ONE-UNIT CONTRAST FUNCTIONS FOR INDEPENDENT COMPONENT ANALYSIS: A STATISTICAL ANALYSIS

Aapo Hyvärinen

Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 2200, FIN-02015 HUT, Finland
Email: aapo.hyvarinen@hut.fi

ABSTRACT

The author introduced previously a large family of one-unit contrast functions to be used in independent component analysis (ICA). In this paper, the family is analyzed mathematically in the case of a finite sample. Two aspects of the estimators obtained using such contrast functions are considered: asymptotic variance, and robustness against outliers. An expression for the contrast function that minimizes the asymptotic variance is obtained as a function of the probability densities of the independent components. Combined with robustness considerations, these results provide strong arguments in favor of the use of contrast functions based on slowly growing functions, and against the use of kurtosis, which is the classical contrast function.

1. INTRODUCTION

Independent Component Analysis (ICA) [1] is a statistical signal processing technique whose main applications are blind source separation, blind deconvolution, and feature extraction. In the simplest form of ICA [2], one observes m scalar random variables $x_1, x_2, ..., x_m$ which are assumed to be linear combinations of n unknown independent components, or ICs, denoted by $s_1, s_2, ..., s_n$. These ICs s_i are assumed to be mutually *statistically independent*, and zero-mean. Arranging the observed variables x_j into a vector $\mathbf{x} = (x_1, x_2, ..., x_m)^T$ and the IC variables s_i into a vector $\mathbf{s}$, the linear relationship can be expressed as

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

Here, $\mathbf{A}$ is an unknown $m \times n$ matrix of full column rank, called the mixing matrix. The basic problem of ICA is then to estimate both the mixing matrix $\mathbf{A}$ and the realizations of the ICs s_i *using only observations of the mixtures* x_j.

Estimation of ICA requires the use of higher-order information, i.e., other information than that contained in the covariance matrix of $\mathbf{x}$. This higher-order information is usually incorporated in the estimation procedures by means of 'contrast' functions based on higher-order cumulants [2, 3]. However, little justification has been provided in the literature for the choice of using higher-order cumulants for the construction of the contrast functions. The main reason for their popularity seems to be that they are easy to analyze mathematically. No statistical or practical arguments in favor of cumulants have been put forth, except for the fact that they may be more resistant to Gaussian noise, because the higher-order cumulants of Gaussian noise vanish.

In this paper, we analyze mathematically a large family of one-unit contrast functions introduced in [4]. The asymptotic variance of the obtained estimators is evaluated, and it is shown that for super-Gaussian ICs, the asymptotic variance is minimized for contrast functions that grow much slower than the 4-th power inherent in the fourth-order cumulant or kurtosis. Furthermore, robustness against outliers also requires slowly growing contrast functions. As most ICs encountered in practice seem to be super-Gaussian, this means that kurtosis may be a rather inadequate contrast function in most cases. For neural learning rules, the results imply that better estimates are usually obtained using (anti-)Hebbian learning functions that are sigmoidal, or even go to zero at infinity. Simulations back up our theoretical arguments.

2. GENERAL ONE-UNIT CONTRAST FUNCTIONS

Consider a linear combination of the observed mixtures x_j, say $\mathbf{w}^T\mathbf{x}$, where the (weight) vector $\mathbf{w}$ is constrained so that $E\{(\mathbf{w}^T\mathbf{x})^2\} = 1$. Many ICA algorithms are based on finding the extrema of the square of the kurtosis $\text{kurt}^2(\mathbf{w}^T\mathbf{x}) = (E\{(\mathbf{w}^T\mathbf{x})^4\}-3)^2$ of such a linear combination [2, 3]. This can be motivated by information-theoretic arguments: the square of the kurtosis can be shown to approximate the negentropy of $\mathbf{w}^T\mathbf{x}$ [2]. Moreover, it can be proven that the square of the kurtosis of $\mathbf{w}^T\mathbf{x}$ is maximized exactly in the points where the linear combination equals, up to the sign, one of the ICs, i.e., $\mathbf{w}^T\mathbf{x} = \pm s_i$ for some i [3, 5].

This approach was generalized in [4, 6, 7], where it was shown that instead of kurtosis, practically any non-quadratic, well-behaving even function, say G, can be used to construct a contrast function for ICA. Such a general contrast function can be defined as

$$J_G(\mathbf{w}) = [E_\mathbf{x}\{G(\mathbf{w}^T\mathbf{x})\} - E_\nu\{G(\nu)\}]^2 \tag{2}$$

where ν is a standardized Gaussian variable. The second term in brackets is a normalization constant that makes J_G equal to zero if $\mathbf{w}^T\mathbf{x}$ has a Gaussian distribution. Clearly, J_G can be considered a generalization of the square of kurtosis, as for $G(u) = u^4$, J_G becomes simply the square of kurtosis of $\mathbf{w}^T\mathbf{x}$. It was shown in [6], using a generalization of the Gram-Charlier expansion, how J_G approximates the negentropy of $\mathbf{w}^T\mathbf{x}$ in the same way as the square of the kurtosis. Furthermore, it was shown in [7] that under weak assumptions,

J_G is locally maximized when $\mathbf{w}^T\mathbf{x} = \pm s_i$, i.e. when the linear combination equals one of the ICs. Therefore, J_G can be used as a contrast function for ICA in the same way as the square of the kurtosis. Note that for simplicity, we shall also refer to G as a contrast function.

Thus, we estimate one IC by solving the following optimization problem:

$$\widehat{\mathbf{w}} = \arg \max_{E\{(\mathbf{w}^T\mathbf{x})^2\}=1} J_G(\mathbf{w}) \tag{3}$$

where in practice the expectations are replaced by sample averages. Note that this boils down to maximizing or minimizing $E\{G(\mathbf{w}^T\mathbf{x})\}$, where the type of extrema searched for depends on the sign of $E_{\mathbf{x}}\{G(\mathbf{w}^T\mathbf{x})\} - E_\nu\{G(\nu)\}$. To estimate all the ICs, one needs only to find all the local solutions of this optimization problem. We shall not consider here in detail how to solve this optimization. Two simple methods are possible. First, one can use a simple gradient descent/ascent with a decreasing learning rate, as is considered in more detail in [7]. In that case it may be useful to first whiten (or sphere) the data, which simplifies the constraint to $\|\mathbf{w}\| = 1$. A second possibility is the fixed-point algorithm introduced for kurtosis in [8] and generalized for any G in [4]. However, the statistical properties of the estimator defined in (3) do not depend on the method of optimization.

In the following, we shall analyze two fundamental statistical properties of $\widehat{\mathbf{w}}$, which are asymptotic variance and robustness. Though in principle almost any non-quadratic even function G can be used, in practice the performance of different contrast functions may be very different due to limited sample sizes and deviations from the model (1). Therefore, some analysis is needed to provide guidelines on how to choose the function G to obtain a statistically adequate estimator.

3. ASYMPTOTIC VARIANCE

In practice, one usually has only a finite sample of N observations of the vector $\mathbf{x}$. Therefore, the expectations in the definition of J_G are in fact replaced by sample averages. This results in certain errors in the estimator $\widehat{\mathbf{w}}$, and it is desired to make these errors as small as possible. A classical measure of this error is asymptotic (co)variance, which means the limit of the covariance matrix of $\widehat{\mathbf{w}}\sqrt{N}$ as $N \to \infty$. This gives an approximation of the mean-square error of $\widehat{\mathbf{w}}$. Comparison of, say, the traces of the asymptotic variances of two estimators enables direct comparison of the accuracy of two estimators. One can solve analytically for the asymptotic variance of $\widehat{\mathbf{w}}$, obtaining the following theorem:

Theorem 1 *The trace of the asymptotic variance of $\widehat{\mathbf{w}}$ as defined in (3) for the estimation of the independent component s_i, equals*

$$V_G = C(\mathbf{A}) \frac{E\{g^2(s_i)\} - (E\{s_i g(s_i)\})^2}{(E\{s_i g(s_i) - g'(s_i)\})^2}, \tag{4}$$

where g is the derivative of G, and $C(\mathbf{A})$ is a constant that depends only on $\mathbf{A}$.

Proof: Making the change of variable $\mathbf{z} = \mathbf{A}^T\mathbf{w}$, the equation defining the optimal solutions $\hat{\mathbf{z}}$ becomes

$$\sum_t \mathbf{s}_t g(\hat{\mathbf{z}}^T\mathbf{s}_t) = \lambda \sum_t \mathbf{s}_t\mathbf{s}_t^T\hat{\mathbf{z}} \tag{5}$$

where $t = 1, .., T$ is the sample index, T is the sample size, and λ is a Lagrangian multiplier.. Without loss of generality, let us assume that $\hat{\mathbf{z}}$ is near the ideal solution $\mathbf{z} = (1, 0, 0, ...)$. Note that due to the constraint $E\{(\mathbf{w}^T\mathbf{x})^2\} = \|\mathbf{z}\|^2 = 1$, the variance of the first component of $\hat{\mathbf{z}}$, denoted by $\hat{z}_1$, is of a smaller order than the variance of the vector of other components, denoted by $\hat{\mathbf{z}}_{-1}$. Excluding the first component in (5), and making the first-order approximation $g(\hat{\mathbf{z}}^T\mathbf{s}) = g(s_1) + g'(s_1)\hat{\mathbf{z}}_{-1}^T\mathbf{s}_{-1}$, where also $\mathbf{s}_{-1}$ denotes $\mathbf{s}$ without its first component, one obtains after some simple manipulations

$$\frac{1}{\sqrt{T}}\sum_t \mathbf{s}_{-1}[g(s_1) - \lambda s_1] = \frac{1}{T}\sum_t \mathbf{s}_{-1}[-\mathbf{s}_{-1}^T g'(s_1) + \lambda\mathbf{s}_{-1}^T]\hat{\mathbf{z}}_{-1}\sqrt{T} \tag{6}$$

where the sample index t has been dropped for simplicity. Making the first-order approximation $\lambda = E\{s_1 g(s_1)\}$, one can write (6) in the form $u = v\hat{\mathbf{z}}_{-1}\sqrt{T}$ where v converges to the identity matrix multiplied by $E\{s_1 g(s_1)\} - E\{g'(s_1)\}$, and u converges to a variable that has a normal distribution of zero mean whose covariance matrix equals the identity matrix multiplied by $E\{g^2(s_1)\} - (E\{s_1 g(s_1)\})^2$. This implies the theorem, since $\hat{\mathbf{z}}_{-1} = \mathbf{B}\hat{\mathbf{w}}$, where $\mathbf{B}$ is the inverse of $\mathbf{A}^T$ without its first row.

Thus the comparison of the asymptotic variances of two estimators of the form in (3), but for two different contrast functions G, boils down to a comparison of the V_G's. In particular, one can use variational calculus to find a G that minimises V_G. Thus one obtains the following theorem:

Theorem 2 *The trace of the asymptotic variance of $\widehat{\mathbf{w}}$ is minimized when G is of the form*

$$G_{opt}(u) = c_1 \log f(u) + c_2 u^2 + c_3 \tag{7}$$

where f is the density function of s_i, and c_1, c_2, c_3 are arbitrary constants.

For simplicity, one can choose $G_{opt}(u) = \log f(u)$. Thus one sees that the optimal contrast function is the same as the one obtained for several units by the maximum likelihood approach [9], or the infomax approach [10]. Almost identical results have also been obtained in [11] for another multi-unit algorithm. Our results treat, however, the one-unit case instead of the multi-unit case, and are thus applicable to estimation of a subset of the ICs, and to blind deconvolution [7].

4. ROBUSTNESS

Another very desirable property of an estimator is robustness against outliers [12]. This means that single, highly erroneous observations do not have much influence on the estimator.

In this paper, we shall treat the question: How does the robustness of the estimator $\hat{\mathbf{w}}$ depend on the choice of the function G? Note that the robustness of $\hat{\mathbf{w}}$ depends also on the method of estimation used in constraining the variance of $\mathbf{w}^T\mathbf{x}$ to equal unity in (3). This is a problem independent of the choice of G. In the following, we assume that this constraint is implemented in a robust way. In particular, we assume that the data is sphered (whitened) in a robust manner, in which case the constraint reduces to $\|\mathbf{w}\| = 1$. Several robust estimators of the variance of $\mathbf{w}^T\mathbf{x}$ or of the covariance matrix of $\mathbf{x}$ are presented in the literature; see [12].

The robustness of the estimator $\hat{\mathbf{w}}$ in (3) can be analyzed using the theory of M-estimators [12]. Without going into technical details, the definition of an M-estimator can be formulated as follows: an estimator is called an M-estimator if it is defined as the solution $\hat{\theta}$ for θ of

$$E\{\psi(\mathbf{x}, \theta)\} = 0 \tag{8}$$

where $\mathbf{x}$ is a random vector and ψ is some function defining the estimator. The estimator $\hat{\mathbf{w}}$ in (3) is an M-estimator. To see this, define $\theta = (\mathbf{w}, \lambda)$, where λ is the Lagrangian multiplier associated with the constraint. Using the Kuhn-Tucker conditions, the estimator $\hat{\mathbf{w}}$ can then be formulated as the solution of equation (8) where $\psi = \psi_J$ is defined as follows (for sphered data):

$$\psi_J(\mathbf{x}, \theta) = \begin{pmatrix} \mathbf{x}g(\mathbf{w}^T\mathbf{x}) + c\lambda\mathbf{w} \\ \|\mathbf{w}\|^2 - 1 \end{pmatrix} \tag{9}$$

where $c = (E_{\mathbf{x}}\{G(\hat{\mathbf{w}}^T\mathbf{x})\} - E_\nu\{G(\nu)\})^{-1}$ is an irrelevant constant.

The analysis of robustness of an M-estimator is based on the concept of an infuence function, $\mathrm{IF}(\mathbf{x}, \hat{\theta})$. Intuitively speaking, the influence function measures the influence of single observations on the estimator. It would be desirable to have an influence function that is bounded as a function of $\mathbf{x}$, as this implies that even the influence of a far-away outlier is 'bounded', and cannot change the estimate too much. This requirement leads to one definition of robustness, which is called B-robustness. An estimator is called B-robust, if its influence function is bounded as a function of $\mathbf{x}$, i.e., $\sup_{\mathbf{x}} \|\mathrm{IF}(\mathbf{x}, \hat{\theta})\|$ is finite for every $\hat{\theta}$. Even if the influence function is not bounded, it should grow as slowly as possible when $\|\mathbf{x}\|$ grows, to reduce the distorting effect of outliers.

It can be shown [12] that the influence function of an M-estimator equals

$$\mathrm{IF}(\mathbf{x}, \hat{\theta}) = \mathbf{B}\psi(\mathbf{x}, \hat{\theta}) \tag{10}$$

where $\mathbf{B}$ is an irrelevant invertible matrix that does not depend on $\mathbf{x}$. On the other hand, using our definition of ψ_J, and denoting by $\gamma = \mathbf{w}^T\mathbf{x}/\|\mathbf{x}\|$ the cosine of the angle between $\mathbf{x}$ and $\mathbf{w}$, one obtains easily

$$\|\psi(\mathbf{x}, (\mathbf{w}, \lambda))\|^2 = C_1 \frac{1}{\gamma^2} h^2(\mathbf{w}^T\mathbf{x}) + C_2 h(\mathbf{w}^T\mathbf{x}) + C_3 \tag{11}$$

where C_1, C_2, C_3 are constants that do not depend on $\mathbf{x}$, and $h(u) = ug(u)$. Thus on sees that the robustness of $\hat{\mathbf{w}}$ essentially depends on the behavior of the function $h(u)$. The slower $h(u)$ grows, the more robust the estimator. However, the estimator cannot be really B-robust, because the γ in the denominator prevents the influence function from being bounded for all $\mathbf{x}$. In particular, outliers that are almost orthogonal to $\hat{\mathbf{w}}$, and have large norms, may still have a large influence on the estimator. These results are stated in the following theorem:

Theorem 3 *Assume that the data* $\mathbf{x}$ *is whitened (sphered) in a robust manner. Then the influence function of the estimator* $\hat{\mathbf{w}}$ *is never bounded for all* $\mathbf{x}$. *However, if* $h(u) = ug(u)$ *is bounded, the influence function is bounded in sets of the form* $\{\mathbf{x} \mid \hat{\mathbf{w}}^T\mathbf{x}/\|\mathbf{x}\| > \epsilon\}$ *for every* $\epsilon > 0$, *where* g *is the derivative of* G.

In particular, if one chooses *a contrast function* $G(u)$ *that is bounded*, h is also bounded, and $\hat{\mathbf{w}}$ is quite robust against outliers. If this is not possible, one should at least choose a contrast function $G(u)$ that does not grow very fast when $|u|$ grows. If, in contrast, $G(u)$ grows very fast when $|u|$ grows, the estimates depend mostly on a few observations far from the origin. This leads to highly non-robust estimators, which can be completely ruined by just a couple of bad outliers. This is the case, for example, when kurtosis is used as a contrast function, which is equivalent to using $\hat{\mathbf{w}}$ with $G(u) = u^4$.

5. CHOOSING THE CONTRAST FUNCTION IN PRACTICE

It is useful to analyze the implications of the theoretical results of the preceding sections by considering the following family of density functions:

$$f_\alpha(x) = C_1 \exp(C_2 |x|^\alpha) \tag{12}$$

where α is a positive constant, and C_1, C_2 are normalization constants that ensure that f_α is a probability density of unit variance. For different values of alpha, the densities in this family exhibit different shapes. For $.5 < \alpha < 2$, one obtains a sparse, super-Gaussian density (i.e. a density of positive kurtosis). For $\alpha = 2$, one obtains the Gaussian distribution, and for $\alpha > 2$, a sub-Gaussian density (i.e. a density of negative kurtosis). Thus the densities in this family can be used as examples of different non-Gaussian densities.

Using Theorem 2, one sees that in terms of asymptotic variance, an optimal contrast function for estimating an IC whose density function equals f_α, is of the form:

$$G_{opt}(u) = |u|^\alpha \tag{13}$$

where the arbitrary constants have been dropped for simplicity. This implies roughly that for super-Gaussian (resp. sub-Gaussian) densities, the optimal contrast function is a function that grows *slower than quadratically* (resp. *faster than quadratically*). Next, recall from Section 4 that if $G(u)$

grows fast with $|u|$, the estimator becomes highly non-robust against outliers. Taking also into account the fact that most ICs encountered in practice are super-Gaussian, one reaches the conclusion that as a general-purpose contrast function, one should choose a function G that resembles rather

$$G_{opt}(u) = |u|^{\alpha}, \text{where } \alpha < 2. \tag{14}$$

The problem with such contrast functions is, however, that they are not differentiable at 0 for $\alpha \leq 1$. Thus it is better to use approximating differentiable functions that have the same kind of qualitative behavior. Considering $\alpha = 1$, in which case one has a double exponential density, one could use instead the function $G_1(u) = \log\cosh a_1 u$ where $a_1 > 1$ is a moderately large constant. Note that the derivative of G_1 is then the familiar tanh function (for $a_1 = 1$). In the case of $\alpha < 1$, i.e. highly super-Gaussian ICs, one could approximate the behavior of G_{opt} for large u using a Gaussian function (with a minus sign): $G_2(u) = -\exp(-a_2 u^2/2)$ where a_2 is a constant. The derivative of this function is like a sigmoid for small values, but goes to 0 for larger values. Note that this function also fulfills the condition in Theorem 3, thus providing an estimator that is as robust as possible in this framework. We have found $a_1 = 2$ and $a_2 = 1$ to provide 'good' approximations of G_1 and G_2. Note that there is a trade-off between the precision of the approximation and the smoothness of the resulting objective function.

Thus, we reach the following general conclusion:

- a good general-purpose contrast function is $G(u) = \log\cosh a_1 u$, where $a_1 \geq 1$ is a constant.
- when the ICs are highly super-Gaussian, or when robustness is very important, $G(u) = -\exp(-a_2 u^2/2)$ with $a_2 \approx 1$ may be better.
- using kurtosis is justified only if the ICs are sub-Gaussian and there are no outliers.

In this paper, we have used purely statistical criteria for choosing the contrast function. One important criterion that is completely independent of statistical considerations is computational simplicity. For example, the calculation of the tanh function is rather slow in many environments. The convergence may be speeded up if one uses instead piecewise linear approximations of the derivatives of the contrast functions. In the case of $g(u) = \tanh(a_2 u)$, one may define g so that $g(u) = a_3 u$ for $|u| < 1/a_3$ and $g(u) = \text{sign}(u)$ otherwise, where $a_3 \geq 1$ is a constant. This amounts to using the so-called Huber function [12] as G.

6. SIMULATIONS

We performed simulations in which 3 different contrast functions were used to estimate one IC from a mixture of 4 i.i.d. ICs. The contrast functions used were kurtosis, and the two functions proposed in the preceding section:

log cosh (or G_1) and the Gaussian function (or G_2). The constants were set as suggested in the preceding section. We also used three different distributions of the ICs: uniform, double exponential (or Laplace), and the distribution of the third power of a Gaussian variable. The sample size was fixed at 1000 and the fixed-point algorithm in [4] was used to maximize the contrast function. The asymptotic mean absolute deviations (MAD) between the components of the obtained vectors and the correct solutions were estimated and averaged over 1000 runs for each combination of non-linearity and distribution of IC. MAD was used instead af variance because it is a more robust measure of deviation.

The results in the basic, noiseless case are depicted in Fig. 1. As one can see, the estimates using kurtosis were essentially worse for super-Gaussian ICs. Especially the strongly super-Gaussian IC (cube of Gaussian) was estimated considerably worse using kurtosis. Only for the sub-Gaussian IC, kurtosis was better than the other contrast functions. There was no clear difference between the performances of the contrast functions G_1 and G_2.

Next, the experiments were repeated with added Gaussian noise whose energy was 10% of the energy of the ICs. The results are shown in Fig. 2. This time, kurtosis did not perform better even in the case of the sub-Gaussian density. This result goes against the view that kurtosis would tolerate Gaussian noise well. Indeed, the theoretical arguments supporting that view neglect any finite-sample effects, and may thus have rather limited validity.

No outliers were added in these experiments. Experiments confirming the robustness of the non-linearities proposed in section 5 can be found in [4].

7. CONCLUSION

The problem of choosing the contrast function for ICA was treated. The behavior of a large family of contrast functions, which includes kurtosis as a special case, was analyzed. Combining the results on asymptotic variance and robustness against outliers, it was shown that the use of kurtosis is not justified on statistical grounds, except perhaps for sub-Gaussian independent components. Instead, contrast functions that grow slower than quadratically were found to be better approximations of the optimal ones in most cases. In neural learning rules, this leads, e.g., to the use of tanh-like sigmoids, or functions resembling the derivative of a Gaussian function.

8. REFERENCES

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

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

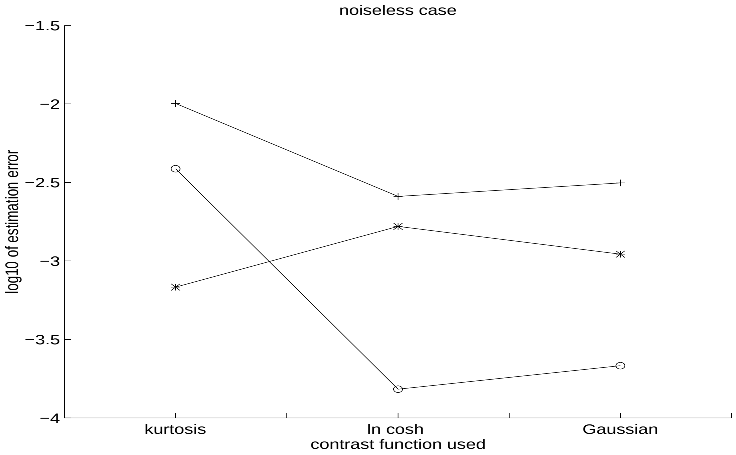

Figure 1: Estimation errors plotted for different contrast functions and distributions of the ICs, in the noiseless case. Asterisk: uniform distribution. Plus sign: Double exponential. Circle: cube of Gaussian.

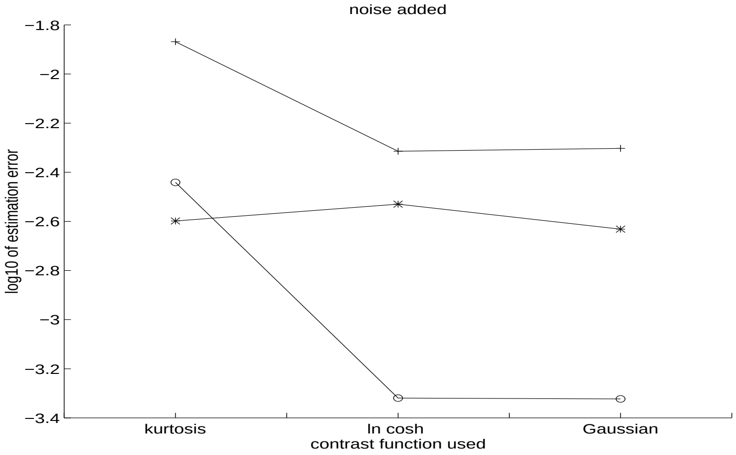

Figure 2: The noisy case. Estimation errors plotted for different contrast functions and distributions of the ICs. Asterisk: uniform distribution. Plus sign: Double exponential. Circle: cube of Gaussian.

[3] N. Delfosse and P. Loubaton, “Adaptive blind separation of independent sources: a deflation approach,” *Signal Processing*, vol. 45, pp. 59–83, 1995.

[4] A. Hyvärinen, “A family of fixed-point algorithms for independent component analysis,” in *Proc. ICASSP'97*, (Munich, Germany), pp. 3917–3920, 1997.

[5] A. Hyvärinen and E. Oja, “One-unit learning rules for independent component analysis,” in *Advances in Neural Information Processing Systems 9 (NIPS*96)*, MIT Press, 1997. To appear.

[6] A. Hyvärinen, “Approximations of differential entropy for independent component analysis and projection pursuit,” 1997. Submitted.

[7] A. Hyvärinen and E. Oja, “Independent component analysis by general non-linear hebbian-like learning rules,” 1997. Submitted.

[8] A. Hyvärinen and E. Oja, “A fast fixed-point algorithm for independent component analysis,” *Neural Computation*, 1997. To appear.

[9] D.-T. Pham, P. Garrat, and C. Jutten, “Separation of a mixture of independent sources through a maximum likelihood approach,” in *Proc. EUSIPCO*, pp. 771–774, 1992.

[10] A. Bell and T. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[11] J.-F. Cardoso and B. H. Laheld, “Equivariant adaptive source separation,” *IEEE Trans. on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.

[12] F. Hampel, E. Ronchetti, P. Rousseuw, and W. Stahel, *Robust Statistics*. Wiley, 1986.