Solution of high-dimensional linear separation problems

F. Herrmann and A. K. Nandi
Dept. of Electrical Engineering and Electronics, The University of Liverpool,
Brownlow Hill, Liverpool L69 3GJ, UK, e-mail: {fherrm,aknandi}@liverpool.ac.uk

ABSTRACT

Blind source separation (BSS) has been one of the emerging research topics within the signal processing community in recent years. Particularly, the maximum squared kurtosis has been found to be a suitable criterion for many technical applications of BSS. Conventionally an elementary Givens rotation estimator is applied to all source pairs in a Jacoby-like algorithm. However, those methods suffer from an escalation of computational expenses as soon as the number of sources becomes large. This paper introduces a novel eigenvector deflation method. It allows the separation of complex and high-dimensional mixtures without such performance penalty.

1 Introduction

1.1 General

Essential to the recovery of sources from a mixture is their independence. This is a very weak limitation, since independence is plausible for the majority of cases, due to their different nature or origin. The mixing of sources is strictly linked to an increase of their entropies due to the additional information cross over. The manner in which this bridging takes place is defined by the underlying model which is often known or can be sufficiently approximated. Natural limitations are found in the accuracy and inversibility of the model. Linear models or respective linear approximations play an exceptional role in this context, because of their mathematical tractability.

1.2 Independence subject to a linear model

A yardstick to measure the degree of mixing of sources is the Kullback discrimination information δ_{KDI} [9] between the joint probability density function (pdf) and an equivalent independent pdf. This divergence is written in terms of entropies $H(f_{X_k})$, characterising the information content of an observation X_k

$$\delta_{\mathrm{KDI}}\left(f_{X_{1:K}}, \prod_{k=1}^{K} f_{X_k}\right) = -H(f_{X_{1:K}}) + \sum_{k=1}^{K} H(f_{X_k}), \quad (1)$$

$$\text{with} \quad H(f_{X_k}) = -\int_{-\infty}^{\infty} f_{X_k}(\xi) \log f_{X_k}(\xi) d\xi. \quad (2)$$

Imposed orthonormal transformations of the random vector ensure that the joint entropy $H(f_{X_{1:K}})$ remains constant and hence the divergence is a function of their individual marginal entropies $H(f_{X_k})$, for $k = 1, 2, \ldots, K$ random elements. This important outcome allows a divergence optimisation within the multiple parameter space of an orthonormal transformation θ, with respect to marginal pdfs only

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{K} H(f_{X_k}). \quad (3)$$

1.3 Maximising non-Gaussianity

Despite the apparently attractive simplicity of the above approach, the difficulty arises in the blind scenario. The lack of prior knowledge makes the description of the source pdf by a conveniently small number of characteristic parameters almost impossible. To circumvent this, the entropy (2) is rewritten in relation to a standardised Gaussian pdf $\alpha(\xi)$

$$H(f_{X_k}) = -\int_{-\infty}^{\infty} f_{X_k}(\xi) \log \alpha(\xi) d\xi - \int_{-\infty}^{\infty} f_{X_k}(\xi) \log \frac{f_{X_k}(\xi)}{\alpha(\xi)} d\xi. \quad (4)$$

The left-hand side integral depends only on the covariance of the pdf, whereas the right-hand side integral in (4) is again a divergence $\delta_{\mathrm{KDI}}(f_{X_k}, \alpha)$

$$H(f_{X_k}) = \frac{1}{2}\Big(\log 2\pi + \int_{-\infty}^{\infty} f_{X_k}(\xi)\xi^2 d\xi\Big) - \delta_{\mathrm{KDI}}(f_{X_k}, \alpha). \quad (5)$$

From here onwards let it be assumed that the variance is kept constant. That corresponds to decorrelation and standardisation of the pdf $f_{X_k}(\xi)$ which is achieved by sphering the observation. Thus the maximisation of the non-negative divergence $\delta_{\mathrm{KDI}}(f_{X_k}, \alpha)$, also implies (3). Indeed the emerging phenomenon has been anticipated as reversal of the central limit theorem of statistics. Unfortunately, the maximisation of $\delta_{\mathrm{KDI}}(f_{X_k}, \alpha)$ eludes easy mathematical evaluation. Alternatively, the

least square (LS) norm

$$\delta_{\mathrm{LS}}(f_{X_k},\alpha)=\int_{-\infty}^{\infty} w(\xi)\Big(f_{X_k}(\xi)-\alpha(\xi)\Big)^2 d\xi \qquad (6)$$

describes similar distance measure between two functions. Here, $w(\xi)$ represents a suitable weight function. Note, that the unknown pdf f_{X_k} has not yet been characterised by statistical parameter inferable from an observation. Respectively the primary aim, the optimisation of the distance $\delta_{\mathrm{LS}}(f_{X_k},\alpha)$, the Gram-Charlier series of Type A [8] derives exactly the required statistical parameter

$$f_{X_k}(\xi)=\sum_{j=0}^{\infty} c_j H_j(\xi)\alpha(\xi), \qquad (7)$$

which are given in terms of independent coefficients c_j. With $w(\xi)=1/\alpha(\xi)$ in (6), the differences over the entire ξ-domain are equalised and homogeneously evaluated for near-Gaussian pdfs. Evaluating (6) becomes an objective function

$$\delta_{\mathrm{LS}}(f_{X_k},\alpha)=\sum_{j=1}^{\infty} j!c_j^2. \qquad (8)$$

It can be shown that the maximisation of any coefficient c_j provides a potentially accurate solution. Nevertheless, the coefficients depend to various degrees on the parameter vector θ. Hence for single squared coefficient maximisation, the choice of the right one matters. The coefficients are given by their cumulant substitutes in standard measure ($\kappa_{kk}=1$) up to fourth order [8]

$$c_0=1,\ c_1=0=c_2=0,\ c_3=\frac{1}{6}\kappa_{kkk},\ c_4=\frac{1}{24}\kappa_{kkkk}. \qquad (9)$$

1.4 Separation criteria

The maximum squared kurtosis criterion (maximising c_4^2) is a good trade off between cumulant estimation effort and estimation outcomes for many practical applications. The cumulants of the source estimates enter the criterion via their kurtosis values — the parameter of the marginal posterior pdfs. The criterion is given by

$$\hat{\theta}=\arg\max_{\theta}\sum_{k=1}^{K}\kappa_{kkkk}^2(\theta), \qquad (10)$$

where θ are the parameters of an orthonormal transformation. Alternatively, a minimum squared cross kurtosis criterion could be used instead. A similar criterion was published in [4].

2 Implementation

2.1 Linear model

The following linear model is used for implementation of an algorithm

$$\hat{\mathbf{x}}_i=\mathbf{B}\mathbf{y}_i=\mathbf{B}\mathbf{A}\mathbf{x}_i. \qquad (11)$$

It consists of the separation matrix $\mathbf{B}$ and the unknown mixing matrix $\mathbf{A}$. The vectors $\mathbf{x}_i$, $\mathbf{y}_i$ and $\hat{\mathbf{x}}_i$ represent the source, the observation and the source estimate at the sampling instants $i=0,1,\ldots,(N-1)$. The constancy of the parameter $\mathbf{A}$ over this interval is crucial to allow cumulant estimation under stationary conditions. The mixture matrix can be decomposed into a Hermite ($\mathbf{H}$) and an unitary ($\mathbf{U}$) matrix $\mathbf{A}=\mathbf{H}\mathbf{U}=\mathbf{V}\mathbf{S}^2\mathbf{V}^H$, where $\mathbf{H}$ can be further decomposed into the unitary matrix $\mathbf{V}$ and the diagonal matrix $\mathbf{S}$. The parameter of the Hermite matrix can be estimated independently and the observation is sphered $\tilde{\mathbf{y}}_i=\mathbf{S}^{-1}\mathbf{V}^H\mathbf{y}_i$. This approach reduces the number of remaining parameters to the ones of the now unitary mixture matrix $\tilde{\mathbf{A}}=\mathbf{V}^H\mathbf{U}$ and its inverse, the separation matrix $\tilde{B}=\tilde{\mathbf{A}}^{-1}$.

2.2 The eigenvector deflation algorithm (EDA)

As mentioned before, a sound first step towards independence is the sphering (removal of second order correlations $\kappa_{kl}(\boldsymbol{\xi})\stackrel{\mathrm{def}}{=}\mathrm{cum}\,(\xi_k^*,\xi_l)$ from the observation) in order to obtain unit variance $\kappa_{kl}(\tilde{\mathbf{y}})=1$ for $k=l$ and 2nd-order decorrelation $\kappa_{kl}(\tilde{\mathbf{y}})=0$ for $k\neq l$. The cumulants of the random variables in (11) obey a multi-linear relationship

$$\kappa_{klmn}(\hat{\mathbf{x}})=\sum_{o,p,q,r} b_{ko}^* b_{lp} b_{mq}^* b_{nr}\kappa_{opqr}(\mathbf{y}) \qquad (12)$$

with the cumulants $\kappa_{klmn}(\boldsymbol{\xi})\stackrel{\mathrm{def}}{=}\mathrm{cum}\,(\xi_k^*,\xi_l,\xi_m^*,\xi_n)$ and b_{ko} being the entry in the kth row and oth column of the separation matrix $\mathbf{B}$.

To recover a single source, the criterion (10), can be downgraded to a simple kurtosis optimisation (instead a sum squared kurtosis maximisation, see also MaxKurt in [2])

$$\kappa_{kkkk}(\hat{\mathbf{x}})=\mathbf{b}_k^H\mathbf{Q}\mathbf{b}_k,\qquad q_{op}=\sum_{q,r} b_{kq}^* b_{kr}\kappa_{opqr}. \qquad (13)$$

with the "quadricovariance" matrix $\mathbf{Q}=(q_{op})$ [3]. The introduced linearisation in (13) leads to mathematical tractability by the price of giving up a close form solution. In consequence an iterative approach has to be applied [7], where $\mathbf{Q}$ remains fixed throughout optimisation and must be newly estimated for each iteration step. Since the only changes affect the symmetric tensor $\mathbf{T}=\mathbf{b}\mathbf{b}^H$, a relatively low cost estimate is obtained by

$$\mathbf{Q}=\left\langle \mathbf{y}^*\mathbf{y}^T\mathbf{T}\mathbf{y}^*\mathbf{y}^T\right\rangle-\mathbf{R}\mathbf{T}\mathbf{R}^H-\mathbf{C}\mathbf{T}\mathbf{C}^H-\mathbf{R}\,\mathrm{trace}\Big(\mathbf{T}\mathbf{R}\Big) \qquad (14)$$

with the covariance matrices $\mathbf{R}=\left\langle \mathbf{y}\mathbf{y}^H\right\rangle$ and $\mathbf{C}=\left\langle \mathbf{y}\mathbf{y}^T\right\rangle$. Note that for the sphered observation the covariance is an identity matrix and for circular distributed sources $\mathbf{C}$ completely vanishes. However optimisation and sphering could be done in a single step and this does not necessarily apply. The procedure to solve the BSS problem becomes obvious with (13). The row vector $\mathbf{b}$ of

the separation matrix $\mathbf{B}$ can be extracted as the dominant eigenvector of the symmetric tensor $\mathbf{T}$ from the quadricovariance matrix $\tilde{\mathbf{Q}}$ of the sphered observation $\tilde{\mathbf{y}}$

$$\lambda_k \tilde{\mathbf{b}}_k = \tilde{\mathbf{Q}} \tilde{\mathbf{b}}_k, \tag{15}$$

or $\mathbf{Q}\mathbf{b}_k = \lambda_k \mathbf{R}\mathbf{b}_k$ for the raw observation without preceding sphering. The eigenvector associated with the maximum squared eigenvalue $\max_k |\lambda_k|$ is the desired solution. This task of optimisation and updating $\mathbf{Q}$ has to be repeated, until changes of $\mathbf{b}_k$ are negligible and $\lambda_k = \kappa_{kkkk}(\hat{\mathbf{x}}) \rightarrow \kappa_{kkkk}(\tilde{\mathbf{x}})$. As simulation results indicate, this happens after not more than 2 ... 5 iterations. After successful extraction of the first source, the remaining $(K-1)$ eigenvectors are used to generate the new observations, where the formerly recovered source is removed from (deflation). The method is subsequently applied to the remaining modified observations in the same manner.

2.3 Algorithm summary

1) Sphere observation $\tilde{\mathbf{y}}_i = \mathbf{S}^{-1}\mathbf{V}^H \mathbf{y}_1$.

For $k = 1, 2, \ldots, (K-1)$ do:

2) Initialise a vector $\tilde{\mathbf{b}}$ with one as the first entry and zeros elsewhere.
3) Estimate quadricovariance $\mathbf{Q}$ (14) for the tensor $\mathbf{T} = \tilde{\mathbf{b}}_k \tilde{\mathbf{b}}_k^H$ with the row vector $\mathbf{b}_k$ of the separation matrix $\tilde{\mathbf{B}}$.
4) Extract the dominant eigenvector $\mathbf{e}$ from $\mathbf{Q}$, set $\tilde{\mathbf{b}}_k = \mathbf{e}$, and go to step (3) until changes of $\tilde{\mathbf{b}}_k$ are negligible.
5) Estimate source vector $\hat{\mathbf{x}} = \mathbf{E}\tilde{\mathbf{y}}$, where $\mathbf{E}$ is the eigenvector matrix, ordered by decreasing dominance. Separate the first vector element $\hat{x}_1$ of $\hat{\mathbf{x}}$ as the first source estimate and let the remaining elements be the new observations.

end.

Accordingly an orthonormal separation matrix $\tilde{\mathbf{B}}$, initialised with an identity matrix, can be updated within the loop. As the last step the separation matrix is obtained from $\mathbf{B} = \tilde{\mathbf{B}}\mathbf{S}^{-1}\mathbf{V}^H$.

3 Simulations

3.1 Simulation setup

The algorithm is tested with binary distributed sources with a kurtosis $\kappa_{kkkk}(\tilde{\mathbf{x}}) = 1$. The mixture matrix is given in Fig. 1. The separation success is measured by the performance index ϵ [1] (also cross-talking error [11])

$$\epsilon = \sum_{i=1}^{K} \sum_{j=1}^{K} \left(\frac{|p_{ij}|}{\max_k |p_{ik}|} - 1 \right) + \sum_{i=1}^{K} \sum_{j=1}^{K} \left(\frac{|p_{ij}|}{\max_k |p_{kj}|} - 1 \right) \tag{16}$$

with the transition matrix $(p_{ij}) = P = \mathbf{B}\mathbf{A}$. Its small value, ideally zero, earmarks a successful source separation unaffected by granted permutations or scalings of the sources. The product of number of observations and independent runs is kept constant at $200,000$. A white uniformly distributed noise of given relative noise power is added to the observation.

3.2 Simulation results

Fig. 2 shows the performance index depending on the number of sources. Because of the quadratically increasing number of matrix elements the performance index is normalised $\epsilon_K = \epsilon / K^2$. Apparently the number of sources does not substantially effect the outcomes. Additive noise, as in Fig. 3 affects the estimation performance only little up to a certain power level. Beyond that, the algorithm suddenly fails to provide appropriate estimates. This noise power level is about 10dB below the accumulated source power in the shown simulations. By increasing the number of observations to reduce cumulant fluctuations the failing threshold does not change, but it significantly determines the performance index as it can be seen in Fig. 4. Nonetheless the achieved benefits are relatively small compared to the increased effort for excessively large block lengths.

3.3 Computations

To illustrate the performance in comparison with well-established algorithms, table 1 shows the achieved performance index as well as the number of MATLAB floating point operations (Flops). For this example 5000 observations of a mixture of 5 independent identically distributed sources were used.

4 Summary and outlook

The proposed algorithm manifests its benefits in terms of efficiency for a large number of mixed source signals compared to conventional pairwise approaches. The deflation recovers the sources in order of decreasing normalised kurtosis. This comes along with a often called "interestingness in a statistical sense" but it also is a welcome behaviour, since in practice the number of sources is not as well defined as in a simulation environment and the progress can be stopped at any stage. The estimation performance is apparently not substantially depending on the number of sources. Additive noise can be tackled up to a moderate power level.

5 Acknowledgements

Authors would like to thank The University of Liverpool for the financial support of this research.

6 References

[1] S. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems 8*, 1996.

[2] J. Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157–192, Jan 1999.

Table 1: Performance comparison

algorithm	perf. ind. ϵ	Mega Flops
JADE [3]	.31	8.54
EDA	.42	5.57
MaSSFOC [5]	.31	10.07
MaxKurt [2]	.26	59.98
FastICA [6]	.39	5.03
EASI [10]	1.75	3.30

$$\begin{bmatrix} -.12 & .87 & -.73 & -.16 & -.41 & -.25 & -.61 & .25 & .43 & -.77 \\ -.32 & -.47 & .64 & .71 & -.90 & -.98 & .81 & .40 & .02 & .33 \\ -.37 & -.68 & -.14 & -.02 & .39 & -.16 & .14 & -.21 & .55 & -.27 \\ -.27 & .75 & .78 & .63 & .30 & .51 & .26 & -.17 & -.02 & -.72 \\ -.21 & -.52 & .47 & -.08 & .97 & .59 & -.53 & .31 & -.63 & .13 \\ .18 & .29 & .37 & -.09 & .11 & .84 & .10 & .68 & .40 & .65 \\ -.76 & .93 & -.31 & -.10 & -.20 & .69 & .86 & -.26 & .97 & .35 \\ -.92 & .33 & -.67 & -.18 & -.60 & -.26 & -.33 & -.15 & .61 & 1.00 \\ -.08 & .74 & -.69 & .80 & .25 & .24 & .31 & .19 & .41 & .92 \\ .74 & -.98 & -.62 & -.99 & .47 & .46 & -.22 & .13 & -.03 & -.88 \end{bmatrix}$$

Figure 1: Mixing matrix

[3] J. Cardoso and A. Souloumiac. An efficient technique for the blind separation of complex sources. In *Proceedings of the IEEE SP Workshop on Higher-Order Statatistics*, 1993.

[4] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[5] F. Herrmann and A. K. Nandi. Blind separation of linear instantaneous mixtures using closed-form estimators. submitted to IEEE Signal Processing.

[6] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, NN-10(3):626 – 634, May 1999.

[7] B. Jellonek, D. Boss, and K. D. Kammeyer. Generalised eigenvector algorithm for blind equalization. *Signal Processing*, 61(3):237–264, 1997.

[8] S. M. Kendall and A. Stuart. *The advanced theory of statistics*, volume 1. Charles Griffin & Company Limited, 4th edition, 1977.

[9] S. Kullback, J. C. Kegel, and J. H. Kullback. *Topics in statistical information theory*. Number 42 in Lecture Notes in Statistics. Springer Verlag Berlin Heidelberg, 1987.

[10] B. Laheld and J. Cardoso. Adaptive source separation with uniform performance. In *Proc. EUSIPCO'94*, pages 183–186, Edinburgh, Sep. 1994.

[11] H. H. Yang and S. Amari. Adaptive on-line learning algorithms for blind separation - maximum entropy and minimum mutual information. *Neural Computation*, (9):1457–1482, 1997.

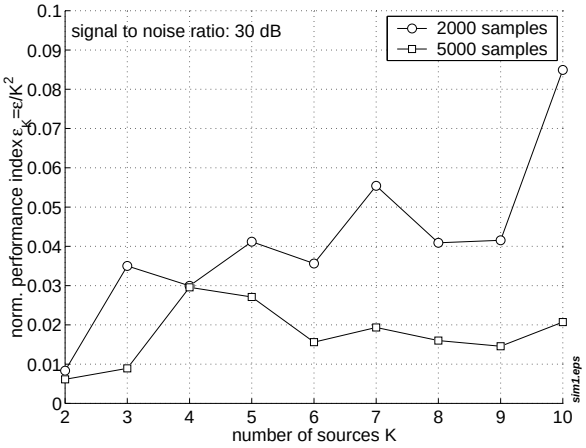

Figure 2: Number of sources

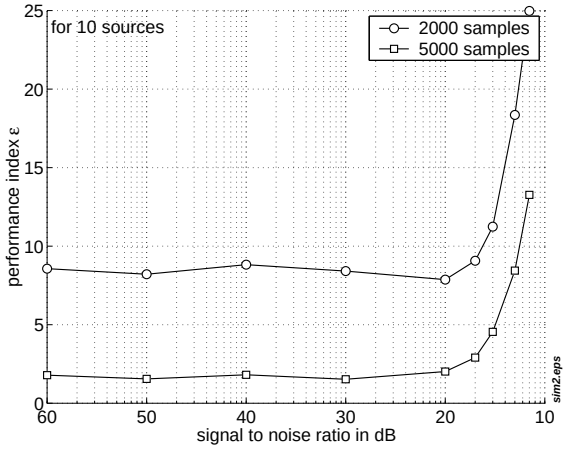

Figure 3: Influence of observation noise

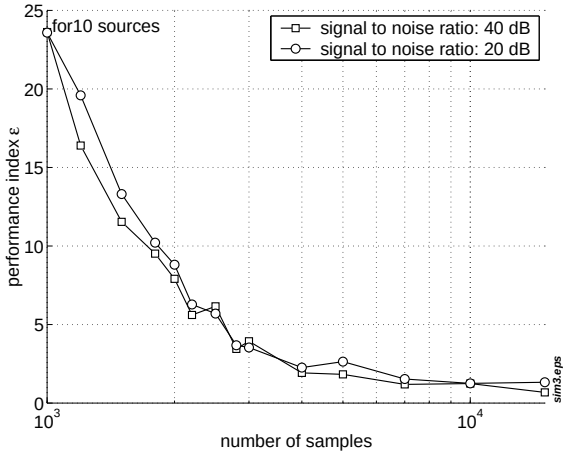

Figure 4: Number of samples