AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 1, 91–100

Robust Independent Component Analysis Based on Two Scatter Matrices

Klaus Nordhausen[1], Hannu Oja[1] and Esa Ollila[2]

[1]University of Tampere, Finland
[2]Helsinki University of Technology, Finland

Oja, Sirkiä, and Eriksson (2006) and Ollila, Oja, and Koivunen (2007) 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 (Cardoso, 1989) 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.

Keywords: Affine Equivariance, Kurtosis, Source Separation.

1 Introduction

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

$$X = \left(x_1 \; x_2 \; \ldots \; x_n \right)$$

for the $p \times n$ data matrix. We assume that X is generated by

$$X = AZ,$$

where $Z = (z_1 z_2 \ldots z_n)$ and $z_1, \ldots, z_n$ are independent and identically distributed latent random 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 the model may also be written as

$$X = A^* Z^*$$

where

$$A^* = AP'D^{-1} \qquad \text{and} \qquad Z^* = DPZ$$

for any diagonal matrix D (with nonzero diagonal elements) and for any permutation matrix P. (A permutation matrix P is obtained from identity matrix I_p by permuting its rows.) If Z has independent components, then also the components of $Z^* = DPZ$ are independent. The problem in the so called *independent component analysis (ICA)* is to find an *unmixing matrix* B such that Bx_i has independent components. Based on the discussion above, the solution is then 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, Karhunen, and Oja, 2001.)

1. To simplify the problem it is first commonly assumed that the x_i are *whitened* so that $\mathrm{E}(x_i) = 0$ and $\mathrm{cov}(x_i) = I_p$. Then
$$X = UZ^*$$
with an orthogonal matrix U and Z^* with (columns having) independent components such that $\mathrm{E}(z_i^*) = 0$ and $\mathrm{cov}(z_i^*) = I_p$
2. For the whitened data X, 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, 1997) 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
$$g(u'X) = [\text{ave}\{h(u'x_i)\} - \mathrm{E}[h(z)]]^2 \tag{1}$$
with $z \sim N(0, 1)$ and with several possible choices for the function $h(\cdot)$.

A different solution to the ICA problem, called FOBI, was given by Cardoso (1989): 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 consists of a joint diagonalization of the regular covariance matrix and of the scatter matrix based on fourth moments. FOBI was generalized in Oja et al. (2006) (real data) and Ollila et al. (2007) (complex data) where any two scatter matrices which have the so called independent components property can be used. An interesting question then naturally arises: How should one choose these two scatter matrices in a good or optimal way?

The paper is organized as follows. First, in Section 2 scatter matrices and their use in the estimation of an unmixing matrix is reviewed. In Section 3 we describe the results from simulation studies where new ICA estimates with several choices of scatter matrices are compared to classical FastICA and FOBI estimates. Also an image analysis example is given. The paper ends with some conclusions in Section 4.

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, Martin, and Yohai (2006).

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. (2006), Ollila et al. (2007) and Sirkiä, Taskinen, and Oja (2007).

An alternative approach to the ICA using two scatter matrices with IC-property (Oja et al., 2006, Ollila et al., 2007) 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
$$X = UZ^*$$
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 X, find an orthogonal matrix U as the matrix of eigenvectors of $S_2(F_{x_i})$.

The resulting data transformation $X \rightarrow \hat{B}X$ then jointly diagonalizes 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

$$S_2^{-1} S_1 B' = B' D^{-1}.$$

The matrix $\hat{B}$ is the matrix of eigenvectors and the diagonal matrix $\hat{D}$ is the matrix of eigenvalues of $S_2^{-1} S_1$. Note the similarity between our ICA procedure and the principal component analysis (PCA): The direction u of the first eigenvector of $S_2^{-1} S_1$ maximizes the criterion function $(u' S_1 u)/(u' S_2 u)$ which is a measure of kurtosis (ratio of two scale measures) rather than a measure of dispersion (as in PCA) in the direction u, etc. The independent components are then ordered according to this specific kurtosis measure. The solution is unique if the eigenvalues of $S_2^{-1} S_1$ are distinct.

Different choices of S_1 and S_2 naturally 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) naturally depend on the choices of S_1 and S_2.

3 Performance Study

3.1 The Estimates $\hat{\mathrm{B}}$ 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

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

in equation (1) serve as a reference. These algorithms will be denoted as *FastICA1* and as *FastICA2*, respectively. According to Hyvärinen and Oja (2000), these choices are more robust than the traditional negentropy estimate with criterion

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

The *FOBI* estimate by Cardoso (1989) assumes that the centering is done using the mean vector, and

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

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 their 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. Simultaneous M-functionals for location and scatter corresponding to chosen weight functions $w_1(r)$ and $w_2(r)$ are functionals which satisfy implicit equations

$$T(F_x) = [\text{E}[w_1(r)]]^{-1}\text{E}[w_1(r)x] \qquad \text{and} \qquad S(F_x) = \text{E}[w_2(r)xx'],$$

where r is the Mahalanobis distance between x and $T(F_x)$, i.e.

$$r^2 = (x - T(F_x))'S(F_x)^{-1}(x - T(F_x)).$$

In this paper we consider Huber's M-estimator (Maronna et al., 2006) with

$$w_1(r) = \begin{cases} 1 & r \le c \\ c/r & r > c \end{cases} \qquad \text{and} \qquad w_2(r) = \begin{cases} 1/\sigma^2 & r \le c \\ c^2/\sigma^2 r^2 & r > c. \end{cases}$$

The tuning constant c is chosen to satisfy $q = \Pr(\chi_p^2 \le c^2)$ and the scaling factor σ^2 so that $\text{E}[\chi_p^2 w_2(\chi_p^2)] = p$. Tyler's shape matrix (Tyler, 1987) is often called the most robust M-estimator. Tyler's shape matrix and simultaneous spatial median estimate, see (Hettmansperger and Randles, 2002), have the weight functions

$$w_1(r) = \frac{1}{r} \qquad \text{and} \qquad w_2(r) = \frac{p}{r^2}.$$

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 (Dümbgen, 1998).
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 (R Development Core Team, 2006); the package fastICA (Marchini, Heaton, and Ripley, 2006) was used for the FastICA solutions and the package ICS (Nordhausen, Oja, and Tyler, 2006) 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, Cichocki, and Yang, 1996)

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

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 $B = (b_1 \dots b_p)'$ ($B \to PDB$) such that

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

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 $\hat{B}$.

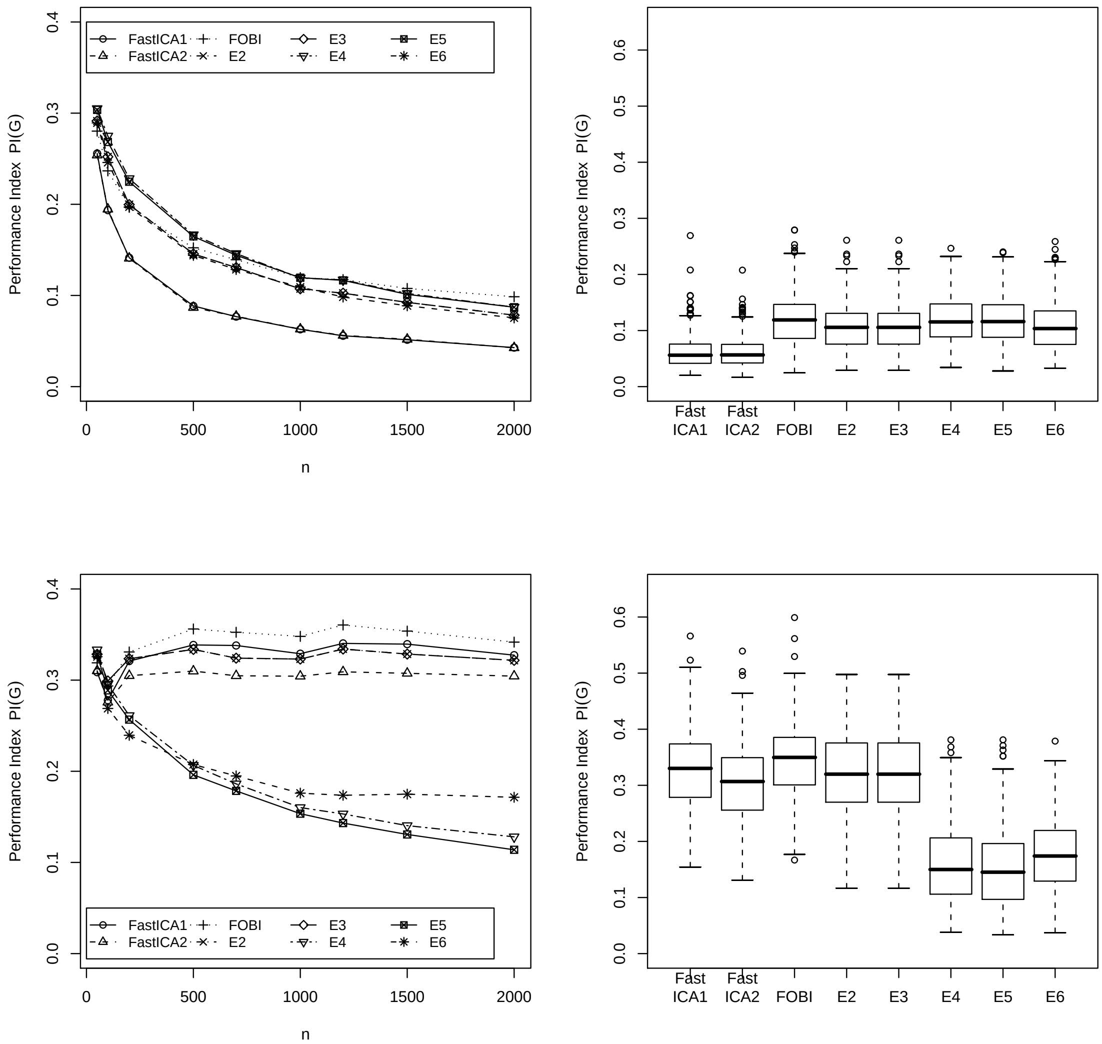

Figure 1: Results of the simulations. The top row shows the results for Design I and the bottom row for Design II. The left column shows the mean of $PI(\hat{B}A)$ for 300 repetitions and the right column boxplots of $PI(G)$ when $n = 1000$. The estimates based on two scatter matrices besides $FOBI$ 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.

3.4 Simulation 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 on two symmetrized scatter matrices. The differences are minor, however. The results change considerably 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.

3.5 An Example

To demonstrate the effect of outliers in a real example we will attempt to unmix three mixed images. The original images which show a cat, a forest track and a sheep, are all in a greyscale having each 130×130 pixels and are part of the the R-package ICS. In the analysis of image data, the pixels are thought to be individuals ($n = 130 \times 130$), and each individual has three measurements corresponding to the three pictures ($p = 3$). The three pictures are first mixed with a random 3×3 matrix using the vector representation of the pictures. Contamination is added to the first mixed image by blackening 60 pixels in the right upper corner, which corresponds to less than 1 percent of outliers. The algorithms $E5$ and $FastICA2$ are then applied to recover the original images. To retransform the independent components to a reasonable greyscale, for all independent components, values smaller than the 2.5% quantile are replaced by the quantile and the same was done for values larger than the the 97.5% quantile. The result is shown in Figure 2.

As can be seen, some images are negatives of the original images. This is due to the arbitrary sign of the independent components. Nevertheless, it can be observed, that $E5$ performs better than $FastICA2$ even when the amount of contamination is so small. The algorithm $E5$ recovers the two images with the sheep and the cat well and only in the image of the forest track the head of the cat is slightly present. In the images recovered by $FastICA2$ however none could be called well separated. The picture with the cat has still the windows that belong to the picture with the sheep and in the picture of the sheep and of the forest track the head of the cat is still visible. The good performance of $E5$ is noteworthy here especially when considering that the images probably do not have underlying symmetric distributions. Using two robust scatter matrices having the IC-property like symmetrized scatter matrices might therefore even improve the result. However the dimension of this example with 16900 observations and three variates is currently too large to apply symmetrized scatter matrices since the resulting large number of pairwise differences is a too huge computational task and hence not feasible.

4 Conclusion

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. However, as the image example shows, ICA problems have easily several thousand observations and therefore this is not feasible yet. To relieve the computational burden, the original estimate may then be re-

Figure 2: ICA for three pictures. The first row shows the original pictures, the second row the mixed pictures including some contamination. The third row used two robust scatter matrices ($E5$) to recover the pictures and the fourth row the $FastICA2$ algorithm.

placed 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

Amari, S., Cichocki, A., and Yang, H. (1996). A new learning algorithm for blind source separation. In *Advances in neural information processing systems 8* (p. 757-763). Cambridge, MA.: MIT Press.

Cardoso, J. (1989). Source separation using higher order moments. In *Proceedings of IEEE international conference on acustics, speech and signal processing* (p. 2109-2112). Glasgow.

Dümbgen, L. (1998). On Tyler's M-functional of scatter in high dimension. *Annals of Institute of Statistical Mathematics*, *50*, 471-491.

Hettmansperger, T. P., and Randles, R. H. (2002). A practical affine equivariant multivariate median. *Biometrika*, *89*, 851–860.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Hyvärinen, A., and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*, 1483-1492.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*, 411-430.

Marchini, J., Heaton, C., and Ripley, B. (2006). fastICA: FastICA algorithms to perform ICA and projection pursuit [Computer software manual]. (R package version 1.1-8)

Maronna, R., Martin, R., and Yohai, V. (2006). *Robust statistics*. Chichester: Wiley.

Nordhausen, K., Oja, H., and Tyler, D. (2006). ICS: ICS / ICA computation based on two scatter matrices [Computer software manual]. (R package version 0.1-2)

Oja, H., Sirkiä, S., and Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, *35*, 175-189.

Ollila, E., Oja, H., and Koivunen, V. (2007). *Complex-valued ICA based on a pair of generalized covariance matrices.* (Conditionally accepted by Computational Statistics & Data Analysis)

R Development Core Team. (2006). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Sirkiä, S., Taskinen, S., and Oja, H. (2007). Symmetrized M-estimators of multivariate scatter. *Journal of Multivariate Analysis*, *98*, 1611-1629.

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, *15*, 234-251.

Authors' Addresses:

Klaus Nordhausen
Tampere School of Public Health
FIN-33014 University of Tampere
Finland
E-mail: klaus.nordhausen@uta.fi

Hannu Oja
Tampere School of Public Health
FIN-33014 University of Tampere
Finland
E-mail: hannu.oja@uta.fi

Esa Ollila
Signal Processing Laboratory
Helsinki University of Technology
P.O. Box 3000
FIN-02015 HUT
Finland
E-mail: esollila@wooster.hut.fi