Statistical Analysis of Sample-Size Effects in ICA

J. Michael Herrmann[1,2⋆] Fabian J. Theis[1,3⋆⋆]

[1] Bernstein Center for Computational Neuroscience Göttingen
[2] Göttingen University, Institute for Nonlinear Dynamics
[3] Max Planck Institute for Dynamics and Self-Organization
Bunsenstraße 10, 37073 Göttingen, Germany
michael@chaos.gwdg.de, fabian@theis.name

Abstract. Independent component analysis (ICA) solves the blind source separation problem by evaluating higher-order statistics, e.g. by estimating fourth-order moments. While estimation errors of the kurtosis can be shown to asymptotically decay with sample size according to a square-root law, they are subject to two further effects for finite samples. Firstly, errors in the estimation of kurtosis increase with the deviation from Gaussianity. Secondly, errors in kurtosis-based ICA algorithms increase when approaching the Gaussian case. These considerations allow us to derive a strict lower bound for the sample size to achieve a given separation quality, which we study analytically for a specific family of distributions and a particular algorithm (fastICA). We further provide results from simulations that support the relevance of the analytical results.

1 Introduction

Independent component analysis (ICA) aims at the extraction of source signals from measured multi-dimensional data [10, 5]. In order to identify a transformation of the data such that the resulting random vector is statistically independent, probabilistic algorithms for ICA rely on estimates of higher moments of the probability distribution functions of the observations [8]. While for first- and second-order moments many mathematical results (cf. e.g. [12]) are available describing the dependence of the estimation quality on the sample size or on the underlying parameters, the statistical properties of estimates of higher moments are less understood. A useful but typically loose characterization of the sample-size effects is given by the Tschebyscheff inequality, which assures a square-root decay of large deviations, provided that the existence of certain moments is guaranteed. Tighter estimates of higher moments are often designed as tests of normality, i.e. work best in the vicinity of the normal distribution. Probabilistic ICA algorithms are, however, based on the assumption that all but at most one sources are non-Gaussian [5, 15]. Thus on the one hand, ICA algorithms prefer source distributions most of which are sufficiently far from Gaussianity, while on the other hand estimates of higher-order moments such as the kurtosis are increasingly inefficient on deviations from Gaussianity [3].

⋆ present address: University of Edinburgh, Div. of Informatics, Edinburgh, Scotland, UK
⋆⋆ present address: Computational Modeling in Biology, IBI, GSF, Munich, Germany

Theoretical results in independent component analysis commonly focus on either showing existence and uniqueness of solutions [5, 15] or derive conditions that lead to considerable improvements in separation quality. For example, the learning rate can be bounded such as to achieve monotonous convergence [14] or can be modulated in order saturate the Cramér-Rao bound of the variance of the off-diagonal elements of the confusion matrix [11]. These conditions are formulated, however, in terms of certain expectation values of the contrast function, while the estimation errors for these quantities tend to be ignored such that essentially only asymptotic results are obtained. In the case of real-world data sets both the moments of the contrast function are unknown and the set of samples is small [6]. A serious study of data from natural images [4] reports a negligible gain in statistical independence by ICA when compared to a decorrelation of the set of image patches by principal component analysis. Reasons for this failure may be found in the questionable validity of the assumptions underlying the ICA algorithm.

The present study identifies several causes of ICA-based component estimation errors, namely sample-size, the statistical properties of the sources such as the distribution of kurtoses, and temporal correlations. We restrict the analysis to a single popular algorithm, fastICA [9], and to a specific family of distributions in order to obtain analytical results. The obtained results can be understood as counterexamples to assumptions which are also made in a larger class of algorithms and for other data distributions. We have chosen Pearson's second and seventh family [1, 7], see section 3.2 and Fig. 1. The choice of this example is suggestive: Zero mean and unit variance of the observations can be assumed without much loss of generality, since centering and whitening the data are non-critical in most cases. If additionally symmetry is assumed, then the kurtosis κ is the most relevant characteristic of the distribution function. κ is defined by the ratio of the fourth central moment $\mu_4 = E[(x-\mu)^4]$ and the squared variance σ^4, i.e. for a normal (Gaussian) distribution we have $\kappa = 3$. After appropriate reparameterization, Pearson's families depend on κ as the single parameter. The main advantage we are relying on here consists in the controlled access to a wider range of kurtoses than typically found in real data, where many components may be close to Gaussianity.

We start with a brief review of the blind source separation paradigm and discuss the role of estimates of moments in fastICA [9, 8]. Section 3 then presents analytical results concerning the optimally achievable estimation quality in the Cramér-Rao sense, and compares this bound with the expected performance of a particular estimation scheme. In section 4, we illustrate these results by a set of simulations, which demonstrate that sample-size effects are drastic for some parameter settings. Finally, in Section 5, we conclude by discussing implications of the presented results for practical applications of ICA algorithms.

2 Blind source separation

The linear blind source separation model describes the mixture of signals by a matrix multiplication

$$\mathbf{X} = \mathbf{A}\mathbf{S} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^d$ represents the observed data and $\mathbf{S} \in \mathbb{R}^n$ the underlying source signals, both modeled by random vectors. A full-rank matrix $\mathbf{W} \in \mathbb{R}^{n \times d}$ is said to be a solution

of the ICA of $\mathbf{X}$ if $\mathbf{WX}$ is statistically independent. We assume that $n = d$ and that $\mathbf{S}$ has at most one Gaussian component. It can be shown that $\mathbf{W}^{-1}$ equals $\mathbf{A}$ except for right-multiplication by a permutation and a scaling matrix [5, 15].

We are interested in how many samples are needed to guarantee certain recovery rate with a confidence of say $\alpha = 95\%$. In order to study this question, we make the following restrictions in addition to the linearity and independence assumption:

- We only consider higher-order statistics, i.e. assume that the data have been whitened and $\mathrm{Cov}(\mathbf{X}) = \mathrm{Cov}(\mathbf{S}) = \mathbf{I}$, without estimation errors. This allows to consider only orthogonal $\mathbf{A}$ and $\mathbf{W}$. Further, we assume vanishing source skewnesses (such as in the case of sources with symmetric densities) and non-trivial kurtoses, so that we can restrict ourselves to fourth-order statistics.
- We restrict ourselves to 'deflation', i.e. the one-unit approach, where a single column of the mixing matrix is extracted, such as in the fastICA algorithm [9, 8].
- In order to reduce the semi-parametric model to a well-defined estimation problem, we consider sources from Pearson's family, cf. Fig. 1.

Our goal is to invert (1), i.e. to find $\mathbf{W}$ with $\mathbf{WX}$ independent. In deflation mode, this is done one-by-one by only extracting a single source within $\mathbf{X}$ i.e. a single row $\mathbf{w}_i^\top$ of $\mathbf{W}$. It can be shown that this can be achieved by maximizing the kurtosis $\mathrm{kurt}(\mathbf{w}_i^\top \mathbf{X})$ of the transformation [9, 8], where the (normalized) kurtosis $\mathrm{kurt}(S) := \kappa(S) - 3$ is zero if S is normal.

Starting with a minimum-mutual-information approach, we get $I(\mathbf{WX}) = \sum_i H(S_i) - H(\mathbf{X})$ (assuming orthogonal $\mathbf{W}$), with $H(\mathbf{X})$ being constant. This equals maximizing the above general likelihood because $\sum_t \log p(\mathbf{w}_i^\top \mathbf{x}(t)) = E(\log p(\mathbf{w}_i^\top \mathbf{X})) = H(\mathbf{w}_i^\top \mathbf{X})$. So $I(\mathbf{WX}) = c - \sum_i J(S_i)$ with some constant c independent of $\mathbf{W}$ and the negentropy $J(S_i) = H(S_{i,\text{gauss}}) - H(S_i)$. Using Gram-Charlier expansion of densities, the negentropy can be approximated as

$$J(S_i) = \frac{1}{12} E(S_i^3)^2 + \frac{1}{48} \mathrm{kurt}(S_i)^2 + \text{higher-order statistics.} \tag{2}$$

Due to our assumption of vanishing skewness we may approximate $\sum_i J(S_i)$ by

$$\sum\nolimits_i \mathrm{kurt}(\mathbf{w}_i^\top \mathbf{X})^2. \tag{3}$$

Instead of (3) we consider now the related maximization problem $\sum_i \sigma_i \mathrm{kurt}(\mathbf{w}_i^\top \mathbf{X})$ with $\sigma_i := \mathrm{sign}\,\mathrm{kurt}(\mathbf{w}_i^\top \mathbf{X})$. Relying on the independence of the sources we can write

$$\mathrm{kurt}(\mathbf{w}^\top \mathbf{X}) = \mathrm{kurt}(\mathbf{v}^\top \mathbf{S}) = \sum\nolimits_i v_i^4 \,\mathrm{kurt}(S_i), \tag{4}$$

where $\mathbf{v}^\top := \mathbf{w}^\top \mathbf{A}$ with $\|\mathbf{v}\| = 1$ [9]. The extrema of the expression (4) are assumed for $\mathbf{v} = \pm \mathbf{e}_i$ for some i, where $\mathbf{e}_i$ denotes the i-th unit vector. But since $\mathbf{A}$ is unknown, $\mathbf{v}$ is an implicit variable. Therefore the result of the ICA i.e. the extraction of the vector $\mathbf{w}$ strongly depends on the quality of the estimation of the kurtosis.

The performance of the ICA algorithm depends thus essentially on the error of the empirical kurtosis $\widehat{\mathrm{kurt}}(\mathbf{w}^\top \mathbf{X})$. Denoting the random variable $\mathrm{kurt}(\mathbf{w}^\top \mathbf{X}) -$

$\widehat{\mathrm{kurt}}(\mathbf{w}^\top \mathbf{X})$ by Δ, we realize that, instead of the ideal result $\mathbf{v} = \pm\mathbf{e}_i$, Eq. (4) yields

$$\left(v_j \mp e_{ij}\right)^4 \sim \frac{\Delta}{n\,\mathrm{kurt}(S_j)}, \tag{5}$$

where e_{ij} is the jth component of the unit vector $\mathbf{e}_i$. Typically, Δ is of the order of the variance of $\widehat{\mathrm{kurt}}(\mathbf{w}^\top \mathbf{X})$, below we will analyze the latter random variable.

3 Finite-sample effects in the estimation

We will proceed in three steps. First we consider the quality of the commonly used estimator for the kurtosis $\hat{\kappa} = \hat{\mu}_4$. It is unbiased since sample-size dependent weighting is not required when the lower moments are assumed to be known. Next we discuss the minimal variance of an unbiased estimator of the kurtosis based on the Cramér-Rao inequality, cf. 3.3. Finally we will present results from simulations in section 4, emphasizing especially scaling effects in the sample size, whereas the first two analytical consideration concentrate on the effect *per* sample.

3.1 A test for kurtosis

Kurtosis is a relevant property in tests of normality of a empirically given distribution. In this context it is natural to assume kurtosis to be small. Estimators that are efficient far from Gaussianity are not known, and the following theorem provides some insight why this is the case:

Theorem [3]: If the kurtosis κ is estimated by the sample moments $\hat{\mu}_r = N^{-1}\sum_{n=1}^{N}(x_n - \mu)^r = E(x_n - \mu)^r$, the resulting estimator $\hat{\kappa}$ deviates from the true kurtosis as follows:

$$(\hat{\kappa} - \kappa) = \frac{\beta}{\hat{\sigma}^4}\frac{1}{N}\sum\nolimits_{n=1}^{N}\mathbf{B}_n + o\left(1\right), \tag{6}$$

where $\beta = \left(1, -4\mu_3, -2\sigma^2\kappa\right)$ and $\mathbf{B}_n = \left((x_n - \mu)^4 - \mu_4, x_n - \mu, (x_n - \mu)^2 - \sigma^2\right)^\top$. The theorem refers to the estimator of kurtosis that is given by the estimates of the fourth central moment and the variance. Because we have fixed the variance to unity, we can write $\hat{\sigma}^2 = \sigma^2 = 1$ and the kurtosis estimate is the only source of the error. The estimator of the kurtosis is asymptotically normal [3] and its variance is given by $\hat{\sigma}_\kappa^2 = \hat{\beta}^\top E[\hat{\mathbf{B}}\hat{\mathbf{B}}^\top]\hat{\beta}/N$. In our case of a non-skewed distribution with unit variance this expression simplifies to

$$E[(\kappa - 2\kappa x^2 + x^4)^2]/N. \tag{7}$$

This result will be used below as an approximation of Δ in Eq. 5.

3.2 Pearson's families

The family of leptokurtotic (super-Gaussian) distribution functions defined by

$$f_{\mathrm{VII}}\left(x; m\right) = \frac{\left(1 + \frac{\kappa_0 x^2}{6+2\kappa_0}\right)^{-m}\Gamma\left(m\right)}{\sqrt{1 + \frac{3}{\kappa_0}}\sqrt{2\pi}\Gamma\left(m - \frac{1}{2}\right)} \tag{8}$$

is known as Pearson type VII [1]. The distributions are symmetric. For $m = \frac{5}{2} + \frac{3}{\kappa_0}$ with $\kappa_0 = \kappa - 3$ the distribution is singly-parameterized by its kurtosis $\kappa > 3$.

The platykurtotic Pearson type II family is given by

$$f_{\mathrm{II}}(x;r) = \frac{\Gamma\left(r+\frac{3}{2}\right)}{\Gamma(r+1)} \frac{1}{\sqrt{\pi r^2}} \left(1 - \frac{(x-\mu)^2}{r^2}\right)^r. \tag{9}$$

f_{II} is symmetric for $\mu = 0$. With the further substitutions $r = \sqrt{3+2m}$ and $m = \frac{3}{3-\kappa} - \frac{5}{2}$, it is parameterized by its kurtosis $\kappa \in (1,3)$. The Gaussian case yields in both families as the limit $\kappa \to 3$ such that we can consider the two families as connected.

For the platykurtotic family the variance decays with the distance from the Gaussian case, cf. Fig. 2(a). This indicates that a reliable estimate of the kurtosis can be obtained already at relatively small sample sizes. Intuitively, this is due to the dramatic structural changes that the distributions of the platykurtotic family undergo when varying κ. Thus a few samples are sufficient to identify a confidence interval for the kurtosis. For $\kappa \to \infty$, however, the density functions of the leptokurtotic family slowly approach the limiting distribution $3(2+x^2)^{-5/2}$, which does not differ too much from the Gaussian distribution as can be seen in Fig. 1. The variance of the kurtosis increases (cf. Fig. 2(b)) because the changes of the form of the density become smaller and smaller. Note that the absolute values in Fig. 2 refer to the variance of the estimator from a single sample and are to be divided by the sample size.

3.3 Fisher information

The Cramér-Rao inequality $\mathrm{Var}(\hat{\kappa}) \geq J(\kappa)^{-1}$ states that the variance of any specific unbiased estimator for the kurtosis is bounded from below by the inverse of the Fisher information

$$J(\kappa) = \int p(x;\kappa) \left(\frac{\partial \log p(x;\kappa)}{\partial \kappa}\right)^2 dx, \tag{10}$$

where the integral is taken over the support of the density function $p(x)$. The integral (10) can be calculated analytically for Pearson's families in the sense of a principal value. The respective formulas have been obtained using Mathematica and the result is graphically displayed in Fig. 2.

An unbiased estimator is efficient if it meets the Cramér-Rao bound. The above estimator of the kurtosis [3] achieves minimal variance only in the Gaussian case. For non-trivial kurtoses the quality of the estimator deviates considerably, see Fig. 2.

4 Simulations

At first, we study the estimation error by considering a $(n = 2)$-dimensional uniform distribution with covariance $\mathrm{Cov}(\mathbf{S}) = \mathbf{I}$. The differences of the recovered demixing matrix $\mathbf{W}$ from the mixing matrix $\mathbf{A}$ can be evaluated by Amari's performance index

$$E_1(\mathbf{C}) = \sum_{r=1}^{n} \sum_{s=1}^{n} \left(\frac{|c_{rs}|}{\max_i |c_{ri}|} + \frac{|c_{sr}|}{\max_i |c_{ir}|} - 2\right), \tag{11}$$

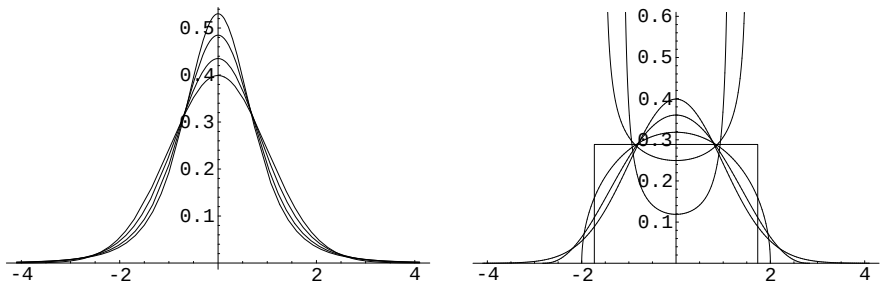


Fig. 1. The left subfigure shows the leptokurtotic (or super-Gaussian) Pearson VII family, which equals the normalized Student-t distribution. All distribution functions (given here for the values $\kappa = 3,\ 4,\ 8$, and $\kappa \to \infty$) are uni-modal and vary slowly in shape for increasing κ. For $\kappa \to \infty$ the family approaches the limit distribution $p(x) = 3(2 + x^2)^{-5/2}$, $\kappa \to \infty$, where strictly speaking the kurtosis does not exist. For $\kappa = 3$, we get the standard normal distribution. The right figure represents the platykurtotic (or sub-Gaussian) Pearson II family. The curve with the largest central peak is again the Gaussian, which connects both families at $\kappa = 3$. Further values of κ are 2.4, 2, 1.8, 1.6 and 1.2. At $\kappa = 2$ the distribution function forms a semi-ellipse, while for $\kappa = 1.8$ it is of box-shape, i.e. the uniform distribution is a member of this family. Eventually, at $\kappa = 1$ the function degenerates into a sum of two δ-distributions which are located at $x = \pm 1$ (not shown). For $1 \leq \kappa < 3$ the support of the functions is bounded by $\pm\sqrt{\frac{6}{3-\kappa} - 2}$.

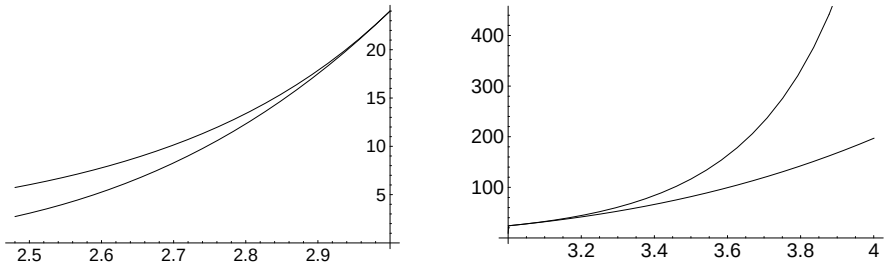


Fig. 2. Cramér-Rao bound (lower curves) for the estimation of the kurtosis κ as the parameter of the platykurtotic Pearson family (left) and the leptokurtotic Pearson family (right). The upper curves represent the respective variances of the kurtosis estimator [3]. The variance of the estimator coincides with the inverse Fisher information and is thus efficient only in the Gaussian limit ($\kappa = 3$), where we find consistently $J^{-1}(3) = 24$.

which quantifies deviations of $\mathbf{C} = \mathbf{WA}$ from a scaling and permutation matrix [2]. This measures the performance of the full ICA algorithm. For analyzing a single deflation step we make use of the single column error

$$E_2(\mathbf{v}) := \min_{\mathbf{e}_i} \|\mathbf{v} \pm \mathbf{e}_i\|. \tag{12}$$

The comparison is made over 1000 runs, in which the coefficients of $\mathbf{A}$ have been drawn uniformly from the orthogonal group. In Fig. 3, we see that for varying number of samples N, indeed the error gets smaller, roughly following the expected $1/\sqrt{N}$-law.

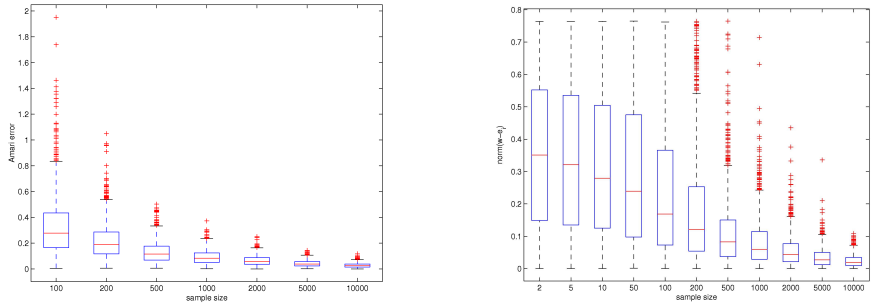


Fig. 3. Cross-talking error $E_1(\mathbf{WA})$ (11) for total matrix (left) and recovery error (12) for a single column (right). The boxes are delimited by the quartiles and intersected by the median. Data are marked by crosses if their distance from the box is more than 1.5 times the interquartile distance.

κ	1.5	1.6	1.8	2.0	2.5	2.9	3.0	3.1	4.0	6.0	9.0	13.0
T_0	3	5	9	17	146	4242	∞	6031	184	80	65	58
T_{min}	71	96	189	319	2251	72829	∞	85464	2557	1026	869	723
$T_{\alpha=5\%}$	457	617	1218	2060	14562	471536	∞	553346	16542	6637	5620	4675

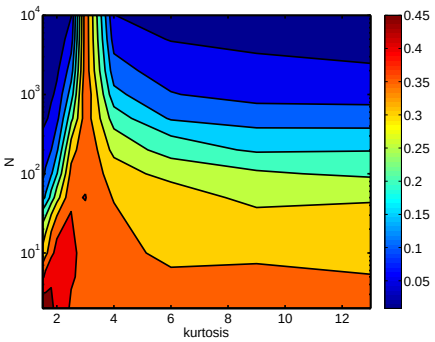
Table 1. T_0 is defined as the maximal number of samples where still the maximum error is likely to be observed, while T_{min} is the minimal number of patters which is needed in order to achieve an average error (12) of 0.1. And $T_{\alpha=5\%}$ is the error (0.039) at the 5%-quantile, when sampling random directions and calculating the resulting error.

Now, we calculate the mean and standard deviation when varying the number of samples N and the source kurtosis κ (assumed to be equal for all source components). This is done in $n = 2$ dimensions for 1000 batch runs to get sufficient statistics. In Fig. 4, the results are presented. Clearly the error increasing to randomness in the case of Gaussians ($\kappa = 3$) as was to be expected due to the symmetry in that case. For values different from 3, recovery error versus N can be analyzed, and we observe that this relationship follows a power-law with exponent $-\frac{1}{2}$, confirming again the asymptotic $1/\sqrt{N}$-decrease, cf. Eq. 7.

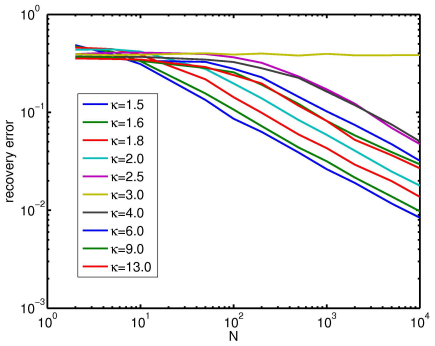
A further set of simulations studies the effect of temporal correlations which is included here because it shows a behavior opposite to the intuition implied by Fig. 2. For the more structured platykurtotic distributions, temporal correlations yield a larger recovery error than the less informative leptokurtotic distributions. The obtained differences are significant for the illustrative case studied here and cannot be ignored in less obvious situations.

5 Discussion

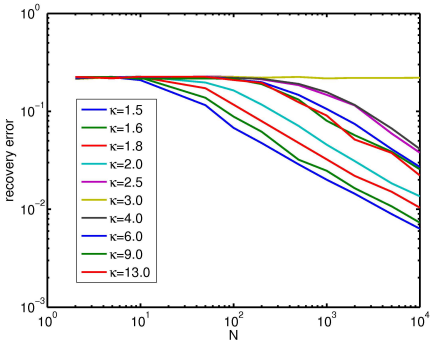
For large sample sizes the square-root decay of both the error and the variance reproduces relations (6) and (7), In Figs. 4b and c it is obvious that the intercept of the log-linear fit to the asymptotic part of the equation depends clearly on the true value of the



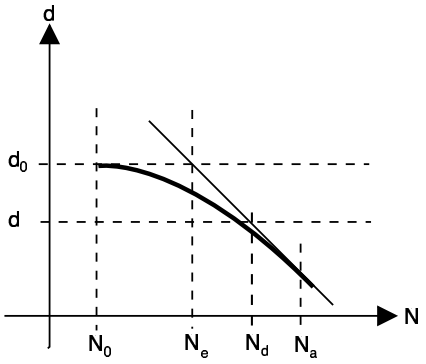
(a) Mean matrix-row estimation error



(b) Mean recovery error versus number of samples N for varying kurtosis



(c) Standard deviation of the recovery error versus number of samples N



(d) Construction of maximal N bounds

Fig. 4. Comparison of row estimation of $\mathbf{W}$ in the symmetric Pearson family for varying number of samples N and kurtosis κ. Mean and variance are taken over 1000 runs. Figures (a) and (c) show the mean, and (b) the corresponding standard deviation. The asymptotic $1/\sqrt{N}$-decrease can be used to fit a line in the log-log plots and to determine its intersection with the maximal error to get bounds for minimal number of samples N_e and N_d depending on the error threshold.

kurtosis that is estimated. For sample sizes below the intercept, no information about the component is extracted from the mixture. The respective values are corrected for the estimation quality as discussed in section 3.3 and are illustrated by the values in Tab. 1. They can be obtained from the simulation as illustrated in Fig. 4 d. Asymptotically we can assume $1/\sqrt{N}$-decay, say at $N = N_a$. We are then interested at the number of samples N_d that produce precisely an error of $d(N_d)$ according to this upper bound. This can be easily derived as $N_d = N_a d_a^2/d^2$. Interesting upper bounds N are N_e for the maximal error $d_{\max}$, which is in practice achieved at much smaller samples, and the maximal error at confidence level of α, where statistics are performed over randomly sample distances and resulting errors d, see Tab. 1.

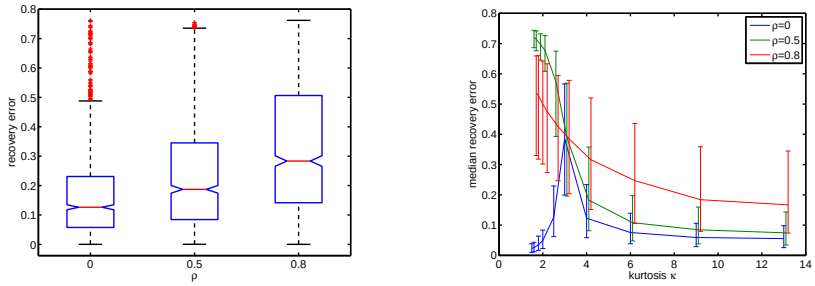


Fig. 5. $T = 1000$ samples of a two-dimensional problem are drawn from a Pearson distribution with kurtosis $\kappa = 4$ with a temporal correlation $X(t) = \rho X(t-1) + \epsilon(t)$, where $\epsilon(t)$ are i.i.d. random variables. Presented is an average over 1000 runs for each value ρ. In the plot on the left, the depths of the dents next to the median (compare Fig. 3) indicate that the significancy of the difference according to the Wilcoxon rank sum test ($\rho = 0.0$ vs. $\rho = 0.5$: $p = 3.1 \times 10^{-15}$, $\rho = 0.0$ vs. $\rho = 0.8$: $p = 3.5 \times 10^{-62}$, and $\rho = 0.5$ vs. $\rho = 0.8$: $p = 4.1 \times 10^{-21}$). The plot on the right shows the median and the quartiles of the recovery error in dependency of the source kurtosis κ for different levels of temporal coherency.

Combining Eqs. 12 and 5 we notice that kurtoses $\kappa \approx 3$ will have a larger effect on the ICA errors than inefficiencies of the estimator. For the estimate of the minimal sample size we can rely on the additivity of the Fisher information provided that samples are independently chosen. In this case we can derive a heuristic correction from the result in Fig. 2. We multiply the minimal sample size by the square root ratio of the variances of the used estimator [3] and the Cramér-Rao bound, which partially compensates the dependency of the minimal sample size on the true kurtosis. For temporally correlated samples, however, the errors are more grave, cf. the following section and Fig. 5.

6 Conclusion

In order to arrive at a practically useful criterion to determine the minimal sample size in ICA, we have chosen Pearson's family as a prototypical case, because it is non-trivially focused on the problem of kurtosis estimation while analytical consideration are still possible. Based on the presented results we claim that often, e.g. in the analysis of fMRI data sample sizes are too small to permit interpretable results. This situation often becomes more critical because of further sources of errors. In kurtotic distributions, deviations from the mean are either the common (platykurtotic case) or drastic because of heavy tails (leptokurtotic case) and the presence (or absence) of extreme values of the observation has a large effect on the estimate of any higher-order moments. This is also indicated by the local scores of the Fisher information tend to be large at these extreme values, such that outlier rejection is a questionable option. On the other hand the study of temporally correlated samples, cf. Fig. 5 shows that less informative families of distributions are less vulnerable to this type of deviations from the standard assumptions of ICA.

The overall picture that emerges here can be summarized as follows: While source distributions of near Gaussianity are naturally difficult, both, lepto- and platykurtotic distributions may cause substantial errors in the reconstructed source signals which persists up to moderate sample sizes. The results obtained here are likely to be characteristic for larger classes of probability distributions, although also other factors may then affect the errors. In addition to the errors due to finite-sample effects, obviously also nonstationarities or violations of the independence assumption may influence the result. Algorithms [13] that exploit temporal correlations require less severe restriction, and may thus avoid some of the problems discussed here. Finally we want to refer the reader to effects of finite sample sizes in the estimation of lower-order moments, for which exist more studies [12] than for the higher-order estimation problem.

Acknowledgment: This work was partially supported by the BMBF, grant number 01GQ0432. Discussions with T. Geisel are gratefully acknowledged.

References

1. M. Abramowitz and I. A. Stegun (eds.). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover, 1972.
2. S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8:757–763, 1996.
3. J. Bai and S. Ng. Tests for skewness, kurtosis, and normality for time series data. *J. Business & Economic Statistics*, 23(1):49–60, 2005.
4. M. Bethge. Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *J. Opt. Soc. Am. A*, 23(6):1253–1268, 2006.
5. P. Comon. Independent component analysis - a new concept? *Signal Processing*, 36:287–314, 1994.
6. S. Dodel, J. M. Herrmann, and T. Geisel. Comparison of temporal and spatial ica in fmri data analysis. In *Proc. Second Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, volume 13, pages 543–547, 2000.
7. N. Farvardin and J. Modestino. Optimum quantizer performance for a class of non-gaussian memoryless sources. *IEEE Trans. Inf. Theory*, IT-30:485–497, 1984.
8. A. Hyvärinen, J. Karhunen, and E. Oja. Independent component analysis. *John Wiley & Sons*, 2001.
9. A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.
10. C. Jutten, J. Hérault, P. Comon, and E. Sorouchiary. Blind separation of sources, parts I, II and III. *Signal Processing*, 24:1–29, 1991.
11. Z. Koldovský, P. Tichavský, and E. Oja. Efficient variant of algorithm fastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Transactions on Neural Networks*, 17(5):1265–1277, 2006.
12. R. C. MacCallum, M. W. Browne, and H. M. Sugawara. Power analysis and determination of sample size for covariance structure models. *Psychological Methods*, 1(2):130–149, 1996.
13. L. Molgedey and G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637, 1994.
14. P. A. Regalia and E. Kofidis. Monotonic convergence of fixed-point algorithms for ICA. *IEEE Transactions on Neural Networks*, 14(4):943–949, 2003.
15. F.J. Theis. A new concept for separability problems in blind source separation. *Neural Computation*, 16:1827–1850, 2004.