Component Ordering in Independent Component Analysis Based on Data Power

Anne Hendrikse
University of Twente
Fac. EEMCS, Signals and Systems Group
Hogekamp Building, 7522 NB Enschede
The Netherlands
a.j.hendrikse@ewi.utwente.nl

Raymond Veldhuis
University of Twente
Fac. EEMCS, Signals and Systems Group
Hogekamp Building, 7522 NB Enschede
The Netherlands
r.n.j.veldhuis@ewi.utwente.nl

Luuk Spreeuwers
University of Twente
Fac. EEMCS, Signals and Systems Group
Hogekamp Building, 7522 NB Enschede
The Netherlands
l.j.spreeuwers@ewi.utwente.nl

Abstract

With Independent Component Analysis (ICA) the objective is to separate multidimensional data into independent components. A well known problem in ICA is that since both the independent components and the separation matrix have to be estimated, neither the ordering nor the amplitudes of the components can be determined.

One suggested method for solving these ambiguities in ICA is to measure the data power of a component, which indicates the amount of input data variance explained by an independent component. This method resembles the eigenvalue ordering of principle components. We will demonstrate theoretically and with experiments that strong sources can be estimated with higher accuracy than weak components.

Based on the selection by data power, a method is developed for estimating independent components in high dimensional spaces. A test with synthetic data shows that the new algorithm can provide higher accuracy than the usual PCA dimension reduction.

1 Introduction

Independent component analysis (ICA) is a method to estimate the independent components or sources from which the data is generated. ICA assumes the data is generated by making linear combinations of a number of independent sources, denoted by s. This can be described by:

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

where x denotes the data or the vector of mixtures and $\mathbf{A}$ denotes a mixing matrix. The objective of an ICA algorithm is to reverse the linear mixture by estimating a separation matrix, $\mathbf{W}$, which separates estimates of the independent sources, denoted y, from the mixture.

$$\mathrm{y} = \mathbf{W} \cdot \mathrm{x} \tag{2}$$

Comon [5] shows that the ICA estimate has some ambiguities: any multiplication of the separation matrix with a permutation matrix and a scaling matrix results in another valid ICA estimation. Therefore the source estimates are usually supposed to have unit variance.

The estimation of the separation matrix can be split in two stages. In the first stage the data is whitened. This is often done by Principle Component Analysis (PCA). Since the covariance matrix of the whitened input data and the source estimates are identity matrices, the remaining mixing matrix after whitening has to be an orthogonal matrix [6]. Because of the ambiguities of ICA only a rotation matrix has to be estimated. The estimation of this rotation matrix forms the second stage. Several algorithms have been proposed to estimate this rotation matrix. The general approach is to use a contrast function which achieves an extremum when the whitened data is rotated such that the marginals are independent [4].

In order to solve the ambiguity of the order of the estimated sources, it has been suggested to consider the contrast function, used in ICA estimation [8], for example kurtosis. The argument is that sources which have low contrast, are Gaussian and are thus hard to separated. However, depending on the contrast function used, source distributions get different contrast levels assigned. For example, there are distributions which have a zero kurtosis while being far from Gaussian.

Comon [5] suggested to remove the ambiguity of ICA by ordering the eigenvectors matrix columns, the eigenvalue diagonal and the ICA rotation matrix rows such that the eigenvalues in the eigenvalue matrix are in descending order. The only reason for doing so is to make the components appear in the same order every estimation. In Bayesian ICA, it has also been suggested to use the values of the elements of the estimated mixing matrix to determine whether an estimated source is a real source [2].

In the remainder of the article we will focus on solving the source ordering problem by data power. We will demonstrate that strong sources can be estimated with higher accuracy. Based on this ordering an algorithm will be developed which can provide a solution to overtraining behaviour in high dimensional ICA problems.

1.1 Data power definition

In ICA estimation without Bayesian inference, one possibility is to consider the data power of the components. The variance of the input data can be described by individual contributions of the independent components:

$$\sum_{i=1}^{I} \mathrm{E}\left[x_i^2\right] = \sum_{i=1}^{I} \mathrm{E}\left[(\mathrm{a}_{i\bullet} \cdot \mathrm{s})^2\right] = \sum_{j=1}^{J} \left\{ \mathrm{E}\left[s_j^2\right] \cdot \sum_{i=1}^{I} a_{i,j}^2 \right\} \tag{3}$$

where x is the input data, s represents the independent sources and **A** is the mixing matrix. In equation 3 it can be seen that each component makes an independent contribution to the total variation of the input data. Sources which provide large contributions are considered strong sources and those which provide small contributions are considered weak sources.

There are a few reasons for selecting the components with the highest data power. The procedure is related to the selection based on eigenvalues in PCA, and is actually the same in certain mixtures. It therefore shares some of the arguments for using eigenvalue selection. For example, data power selection provides the best description of the data in independent components in the least squared sense.

2 Analysis on relation estimation accuracy and data power

Another reason to select only the strong sources is that they are more robust to errors in the estimation. The weak sources are very sensitive to errors in the eigenvectors estimation. Recall that a common approach to ICA estimation is to first whiten the data after which an ICA rotation matrix is estimated. The separation matrix can thus be decomposed into:

$$\mathbf{W} = \mathbf{A}'^{T} \cdot \mathbf{D}^{-\frac{1}{2}} \cdot \mathbf{E}^{T} \tag{4}$$

where $\mathbf{E}$ and $\mathbf{D}$ are the eigenvector matrix and the eigenvalue matrix respectively, which can be found by performing PCA on the input data. $\mathbf{A}'$ denotes the ICA rotation matrix.

Data power differences between independent components can only occur if the ICA rotation projects different sources on different diagonal elements of the scaling matrix. This causes the sources to be scaled with different factors. Nadal et al.[9] considered the situation in which the mixing matrix was nearly singular and also noise was present. They use the mutual information between the inputs and the outputs of a neural network to demonstrate that weak sources disturb the estimation of strong sources. We will focus more on the influence of the whitening stage in this error. We will also assume the errors are introduced by limited sample behaviour, instead of noise.

While in [9] only large differences in strength are considered, it still makes sense to select sources with large data power when the differences are small as will be shown next. When the number of data samples is sufficiently large, the matrices in equation 4 can be estimated accurately, but when only a limited number of samples is available, errors are introduced. Small errors in $\mathbf{E}$ have a different effect on strong sources than on weak sources. Consider the 2D mixing problem where a strong and a weak source are mixed with a scaling matrix:

$$\mathrm{x} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\alpha} \end{bmatrix} \cdot \mathrm{s} \tag{5}$$

where $\alpha > 1$. An error in the eigenvector matrix can be represented as an additional rotation of the input data by rotation matrix $\mathbf{R}$ before the separation is performed. Therefore the "whitened" data, denoted by z, is no longer white:

$$\begin{aligned} \mathrm{z} &= \begin{bmatrix} 1 & 0 \\ 0 & \alpha \end{bmatrix} \cdot \begin{bmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\alpha} \end{bmatrix} \cdot \mathrm{s} && (6) \\ &= \begin{bmatrix} \cos\varphi\, s_1 + \frac{\sin\varphi}{\alpha} s_2 \\ \cos\varphi\, s_2 - \alpha \sin\varphi\, s_1 \end{bmatrix} && (7) \end{aligned}$$

Note that crosstalk of source 2 on the estimate of source 1 differs a factor α^2 from the crosstalk of source 1 on the estimate of source 2. The difference between the power caused by the crosstalk is a factor α^4. After whitening an ICA rotation matrix still has to be estimated. Since the data is no longer white, it depends on the specific ICA algorithm chosen what happens with the ICA rotation matrix. Cardoso [3] reported the effects of non white data on some objective functions. He reported that some objective functions acquire an additional correlation term. These objective functions thus attempt to find a rotation which partly minimizes the correlation between the two estimates even though the data is not white.

Besides the limited sample behaviour, another possibility for an incorrect estimate of the eigenvalue rotation matrix is when the estimation is done on a subset of the data.

This is for instance the case when ICA is used to determine features for recognition purposes. ICA is then performed only on training data, which may not accurately represent the entire data set. The ICA rotation can not react on the error in the eigenvector estimate in such situations.

The estimation of sources can also be corrupted by the presence of noise [9]. Learned-Miller [7] already indicated that Gaussian noise filters the probability density functions of the independent components, so they become more Gaussian. Weak sources are more affected than strong sources when equal powered noise is added to every dimension of the observation data, so the estimation of weak sources gets more difficult in the presence of noise and leads to more errors.

3 Experiments with data power

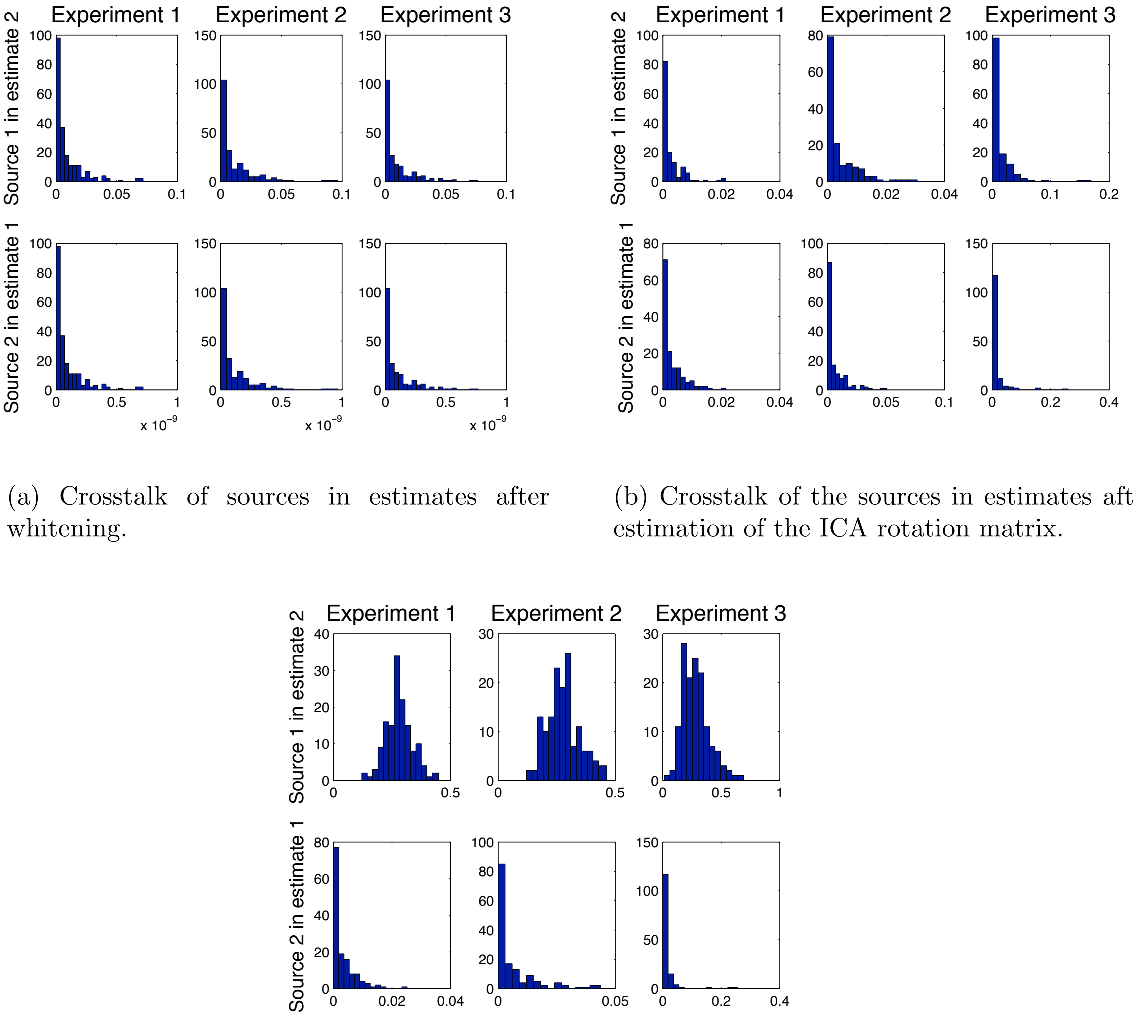
In this section several experiments with synthetic data are described to verify the theory of section 2. To verify the crosstalk between strong and weak sources, an experiment has been performed in which two sources are mixed with a diagonal mixing matrix with a diagonal $[1, 0.01]$. This mixing matrix causes source 1 and 2 to be a strong and a weak source respectively. Three source configurations are used. In experiment 1 both sources are sub Gaussian. In experiment 2 source 1 is super Gaussian and source 2 is sub Gaussian. In experiment 3 both sources are super Gaussian. Sources are sub (super) Gaussian when the kurtosis of the source is negative (positive) [6]. For all tests all sources consist of 100 samples.

The separation of sources is performed by only performing the whitening step, which should be sufficient to separate the sources. According to equation 7 the power of the crosstalk in the two estimates should differ by a factor 10^8. The experiments are repeated a few hundred times. Each time the crosstalk of the sources in the estimates is measured. Histograms of these measurements are plotted in figure 1(a). The horizontal axis in each plot displays the measured crosstalk power. The vertical axis indicates the number of experiments which had an amount of crosstalk as indicated on the horizontal axis. The two plots in each column belong to the same source configuration. The three plots in each row belong to the same source crosstalk in estimate measurement (for example row 1 indicates the crosstalk of source 2 on the estimate of source 1 for the different source configurations). The shapes are the same for every source configuration, but the crosstalk differs indeed a factor 10^8 in the two estimates.

After whitening the strong components have less distortion than the weak components. Next ICA estimates the ICA rotation matrix. In the next experiment the same setup is used as in the previous experiment, but now the ICA rotation is also estimated. This is done using the fastICA algorithm with a tanh nonlinearity function.

In figure 1(b) the histograms of the crosstalk power are given. The ICA rotation matrix in the generating process is an identity matrix, so an accurate estimate of the rotation matrix would result in similar results as in experiment 1. However, the crosstalk between both sources is equal in strength. Apparently ICA reduces the accuracy of the strong source estimate in order to increase the accuracy of the weak source estimate. Nadal et al.[9] also found that the presence of small sources disturbed the estimation of the strong sources when using the infomax criterion. However they assume a clear distinction between weak and strong sources. Since the small components are largely present in the small eigenvalues, this result suggests to remove the smallest principle components before performing ICA.

The second possibility of an error in the estimation of the eigenvectors suggested was in a training-test situation. In figure 1(c) the results are shown of the same test as in the previous experiment, but now an addition rotation of 0.3 degrees is introduced to the mixing matrix after separation estimation. Clearly a bias is introduced in the estimate of the components. The strong component causes on average 30 percent of the power of the weak source estimate, while the weak source is hardly present in the strong source estimate, although the influence difference is not a factor 10^8 as in the first situation.



(a) Crosstalk of sources in estimates after whitening.

(b) Crosstalk of the sources in estimates after estimation of the ICA rotation matrix.

(c) Crosstalk of the sources after a rotation error is introduced to the model matrix of 0.3 degrees.

Figure 1: Histograms of the crosstalk power between two source estimates. Source 1 and 2 have mixing factors 1 and 0.01. Three source configurations are used: both sources sub Gaussian, source 1 super Gaussian and source 2 sub Gaussian and both sources super Gaussian.

4 Application in blocked ICA (blICA)

ICA is known for its overtraining behaviour in high dimensional data with limited sample size. Särelä and Vigário [10] provided some analysis on the subject, especially when kurtosis is used as contrast function. A solution to the problem is to reduce the dimension of the data before ICA is performed. Several authors suggested the use of PCA for this reason [10], [6] chapter 13. Source selection is performed after ICA, so it cannot prevent overtraining in the proposed implementation of section 2.

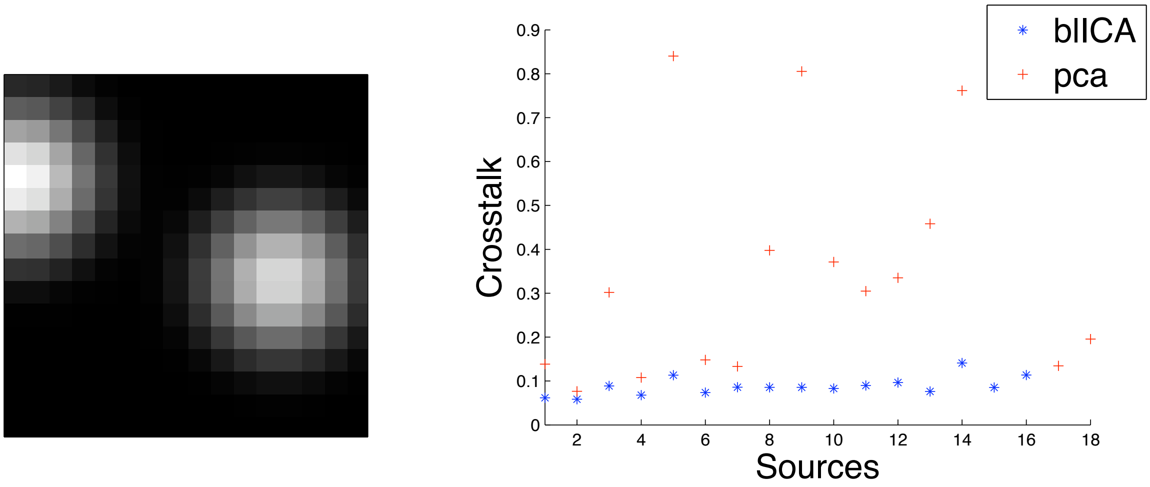
When the number of sources present in a group of dimensions of the data is limited, another possibility exists. Instead of considering all data dimensions at once, ICA can be performed on separate dimension clusters of the data. Consider the problem of performing ICA on image data, in which each image is an observation. Every pixel is considered one dimension, so the number of dimensions is high, while the number of images is in general low. Clusters of dimensions can be formed by cutting the images into blocks of equal size. On each block ICA can be performed. Using the data power criterion, the strongest sources of each block are selected and the rest is discarded. The remaining sources of neighbouring blocks are placed into new blocks on which the same operations are performed. This process is repeated until only one block remains. Attias [1] described such an approach for Bayesian ICA.

The separation process defines one separation matrix. However, an estimation of the mixing matrix is not defined. This also leaves the data power undefined. The mixing matrix elements can be estimated by the cross correlation between the estimated sources and the data components. This definition allows the use of overlapping blocks, for example block two partly contains dimensions already used in block one. This may prevent border effects: when a strong source is present on the border of block one and two it might be rejected in both blocks since its power is only half in both blocks, while it might be retained if it is in the center of one of the blocks.

5 Experiments with blICA

To demonstrate the usefulness of the blICA algorithm, the following experiment has been performed. Synthetic images are created of size 16x16 pixels from 256 white super Gaussian sources, each consisting of 10,000 samples. A mixing matrix is constructed with its elements Gaussian distributed. Next masks are applied to the columns of the mixing matrix. The masks have a 2D gaussian shape with uniform random means. 30 strong masks are generated which have both a larger amplitude and a larger spread value. To get an overlap of strong sources in principle components, each strong mask gets another strong mask added. An example of the resulting strong masks is given in figure 2(a). The resulting mixing matrix is used to mix the 256 sources into the synthetic images.

From the mixture data 16 sources are estimated using both the PCA ICA method and the blICA method. Of the estimated components the amount of crosstalk power is measured. The result is given in figure 2(b). blICA estimated all the components PCA also estimated with higher accuracy, and all sources have about the same amount of crosstalk. The PCA method on the other hand has a few sources which consist mostly of crosstalk. The result gets even worse for PCA when the number of dimensions is reduced to 4. It is very difficult to identify a single source which is estimated, most estimates are best described as a linear combination of sources. The result for blICA is the crosstalk of the first 4 elements in figure 2(b). Thus if the estimated number of dimensions is incorrect, blICA performs better in such mixture configurations.



(a) Example of the mask for strong sources. A strong mask is composed of two Gaussian shapes. Each Gaussian is shared with another source mask.

(b) Comparison between the crosstalk in estimated components of both the ICA with PCA dimension reduction and the blICA dimension reduction. Estimates of the same components are placed on top of each other

Figure 2: blICA versus ICA with PCA results.

6 Conclusion

We proposed to solve the ambiguity of the ordering of ICA components based on the data power of a component. Although the method has been suggested before, we provided some new arguments in favour of the method: weaker sources are more sensitive to variations in the estimate of the eigenvectors in the whitening stage. In regular ICA, this is partly compensated by ICA rotation, which introduces errors on the estimate of strong sources, but in training-test situations weak sources get considerable more crosstalk from strong sources than visa versa. In Nadal et al. [9] also the situation with strong and weak sources is considered, but only with large differences in strength and presence of noise. The training-test situation is not considered at all.

In Nadal et al. [9] it was suggested to remove the small principle components, since they would contain the small sources. However, as they noted, when two strong sources could only be separated using a smaller principle component, the dimension reduction could lead to separation problems. Making a selection of the strongest sources after ICA prevents this problem.

A part of the problem of ICA estimation is that after the whitening, ICA rotation transfers part of the error in the weak estimate to the strong estimate. It might be a good idea to modify the symmetric update, like in FastICA [6], with a strength term, so the strong estimates are less effected by the errors in the weak estimates.

Using the component selection criterion, the blICA algorithm has been developed, which reduces the overtraining behaviour of ICA in the case that the sources are only locally present. In such a situation the experiment showed that blICA can estimate the sources with higher accuracy than the general PCA dimension reduction method.

References

[1] H. Attias. Learning in high dimensions: modular mixture models. In *Proceedings of the 8th International Conference on Artificial Intelligence and Statistics*, pages 144–148, 2001.

[2] C. M. Bishop and N. D. Lawrence. Variational bayesian independent component analysis. Technical report, Computer Laboratory, University of Cambridge, 2000.

[3] J. F. Cardoso. Dependence, correlation and gaussianity in independent component analysis. *The Journal of Machine Learning Research*, 4:1177–1203, 2003.

[4] J.F. Cardoso. Blind signal separation: Statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[5] Pierre Comon. Independent component analysis, a new concept? *Signal Process.*, 36(3):287–314, 1994.

[6] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley and Sons, Inc, 2001.

[7] E. G. Learned-Miller and J. W. Fisher III. Ica using spacings estimates of entropy. *The Journal of Machine Learning Research*, 4(7-8):1271–1295, 2004.

[8] Wei Lu and Jagath C. Rajapakse. Eliminating indeterminacy in ica. *Neurocomputing*, 50:271–290, 2003.

[9] J.-P. Nadal, E. Korutcheva, and F. Aires. Blind source separation in the presence of weak sources. *Neural Networks*, 13(6):589–596, 2000.

[10] J. Särelä and R. Vigário. Overlearning in marginal distribution-based ica: analysis and solutions. *The Journal of Machine Learning Research*, 4(7-8):1447–1469, 2004.