_n^{(0)}$ *and* $\hat{\beta}_n^{(0)}$.
2. *When the first* K *samples*[1]*, e.g.* $K = 100$*, are available, set* $p := K$ *and do the following assignments:*

$$\hat{\mu}_n^{(1,0,g_n(.)),(p)} = \sum_{j=1}^{K} x_j^n g_n(x_j), \tag{11}$$

and similarly for $\hat{\mu}_{2n}^{(1,0,g_n(.)),(p)}$ and $\hat{\mu}_{2n}^{(1,0,g_n^2(.)),(p)}$
Then, estimate the optimal parameters by

$$\hat{\alpha}_n^{(p)} = \frac{\hat{\mu}_{2n}^{(1,0,g_n(.)),(p)} - \hat{\mu}_n^{(\alpha_n^{p-1},\beta_n^{p-1},g_n(.)),(p)}\hat{\mu}_n^{(1,0,g_n(.)),(p)}}{\hat{\mu}_{2n}^{(1,0,g_n^2(.)),(p)} - \left(\hat{\mu}_n^{(1,0,g_n(.)),(p)}\right)^2}, \tag{12}$$

$$\hat{\beta}_n^{(p)} = \hat{\mu}_n^{(\alpha_n^{p-1},\beta_n^{p-1},g_n(.)),(p)} - \hat{\alpha}_n^{(p)}\hat{\mu}_n^{(1,0,g_n(.)),(p)}. \tag{13}$$

3. *For every new sample, say* x_{p+1}*, set* $p := p + 1$ *and update the estimations by*

$$\hat{\mu}_n^{(1,0,g_n(.)),(p)} = \frac{p-1}{p}\hat{\mu}_n^{(1,0,g_n(.)),(p-1)} + \frac{1}{p}x_p^n g_n(x_p), \tag{14}$$

$$\hat{\mu}_{2n}^{(1,0,g_n(.)),(p)} = \frac{p-1}{p}\hat{\mu}_{2n}^{(1,0,g_n(.)),(p-1)} + \frac{1}{p}x_p^{2n} g_n(x_p), \tag{15}$$

$$\hat{\mu}_{2n}^{(1,0,g_n^2(.)),(p)} = \frac{p-1}{p}\hat{\mu}_{2n}^{(1,0,g_n^2(.)),(p-1)} + \frac{1}{p}x_p^{2n} g_n^2(x_p), \tag{16}$$

Then, reestimate the optimal parameters by (12) *and* (13) .

[1] We cannot initialize the estimation of the parameters with one sample, see the denominator of (7)

It can be seen immediately that this online algorithm has a constant computational cost per iteration. Moreover, it is still useful for a nonstationary environment since the adaptive approach allows the tracking of the statistics. These are two extremely desirable properties in practice.

It is easy to perform an analysis of the convergence of the algorithm for the case of stationary environments. Since this is a stochastic on-line algorithm, the analysis will be based on the averaged differential equations of the update rules [3]. For example, for the equation (14), we obtain

$$\frac{d\hat{\mu}_n^{(1,0,g_n(.))}}{dt} = \frac{-\hat{\mu}_n^{(1,0,g_n(.))} + \mathbb{E}[X^n g_n(X)]}{t} \tag{17}$$

The fixed point of this equation is $\hat{\mu}_n^{(1,0,g_n(.))} = \mathbb{E}[X^n g_n(X)]$. The analysis is exactly the same for the other update rules of the robust moments. This guarantees the stability of $\hat{\alpha}_n$ but, unfortunately, not much can be said about $\hat{\beta}_n$.

Now that we can estimate moments and cumulants in a robust, fast and presumably accurate way, we will apply them to the estimation of ICs.

3 Independent Component Analysis

The ICA instantaneous noiseless model assumes that the data vector $\mathbf{X} \in \mathbb{R}^M$ is a linear mixture of some latent (unobserved) vector $\mathbf{S} \in \mathbb{R}^N$ whose components are independent random variables

$$\mathbf{X} = \boldsymbol{A}\mathbf{S}. \tag{18}$$

There are some restrictions that are needed to guarantee identifiability (i.e. that it is possible to estimate the ICs up to some trivial indeterminacies). Basically, we need to impose that [4,5] :

1. *The independent components S_j are statistically independent and no more than one of them is Gaussian.*
2. *For simplicity, the unknown matrix is assumed to be square (i.e. $M = N$).*

The first condition is the fundamental hypothesis of ICA. On the other hand, the second condition can be relaxed in some cases [6]. Note that since we assume that the unknown matrix is square, we can also assume that it is invertible if we consider that redundant mixtures are discarded. Then, after estimating the mixing matrix $\boldsymbol{A}$, the ICs can be obtained simply by

$$\mathbf{S} = \boldsymbol{A}^{-1}\mathbf{X}. \tag{19}$$

This means that we can use linear estimators on $\mathbf{X}$ for the independent components. Specifically,

$$\hat{\mathbf{S}}_j = \mathbf{w}_j^T \mathbf{X}, \tag{20}$$

where $\mathbf{w}_j$ must be found by maximization of an independence measure between all the estimates [7]. One possible method to solve the problem is to seek the $\mathbf{w}_j$ as the directions that maximize locally nongaussianity, a measure motivated intuitively by the central limit theorem. Typically, this is done by kurtosis or negentropy approximations [7].

4 Robust Kurtosis Algorithms

Kurtosis for a zero-mean unit-variance random variable is simply given by

$$\kappa_4 = \mu_4 - 3. \tag{21}$$

For Gaussian random variables the kurtosis is zero, but the converse is not true. Kurtosis can be either positive or negative corresponding to supergaussian and subgaussian distributions, respectively. Thus, for measuring nongaussianity the absolute value of kurtosis could be taken. For whitened data $\mathbf{Z}$, the condition on zero-mean and unit-variance gives the following constraint for the weight vectors

$$\mathbb{E}\left[\mathbf{w}_j^T \mathbf{Z}\mathbf{Z}^T \mathbf{w}_j\right] = \|\mathbf{w}_j\|_2^2 = 1 \qquad \forall j \in \{1, \ldots, M\}. \tag{22}$$

Thus, the problem for finding one IC can be formulated as

$$\begin{aligned} \hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} J(\mathbf{w}) = \arg\max_{\mathbf{w}} |\kappa_4(\mathbf{w}^T \mathbf{Z})| \\ &= \arg\max_{\mathbf{w}} |\mathbb{E}\left[(\mathbf{w}^T \mathbf{Z})^4\right] - 3\mathbb{E}\left[(\mathbf{w}^T \mathbf{Z})^2\right]| \\ &\text{s.t.} \|\mathbf{w}\|_2^2 = 1. \end{aligned} \tag{23}$$

The problem can be solved with a fixed-point iteration algorithm that has the advantages of avoiding the necessity of selecting a learning-rate sequence and gives faster convergence [6]. Using the method of Lagrange multipliers it is easily seen that at a stable point of the algorithm, the gradient must point in the direction of $\mathbf{w}$. So the fixed-point algorithm obtained is:

$$\mathbf{w} := \mathbb{E}\left[\mathbf{Z}(\mathbf{w}^T \mathbf{Z})^3\right] - 3\mathbf{w}, \tag{24}$$

$$\mathbf{w} := \frac{\mathbf{w}}{\|\mathbf{w}\|_2}. \tag{25}$$

The final vector gives one of the ICs as the projection of the data in its direction. This algorithm is called FastICA. It can be shown that the convergence of this algorithm is cubic [6]. Moreover, there are no adjustable parameters, which makes it easier to use than gradient ascent, and more reliable.

For the estimation of more than one IC, we first note the following important fact: *the directions of the ICs are orthogonal when the data is sphered.* This can be seen by direct calculation. In fact, suppose we have found the directions for two ICs $\mathbf{w}_i$ and $\mathbf{w}_j$. Then

$$\mathbb{E}\left[\mathbf{w}_i^T \mathbf{Z}\mathbf{Z}^T \mathbf{w}_j\right] = \mathbf{w}_i^T \mathbf{w}_j = 0. \tag{26}$$

So the key to obtain the different ICs is to search in the orthogonal space of the already found independent directions. This can be made by deflationary orthogonalization using the Gram-Schimidt method or by a symmetric orthogonalization in which the directions of the independent components are estimated in parallel [6]. The latter option is very advanteagous in practice, being its more remarkable benefits: i) there is no accumulation of round-off errors; ii)parallel computations can be made making possible to take advantage of any parallel architecture, so the algorithm converges much faster.

The symmetric orthogonalization of $\boldsymbol{W}$ can be achieved by the classic method involving matrix square roots

$$\boldsymbol{W} := (\boldsymbol{W}\boldsymbol{W}^T)^{-1/2}\boldsymbol{W}. \tag{27}$$

which yields obviously an orthogonal matrix. Then the algorithm takes the following form

Algorithm 3. *Symmetric orthogonalization algorithm:*

1. Choose M, the number of ICs to estimate and the tolerances for the weight vectors ϵ.
2. Initialize $\mathbf{w}_i, i = 1, \ldots, M$ (e.g. randomly) with unit 2-norm.
3. Do an iteration of the fixed-point algorithm on every $\mathbf{w}_i$ in parallel.
4. Do a symmetric orthogonalization of the matrix $\boldsymbol{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_M)$ by using equation (27).
5. If the 2-norm of all the weight vectors has changed less than ϵ, end. Otherwise, go back to step 3.

The proposed robust algorithm is obtained by using FastICA with symmetric orthogonalization and replacing moments by the estimators presented in section 2.

5 Robust Negentropy Algorithms

The second measure of nongaussianity that we will consider is negentropy. This quantity is closely related to the information-theoretic concept of differential entropy [8]. The entropy of a random variable is related to the information that the observation of the variable gives. The more unpredictable the variable is, the larger its entropy. The entropy H of a random vector $\mathbf{X} \in \mathbb{R}^M$ is defined as

$$H(\mathbf{X}) = \mathbb{E}\left[-\log(p_{\mathbf{X}}(\mathbf{X}))\right] = -\int_{\mathbb{R}^M} p_{\mathbf{X}}(\mathbf{x}) \log(p_{\mathbf{X}}(\mathbf{x}))d\mathbf{x}. \tag{28}$$

A fundamental result of information theory is that *a Gaussian variable has the largest entropy among all random variables of equal means and variances* [8]. This is why this quantity can be used as a measure for nongaussianity.

Negentropy for a random vector $\mathbf{X}$ is now defined by

$$J(\mathbf{X}) = H(\mathbf{X}_G) - H(\mathbf{X}), \tag{29}$$

where $\mathbf{X}_G$ is a Gaussian random vector with the same mean and covariance matrix of $\mathbf{X}$. Note that this measure is zero only for a Gaussian variable and always nonnegative because of the property just mentioned. Maximization of negentropy is equivalent to minimization of entropy so that the problem is to find the directions of locally minimum entropy that corresponds to the ICs. Another advantage is that negentropy is well justified as a measure for nongaussinity by statistical theory, being optimal in some sense [6].

Observe that to calculate exactly the quantity we should know the probability density function of the variable under consideration, which is not the case here. We will then use some approximations for the evaluation of this quantity.

The classic method of approximating negentropy is based on an expansion of the pdf in the vicinity of a Gaussian density using high-order cumulants. Commonly the Gram-Charlier and the Edgeworth expansions are used. They lead to very similar approximations, but the Edgeworth expansion is preferred because it is a true asymptotic expansion [9]. Using a Edgeworth expansion one can find the following approximation for negentropy of a random variable X [4]

$$J(X) \approx \frac{1}{12}\kappa_3(X)^2 + \frac{1}{48}\kappa_4(X)^2 + \frac{7}{48}\kappa_3(X)^4 - \frac{1}{8}\kappa_3(X)^2\kappa_4(X). \tag{30}$$

Clearly, when the density is symmetric, this approximation reduces to the same criterion as the maximization of kurtosis. In the other case, there is more information present in the negentropy approximation.

So we have to solve the following optimization problem

$$\begin{aligned}\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} J(\mathbf{w}) \\ &= \arg\max_{\mathbf{w}} \frac{1}{12}\kappa_3(\mathbf{w}^T\mathbf{Z})^2 + \frac{1}{48}\kappa_4(\mathbf{w}^T\mathbf{Z})^2 + \frac{7}{48}\kappa_3(\mathbf{w}^T\mathbf{Z})^4 - \frac{1}{8}\kappa_3(\mathbf{w}^T\mathbf{Z})^2\kappa_4(\mathbf{w}^T\mathbf{Z}) \\ &\quad \text{s.t.} \|\mathbf{w}\|_2^2 = 1.\end{aligned} \tag{31}$$

In this case, the fixed-point algorithm is

$$\begin{aligned}\mathbf{w} =& \mathbb{E}\left[\mathbf{Z}(\mathbf{w}^T\mathbf{Z})^2\right]\{\frac{1}{2}\mathbb{E}\left[(\mathbf{w}^T\mathbf{Z})^3\right] + \frac{7}{4}\mathbb{E}\left[(\mathbf{w}^T\mathbf{Z})^3\right]^3 - \frac{3}{4}\left(\mathbb{E}\left[(\mathbf{w}^T\mathbf{Z})^4\right] - 3\right)^2\} + \\ &+\mathbb{E}\left[\mathbf{Z}(\mathbf{w}^T\mathbf{Z})^3\right]\{\frac{1}{6}\left(\mathbb{E}\left[(\mathbf{w}^T\mathbf{Z})^4\right] - 3\right) - \mathbb{E}\left[(\mathbf{w}^T\mathbf{Z})^3\right]^2\}\end{aligned} \tag{32}$$

$$\mathbf{w} := \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \tag{33}$$

For estimating various ICs we can use this update rule in the Algorithm 2.

6 Experiments

We focus the experiments in testing the performance of the proposed robust-MMSE estimators for two different attenuation functions, shown in figure 1, and also to evaluate algorithms 1 and 2. For the simulations, we consider first an exponential distribution with $\lambda = \frac{1}{4}$ which is centered and scaled as a preprocessing step for simplicity. Nevertheless its ordinariness, this is a supergaussian asymmetric distribution which is useful to illustrate some results. We will use a set of 10,000 samples and add a 5% of outliers represented by samples of a normal distribution with mean 30 and unit variance. Then, we repeat the procedure with a less typical distribution, known as the hyperbolic secant distribution.

The results obtained for both algorithms are similar in the stationary case, but the first one depends stronger on the initial condition, as expected. They are shown for the exponential random variable in tables 1 and 2 and for the hyperbolic secant random variable in tables 3 and 4. The attenuation functions used seem to be fairly good choices.

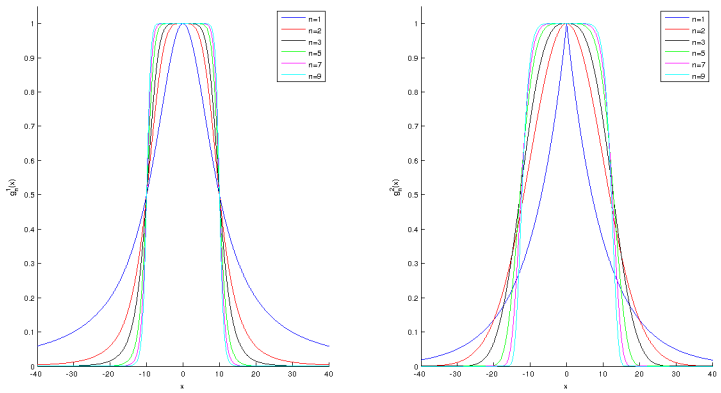

Fig. 1. Plot of $g_n^1(x) = \frac{1}{1+x^{2n}/10^{2n}}$ and $g_n^2(x) = e^{(\frac{-|x|^n}{n10^n})}$ for different values of n

In the simulation process, a lot of observations were made, being the more important ones,

1. The upgrade of the robust moments is remarkable even with high proportions of outliers.
2. The algorithm seems to work well for α_n but becomes sensible in its value for large n. Also, β_n is a parameter that varies a lot and becomes huge for large n in order to compensate for small errors in α_n. In the future, we propose to use regularization to solve this problem.

Table 1. Robust moment estimation of the exponential random variable using the attenuation family function $g_n^1(.)$

n	Sample moment	Robust moment	Theoretical moment	α_n	β_n
1	1.4295E+00	1.3100E-01	0.0000E+00	4.0961E+00	-4.0540E-01
2	4.4016E+01	1.4580E+00	1.0000E+00	4.4952E+01	-6.4083E+01
3	1.2895E+03	3.6993E+00	2.0000E+00	1.8867E+02	-6.9424E+02
4	3.8711E+04	1.5160E+01	9.0000E+00	4.4814E+02	-6.7787E+03
5	1.1648E+06	5.7812E+01	4.4000E+01	1.5015E+03	-8.6750E+04
6	3.5298E+07	2.8999E+02	2.6500E+02	2.4890E+03	-7.2148E+05
7	1.0416E+09	1.6605E+03	1.8540E+03	1.7037E+03	-2.8274E+06
8	3.2477E+10	9.3930E+03	1.4833E+04	4.7129E+03	-4.4259E+07
9	9.6471E+11	8.8853E+04	1.3350E+05	1.7484E+03	-1.5526E+08
10	2.9468E+13	1.7152E+05	1.3350E+06	5.5020E+04	-9.4371E+09

Table 2. Robust moment estimation of the exponential random variable using the attenuation family function $g_n^2(.)$

n	Sample moment	Robust moment	Theoretical moment	α_n	β_n
1	1.4290E+00	3.4700E-02	0.0000E+00	4.0033E+00	-1.0420E-01
2	4.3766E+01	1.4279E+00	1.0000E+00	4.0842E+01	-5.6890E+01
3	1.2910E+03	2.1702E+00	2.0000E+00	2.6098E+01	-5.4467E+01
4	3.8767E+04	7.9262E+00	9.0000E+00	1.2159E+00	-1.7112E+00
5	1.1801E+06	2.8460E+01	4.4000E+01	1.0149E+00	-4.2270E-01
6	3.5518E+07	2.6086E+02	2.6500E+02	1.0860E+00	-2.2430E+01
7	1.0617E+09	1.6529E+03	1.8540E+03	1.0977E+00	-1.6145E+02
8	3.1863E+10	7.9707E+03	1.4833E+04	1.0248E+00	-1.9783E+02
9	9.6192E+11	1.6126E+05	1.3350E+05	1.2196E+00	-3.5406E+04
10	3.0349E+13	1.7427E+05	1.3350E+06	1.0022E+00	-3.8394E+02

Table 3. Robust moment estimation of the hyperbolic secant random variable using the attenuation family function $g_n^1(.)$

n	Sample moment	Robust moment	Theoretical moment	α_n	β_n
1	1.4262E+00	1.4210E-01	0.0000E+00	4.0281E+00	-4.3040E-01
2	4.3805E+01	1.4652E+00	1.0000E+00	5.8375E+01	-8.4064E+01
3	1.2925E+03	1.8837E+00	0.0000E+00	3.7519E+02	-7.0486E+02
4	3.9502E+04	1.1598E+01	5.0000E+00	8.3657E+02	-9.6913E+03
5	1.1749E+06	2.0333E+01	0.0000E+00	2.5753E+04	-5.2360E+05
6	3.5310E+07	1.2032E+02	6.1000E+01	6.9756E+04	-8.3928E+06
7	1.0724E+09	2.6559E+02	0.0000E+00	1.4129E+05	-3.7526E+07
8	3.2385E+10	1.6493E+03	1.3850E+03	1.2864E+06	-2.1215E+09
9	9.7888E+11	-1.0190E+03	0.0000E+00	7.5774E+05	7.7217E+08
10	2.9472E+13	2.5630E+04	5.0521E+04	2.8333E+07	-7.2617E+11

Table 4. Robust moment estimation of the hyperbolic secant random variable using the attenuation family function $g_n^2(.)$

n	Sample moment	Robust moment	Theoretical moment	α_n	β_n
1	1.4196E+00	6.7800E-02	0.0000E+00	3.8272E+00	-1.9180E-01
2	4.3921E+01	1.4295E+00	1.0000E+00	5.5706E+01	-7.8204E+01
3	1.2926E+03	1.9080E-01	0.0000E+00	9.0536E+01	-1.7083E+01
4	3.8999E+04	4.7579E+00	5.0000E+00	1.4074E+00	-1.9385E+00
5	1.1761E+06	-3.0550E+00	0.0000E+00	1.0200E+00	6.1100E-02
6	3.5395E+07	5.0635E+01	6.1000E+01	1.0076E+00	-3.8530E-01
7	1.0655E+09	2.4461E+01	0.0000E+00	1.0022E+00	-5.4800E-02
8	3.2047E+10	1.3779E+03	1.3850E+03	1.0030E+00	-4.0701E+00
9	9.7465E+11	5.0088E+03	0.0000E+00	1.0042E+00	-2.1035E+01
10	2.9301E+13	8.9630E+03	5.0521E+04	1.0001E+00	-7.2969E-01

3. In many cases the final value of the estimation of α_n is close to one, so that may be a good choice for its initial condition.
4. The Gaussian function is not a good choice for $g_n(.)$ since it is not flat at all near the origin as the functions in figure 1.
5. The moments can be efficiently estimated up to high orders in this way, showing that there is a severe attenuation of the effect of the outliers.

7 Conclusions

Estimation of high-order statistics is a difficult problem that arises in many situations. In particular, in the ICA problem, as the algorithms need this information to separate the sources. In this paper, we have defined robust-MMSE moments to deal with this problem and presented two families of attenuation functions that can be useful for this purpose, especially when dealing with supergaussian random variables and in the presence of outliers. We then stated two algorithms to estimate the parameters that define the robust-MMSE moments. The first is a one-step batch algorithm, while the second is a more versatile adaptive algorithm. As we have seen, the results seem promising.

In the future, we will continue to study some approximations to independence measures where one can use higher-order statistics to test the robust-MMSE approximations defined herein. Then, it would be possible to use more information to separate the sources and thus it is expected that this would be done in a more reliable way.

References

1. Welling, M.: Robust higher order statistics. Proc. 10th Int. Workshop on Artificial Intelligence and Statistics, Barbados (Jan 2005) 405–412
2. Papoulis, A.: Probability, Random Variables and Stochastic Processes. 4th edn. McGraw-Hill Europe (2002)

3. Kushner, H. ; Clark, D.: Stochastic approximation methods for constrained and unconstrained systems. 1st edn. Springer-Verlag (1978)
4. Common, P.: Independent component analysis, a new concept? Signal Processing (Apr 1994) 287–314
5. Eriksson, J. ; Koivunen, V.: Identifiability, separability, and uniqueness of linear ica models. IEEE Signal Processing Letters **11**(7) (Jul 2004) 601–604
6. Hyvärinen, A., Oja, E., Karhunen, J.: Independent Component Analysis. 1st edn. Wiley (2001)
7. Hyvärinen, A.; Oja, E.: Independent component analysis: Algorithms and applications. Neural Networks **13**(4-5) (Jun 2000) 411–430
8. Cover, T. ; Thomas, J.: Elements of Information Theory. 2nd edn. Wiley (2006)
9. Blinnikov, S. ; Moessner, R.: Expansions for nearly gaussian distributions. Astron. Astrophys. Suppl. Ser. **130**(1) (May 1998) 193–205