ROBUST ICA BASED ON TWO SCATTER MATRICES

Klaus Nordhausen[†], Hannu Oja[†], Esa Ollila[‡]
[†]*Tampere School of Public Health*
University of Tampere
33014 University of Tampere, FINLAND
[‡]*Signal Processing Laboratory*
Helsinki University of Technology
P.O.Box 3000, 02015 HUT, FINLAND
e-mail: klaus.nordhausen@uta.fi, hannu.oja@uta.fi, esollila@wooster.hut.fi

Abstract

Oja et al. [11] and Ollila et al. [12] showed that, under general assumptions, any two scatter matrices with the so called independent components property can be used to estimate the unmixing matrix for the independent component analysis (ICA). The method is a generalization of Cardoso's [2] FOBI estimate which uses the regular covariance matrix and a scatter matrix based on fourth moments. Different choices of the two scatter matrices are compared in a simulation study. Based on the study, we recommend always the use of two robust scatter matrices. For possible asymmetric independent components, symmetrized versions of the scatter matrix estimates should be used.

1 Introduction

Let $x_1, x_2, \ldots, x_n$ be a random sample from a p-variate distribution, and write

$$X = \left(\begin{array}{cccc} x_1 & x_2 & \ldots & x_n \end{array} \right)$$

for the $p \times n$ data matrix. We assume that the x_i are generated by

$$x_i \;=\; A\, z_i, \quad i = 1, ..., n,$$

where the z_i are independent latent vectors having independent components and A is a full-rank $p \times p$ *mixing matrix*. This model is called the *independent component (IC) model*. The model is not well defined in the sense that

$$X \;=\; A^*\, Z^* \;=\; (AP'D^{-1})\,(DPZ),$$

for all diagonal matrices D (with nonzero diagonal elements) and for all permutation matrices P. Permutation matrix P is obtained from identity matrix I_p by permuting its rows. The problem in the so called *independent component analysis (ICA)* is to find a *unmixing matrix* B such that Bx_i has independent components. The solution is naturally not unique: If B is an unmixing matrix, then so is DPB.

Most ICA algorithms then proceed as follows. (For a recent review of different approaches, see Hyvärinen et al. [7].)

1. To simplify the problem it is first commonly assumed that the x_i are *whitened* so that $E(x_i) = 0$ and $Cov(x_i) = I_p$ Then

$$X \;=\; U\; Z^*$$

with an orthogonal matrix U and Z^* with (columns having) independent components such that $E(z_i^*) = 0$ and $Cov(z_i^*) = I_p$

2. For the whitened data, find a $p \times r$ matrix U with orthonormal columns ($r \leq p$) which maximizes (or minimizes) a chosen criterion function, say $g(U'X)$. Measures of marginal nongaussianity (negentropy, kurtosis measures) $g(u'X)$ and likelihood functions with different choices of marginal distributions are often used.

In the FastICA algorithm (Hyvärinen and Oja [5]) for example in each iteration step (for stage 2) the columns of U are updated one by one and then orthogonalized. The criterion of the FastICA algorithm maximizes the negentropy which is approximated by

$$g(u'X) = [\text{ave}\{h(u'x_i)\} - E[h(v)]]^2 \tag{1}$$

with $v \sim N(0,1)$ and with several possible choices for the function $h(\cdot)$.

A different solution to the ICA problem, called FOBI, was given by Cardoso [2]: After whitening the data as above (stage 1), an orthogonal matrix U is found as the matrix of eigenvectors of a kurtosis matrix (matrix of fourth moments; this will be discussed later). The data transformation then jointly diagonalized the regular covariance matrix and a scatter matrix based on fourth moments. FOBI was generalized in Oja et al. [11] (real data) and Ollila et al. [12] (complex data) where any two scatter matrices which have the so called independence property can be used. An interesting question then naturally arises: How should one choose these two scatter matrices in a good or optimal way?

2 Two Scatter Matrices and ICA

Let x be a p-variate random vector with cdf F_x. A functional $T(F)$ is a p-variate *location vector* if it is affine equivariant in the sense that $T(F_{Ax+b}) = AT(F_x) + b$ for all x, all full-rank $p \times p$ matrices A and all p-variate vectors b. Using the same notation, a matrix-valued $p \times p$ functional $S(F)$ is called a *scatter matrix* if it is positive definite, symmetric and affine equivariant in such way that $S(F_{Ax+b}) = AS(F_x)A'$ for all x, A and b. The regular mean vector $E(x)$ and covariance matrix $Cov(x)$ serve as first examples. There are numerous alternative techniques to construct location and scatter functionals, e.g. M-functionals, S-functionals, etc. See e.g. Maronna et al. [9].

A scatter matrix $S(F)$ is said to have the *independent components (IC-) property* if $S(F_z)$ is a diagonal matrix for all z having independent components. The covariance matrix naturally has the IC-property. Other classical scatter functionals (M-functionals, S-functionals, etc.) developed for elliptical distributions do not generally possess the IC-property. However, if z has independent and symmetrically distributed

components, then $S(F_z)$ is a diagonal matrix for all scatter functionals S. It is therefore possible to develop a symmetrized version of a scatter matrix $S(F)$, say $S_{sym}(F)$, which has the IC-property; just define

$$S_{sym}(F_x) \;=\; S(F_{x_1-x_2})$$

where x_1 and x_2 are two independent copies of X. See Oja et al. [11], Ollila et al. [12] and Sirkiä et al. [14].

An alternative approach to the ICA using two scatter matrices with IC-property (Oja et al. [11], Ollila et al. [12]) has the following two steps:

1. The x_i are whitened using S_1 (instead of the covariance matrix) so that $S_1(F_{x_i}) = I_p$. Then
$$X \;=\; U\; Z^*$$
with an orthogonal matrix U and with Z^* with (columns having) independent components such that $S_1(z_i^*) = I_p$.

2. For the whitened data, find an orthogonal matrix U as the matrix of eigenvectors of $S_2(F_{x_i})$.

The resulting data transformation $X \to \hat{B}X$ then jointly diagonalized S_1 and S_2 ($S_1(\hat{B}X) = I_p$ and $S_2(\hat{B}X) = D$) and the unmixing matrix $\hat{B}$ solves

$$S_2^{-1}S_1B' = B'D^{-1}$$

The matrix $\hat{B}$ is the matrix of eigenvectors and the diagonal matrix $\hat{D}$ is the matrix of eigenvalues of of $S_2^{-1}S_1$; the independent components are then ordered according to their kurtosis. The solution is unique if the eigenvalues of $S_2^{-1}S_1$ are distinct.

Note that different choices of S_1 and S_2 yield different estimates $\hat{B}$. First, the resulting independent components $\hat{B}X$ are rescaled by S_1 and they are given in an order determined by S_2. Also the statistical properties of the estimates $\hat{B}$ (convergence, limiting distributions, efficiency, robustness, etc.) naturally depend on the choices of S_1 and S_2.

3 Performance Study

3.1 The estimates $\hat{\mathbf{B}}$ to be compared

We now study the behavior of the new estimates $\hat{B}$ with different (robust and non-robust) choices for S_1 and S_2. The classical FastICA procedures which use

$$h_1(u'x_i) = \log(\cosh(u'x_i)) \;\text{ or }\; h_2(u'x_i) = -\exp(-u'x_i)$$

in Equation (1) serve as a reference. These algorithms will be denoted as *FastICA1* and as *FastICA2*, respectively. According to Hyvärinen and Oja [6], these choices are more robust than the traditional negentropy estimate with criterion

$$g(u'X) = \frac{1}{12}\left[\text{ave}\left\{(u'x_i)^3\right\}\right]^2 + \frac{1}{48}\left[\text{ave}\left\{(u'x_i)^4\right\} - 3\right]^2$$

In the following we assume that the x_i are centered (using a preliminary or simultaneous location functional). The *FOBI* estimate by Cardoso [2] assumes that the centering is done using the mean vector, and

$$S_1(F_x) = Cov(x) \text{ and } S_2(F_x) = \frac{1}{p+2} E\left[||S_1^{-1/2}x||^2 xx'\right].$$

Then S_2 is a scatter matrix based on the fourth moments, both S_1 and S_2 possess the IC-property, and the independent components are ordered with respect to classical kurtosis measure. The FOBI estimate is member in the new class of estimates but highly non-robust due to the choices of S_1 and S_2.

In our simulation study we consider scatter matrices which are (unsymmetrized and symmetrized) M-functionals. A M-functional of scatter corresponding to a chosen weight function $w(r)$ is a functional which satisfies the implicit equation

$$S(F_x) = E[w(r)xx']$$

where r is the Mahalanobis distance between x and the origin, i.e. $r^2 = x'S^{-1}x$. In this paper we consider Huber's M-estimator [9] with

$$w(r) = \begin{cases} \frac{1}{\sigma^2} & r^2 \leq c^2 \\ \frac{c^2}{\sigma^2 r^2} & r^2 > c^2 \end{cases}$$

The tuning constant c is chosen to satisfy $q = Pr(\chi_p^2 \leq c^2)$ and the scaling factor σ^2 so that $E[\chi_p^2 w(\chi_p^2)] = p$. Tyler's shape matrix [15] is often called the most robust M-estimator and has the weight function

$$w(r) = \frac{p}{r^2}$$

Symmetrized versions of Huber's estimate and Tyler's estimate then possess the IC-property. The symmetrized version of Tyler's shape matrix is also know as Dümbgen's shape matrix [4].

In this simulation study we compare

- FastICA1 and FastICA2 estimates
- E1: FOBI estimate
- E2: Estimate based on the covariance matrix and Tyler's shape matrix
- E3: Estimate based on Tyler's shape matrix and the covariance matrix
- E4: Estimate based on Tyler's shape matrix and Dümbgen's shape matrix
- E5: Estimate based on Tyler's shape matrix and Huber's M-estimator ($q = 0.9$)
- E6: Estimate based on Dümbgen's shape matrix and symmetrized Huber's M-estimator ($q = 0.9$).

All computations are done in R 2.4.0 [13]; the package fastICA [8] was used for the FastICA solutions and the package ICS [10] for the new method.

3.2 Simulation Designs

In this simulation study the independent components are all symmetrically distributed. Therefore all choices of S_1 and S_2 are acceptable. The designs were as follows:

- *Design I:* The $p = 4$ independent components were generated from (i) a normal distribution, (ii) a uniform distribution, (iii) a t_3 distribution, and (iv) a Laplace distribution, respectively (all distributions with unit variance.) The sample sizes ranged from $n = 50$ to $n = 2000$. For each sample size, we had 300 repetitions. For all samples, the elements of a mixing matrix A were generated from a $N(0, 1)$ distribution.

- *Design II:* As Design I but with outliers. The $\max(1, 0.01n)$ observations x_i with the largest L_2 norms were multiplied by $s_i u_i$ where s_i is $+1$ or -1 with probabilities $1/2$ and u_i has a $Uniform(1, 5)$ distribution. This was supposed to partially destroy the dependence structure.

3.3 Performance Index

Let A be the "true" mixing matrix in a simulation and $\hat{B}$ an estimate of an unmixing matrix. For any true unmixing matrix B, $BA = PD$ with a diagonal matrix D and a permutation matrix P. Write $G = (g_{ij}) = \hat{B}A$. The performance index (Amari et al. [1])

$$PI(G) = \frac{1}{2p(p-1)} \left[\sum_{i=1}^{p} \left(\sum_{j=1}^{p} \frac{|g_{ij}|}{\max_h |g_{ih}|} - 1 \right) + \sum_{j=1}^{p} \left(\sum_{i=1}^{p} \frac{|g_{ij}|}{\max_h |g_{hj}|} - 1 \right) \right]$$

is then often used in comparisons. Now clearly $PI(PG) = PI(G)$ but $PI(DG) = PI(G)$ is not necessarily true. Therefore, for a fair comparison, we standardize and reorder the rows of $B = (b_1 ... b_p)'$ $(B \to PDB)$ so that

- $||b_i|| = 1$, $i = 1, ..., p$
- $\max(b_{i1}, ..., b_{ip}) = \max(|b_{i1}|, ..., |b_{ip}|)$, $i = 1, ..., p$
- $\max(b_{i1}, ..., b_{ip}) \geq \max(b_{j1}, ..., b_{jp})$, $1 \leq i \leq j \leq p$.

For the comparison, also A^{-1} is standardized in a similar way.

The performance index $PI(G)$ can take values in $[0, 1]$; the smaller is $PI(\hat{B}A)$ the better is the estimate $\hat{B}$.

3.4 Results

The results of the simulations are summarized in Figure 1 and show, that in the non-contaminated case (Design I) the two versions of the fastICA algorithm dominate all estimates based on two scatter matrices. Surprisingly, in this case, the FOBI estimate seems to be the worst choice among all, whereas the best is estimate E6 which is based

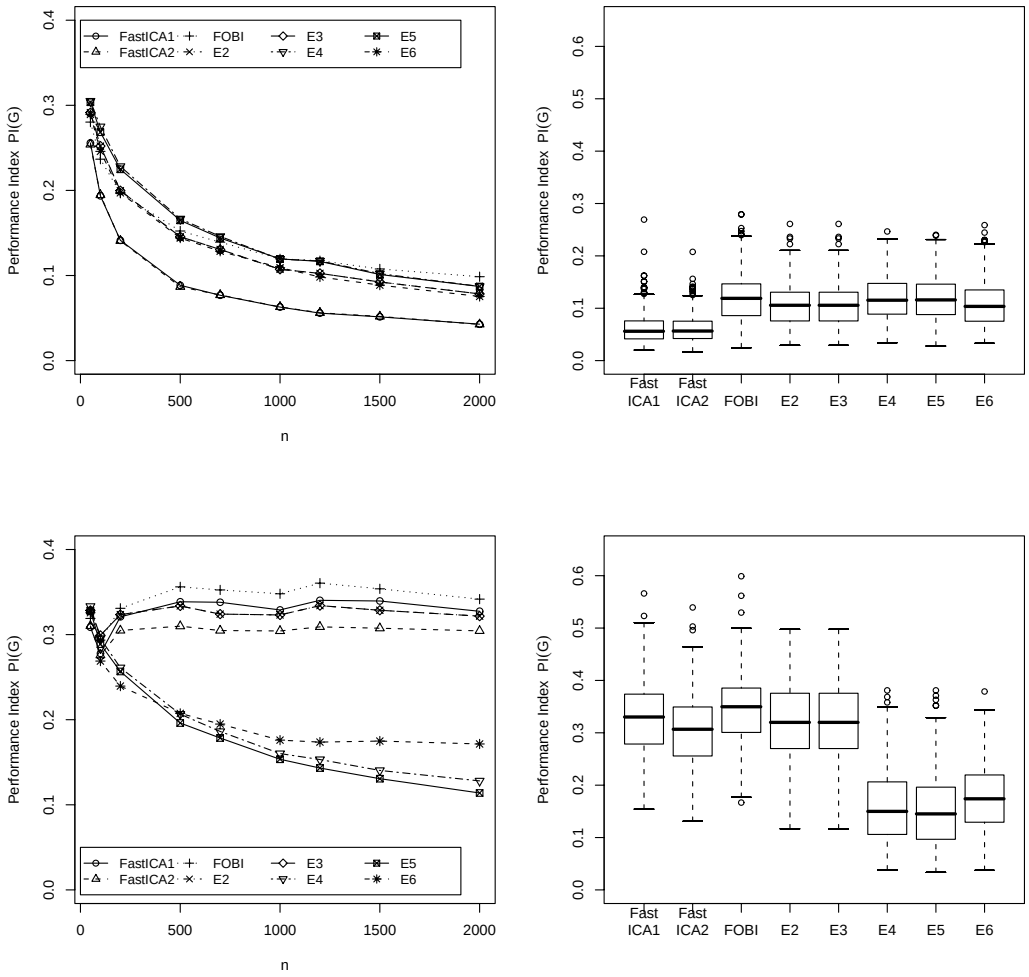

Figure 1: Results of the simulations. The top row shows the results for Design I and the bottom row for Design II. The left column shows the mean of $PI(\hat{B}A)$ for 300 repetitions and the right column boxplots of $PI(G)$ when $n = 1000$. The estimates based on two scatter matrices besides $FOBI$ are $E2$: covariance matrix & Tyler's shape matrix, $E3$: Tyler's shape matrix & covariance matrix, $E4$: Tyler's shape matrix & Dümbgen's shape matrix, $E5$: Tyler's shape matrix & Huber's M-estimator and $E6$: Dümbgen's shape matrix & Symmetrized Huber's M-estimator.

on two symmetrized scatter matrices. The differences are minor, however. The results change a lot when adding outliers (Design II). The procedures $E4$, $E5$ and $E6$ based on two robust scatter matrices are least affected by the outliers. Estimate $E6$ using robust symmetrized estimates presumably has a lowest breakdown point among the robust estimates which may explain its slightly worse behavior here. The order in which the two scatter matrices are used has no effect on the results; $E2$ and $E3$ have naturally the same performance in the simulations.

Based on the simulation results, we recommend the use of two robust scatter matrices in all cases. For possible asymmetric independent components, symmetrized versions of the scatter matrix estimates should be used. Symmetrized scatter matrices are however based on U-statistics and computationally expensive; $n = 1,000$ observations for example means almost $500,000$ pairwise differences. To relieve the computational burden, the original estimate may then be replaced by an estimate which is based on an incomplete U-statistic. Further investigation is needed to examine the situations where the components are not symmetric. For asymmetric independent components, FastICA algorithms for example are known to have a poorer performance.

References

[1] Amari S., Cichocki A., Yang H.H. (1996). A New Learning Algorithm for Blind Source Separation. In *Advances in Neural Information Processing Systems 8, pp. 757-763*. MIT Press, Cambridge, MA.

[2] Cardoso J.F. (1989). Source Separation Using Higher Order Moments. In *Proceedings of IEEE International Conference on Acustics, Speech and Signal Processing, pp. 2109-2112*. Glasgow.

[3] Comon P. (1994). Independent Component Analysis, a New Concept? *Signal Processing*. Vol. **36**, pp. 287-314.

[4] Dümbgen L. (1998). On Tyler's M-functional of Scatter in High Dimension. *Annals of Institute of Statistical Mathematics*. Vol. **50**, pp. 471-491.

[5] Hyvärinen A., Oja E. (1997). A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*. Vol. **9**, pp. 1483-1492.

[6] Hyvärinen A., Oja E. (2000). Independent Component Analysis: Algorithms and Applications. *Neural Networks*. Vol. **13**, pp. 411-430.

[7] Hyvärinen A., Karhunen J., Oja E. (2001). *Independent Component Analysis*. Wiley, New York.

[8] Marchini J.L., Heaton C., Ripley B.D. (2006). *fastICA: FastICA algorithms to perform ICA and Projection Pursuit*. R package version 1.1-8.

[9] Maronna R.A., Martin R.D., Yohai V.J. (2006). *Robust Statistics*. Wiley, Chichester.

[10] Nordhausen K., Oja H., Tyler D. (2006). *ICS: ICS / ICA Computation Based on Two Scatter Matrices*. R package version 0.1-2.

[11] Oja H., Sirkiä S., Eriksson J. (2006). Scatter Matrices and Independent Component Analysis. *Austrian Journal of Statistics*. Vol. **35**, pp. 175-189.

[12] Ollila E., Oja H., Koivunen V. (2007). Complex-valued ICA Based on a Pair of Generalized Covariance Matrices. *Manuscript*.

[13] R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[14] Sirkiä S., Taskinen S., Oja H. (2007). Symmetrized M-estimators of Multivariate Scatter. *Manuscript*.

[15] Tyler D. (1987). A Distribution-free M-estimator of Multivariate Scatter. *Annals of Statistics*. Vol. **15**, pp. 234-251.