MAXIMUM LIKELIHOOD ESTIMATION OF ICA MODEL FOR WIDE CLASS OF SOURCE DISTRIBUTIONS

Juha Karvanen*, Jan Eriksson* and Visa Koivunen
Signal Processing Laboratory
Helsinki University of Technology
P.O. Box 3000, FIN-02015 HUT, Finland
{juha.karvanen, jan.eriksson, visa.koivunen}@hut.fi

Abstract. We propose two blind source separation techniques that are applicable to a wide class of source distributions that may also be skewed and may even have zero kurtosis. Skewed distributions are encountered in many important application areas such as communications and biomedical signal processing. The methods are based on maximum likelihood approach where source distributions are modeled adaptively by the Pearson system and the Extended Generalized Lambda Distribution (EGLD). To compare the developed methods with the existing methods, quantitative measures for the quality of separation are used. Simulation experiments demonstrate the good performance of proposed methods in the cases where the standard BSS methods perform poorly.

INTRODUCTION

This paper addresses the problem of Blind Source Separation (BSS) in the case where sources may have skewed distributions. We propose new signal processing algorithms that can reliable separate also such sources. Skewed source distributions occur e.g. in biomedical, telecommunications and image data. As an example, fading in communications channels is often modeled using Rayleigh and log-normal distributions. Many widely used Independent Component Analysis (ICA) algorithms explicitly or implicitly assume that sources have symmetric distributions [8, 4, 11]. In some cases this assumption may lead to poor results if any of the sources have asymmetric distributions. In practical applications it is desirable to have an algorithm that performs reliably even if the source distributions are skewed.

0* This work was funded by the Academy of Finland. Two first authors have equally contributed to the technical content of the paper.

We consider the ICA model with instantaneous mixing

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where the sources $\mathbf{s} = [s_1, s_2, \ldots, s_m]^T$ are mutually independent random variables and $\mathbf{A}_{m \times m}$ is an unknown invertible mixing matrix. The goal is to find only from the observations, $\mathbf{x}$, a matrix $\mathbf{W}$ such that the output

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

is an estimate of the possibly scaled and permutated source vector $\mathbf{s}$.

The main problem in BSS is that the underlying source distributions are not known. Many approaches including infomax, maximum likelihood and non-linear approximations of negentropy have been proposed. All these principles lead to the same theoretical measure of independence, which is mutual information between the joint distribution and the product of marginal distributions [6]. In the traditional ICA methods where the criterion function includes a fixed non-linearity like kurtosis or the logarithm of hyperbolic cosine, there is an implicit assumption about the model for source distributions [16, 11]. Generally, most of BSS methods are not highly sensitive to model misfit [2] but certain deviations such as asymmetric source distributions with small absolute value of kurtosis, may however degrade the performance of these methods significantly as illustrated in this paper.

In the maximum likelihood estimation source distributions are modeled explicitly. However, we usually do not have any a priori information on the functional form of the source distributions. In source distribution adaptive maximum likelihood estimation we estimate the parameters of source distributions from data. To achieve good models for sources we need a large parametric families of distributions with free parameters to estimate. In the proposed BSS methods the Pearson system [19] and the Extended Generalized Lambda Distribution (EGLD) [14] are used as model families for source distributions. The corresponding BSS methods, called the Pearson-ICA and the EGLD-ICA, can separate wide class of source distributions and what makes the methods especially interesting, they also successfully separate of zero kurtosis sources.

The paper is organized as follows: In Section 2 BSS methods based on the EGLD and the Pearson system are presented and some practical improvements are presented. Benefits and problems of source distribution adaptive methods are discussed. In Section 3 quantitative measures for quality of separation are presented. Finally, in Section 4 source distribution adaptive Maximum Likelihood methods are compared to some widely used BSS methods in simulation. The results demonstrate the reliable performance of the proposed methods even for skewed distributions.

THE PEARSON SYSTEM AND THE EGLD BASED SCORE FUNCTIONS FOR ICA

We have recently introduced two new source distribution adaptive BSS methods called EGLD-ICA [9] and Pearson-ICA [15] employing the EGLD and the Pearson system, respectively. The methods are based on maximum likelihood approach where the score functions of source distributions are chosen as ICA estimating functions. We hypothesize that source distribution can be modeled by a large parametric family of distributions. Contrary to the standard ICA methods, the parameters of source distributions are not fixed a priori but estimated adaptively for every step of algorithm. Consequently, the score functions used as ICA estimating functions change in every step.

The essential question is the selection of the family of distributions. In [4] t-distribution and generalized Gaussian distribution are used as models. These both are symmetric distributions. The families of distributions, the Pearson system and the EGLD, that we propose as models cover significantly larger family of distributions including asymmetric distributions. Due to this adaptation to asymmetric sources, the Pearson-ICA and the EGLD-ICA can separate non-Gaussian sources with zero kurtosis.

The Pearson system is defined by the differential equation [19]

$$f^{'}(x) = \frac{(x-a)f(x)}{b_0 + b_1 x + b_2 x^2}, \tag{3}$$

where a, b_0, b_1 and b_2 are the parameters of the distribution. The score function of the Pearson system is easily solved from equation (3)

$$\varphi_{Pearson}(x) = -\frac{f^{'}(x)}{f(x)} = -\frac{(x-a)}{b_0 + b_1 x + b_2 x^2}. \tag{4}$$

The simplicity of the score makes the Pearson system particularly appealing for ICA. For the derivative of the score function we have

$$\varphi^{'}_{Pearson}(x) = -\frac{b_0 + a\, b_1 + 2\, a\, x\, b_2 - x^2\, b_2}{(b_0 + b_1 x + b_2 x^2)^2}. \tag{5}$$

The parameters a, b_0, b_1 and b_2 may be estimated directly by the method of moments.

The EGLD consist of two parts: the Generalized Lambda Distribution (GLD) and the Generalized Beta Distribution (GBD). When the EGLD is fitted to data the choice between the GLD and the GBD is made based on the values of kurtosis and skewness [9]. The GLD is defined by the inverse distribution function

$$F^{-1}(p) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, \tag{6}$$

where $0 \leq p \leq 1$ and λ_1, λ_2, λ_3 and λ_4 are the parameters of the distribution. It is not possible to present density or distribution functions of GLD in closed form in the general case. On the other hand, observations are easily generated from the GLD employing the inverse distribution function. Estimation of GLD parameters using method of moments is proposed in [14] and the tables needed in the estimation are provided in [13]. Since the density function of the GLD is not available in a closed form, we have to derive the score function from the inverse distribution function (6). If we write

$$p = F(y), \tag{7}$$

where $F(y)$ is the distribution function of a GLD, the following formula is obtained for the score function:

$$\varphi_{GLD}(p) = -\frac{\lambda_2 p^{\lambda_3 - 2}(\lambda_3 - 1)\lambda_3 + \lambda_2(1-p)^{\lambda_4 - 2}(\lambda_4 - 1)\lambda_4}{(p^{\lambda_3 - 1}\lambda_3 + (1-p)^{\lambda_4 - 1}\lambda_4)^2} \tag{8}$$

In some algorithms the derivative of the score function

$$\begin{aligned} \varphi'_{GLD}(p) = {} & \frac{2{\lambda_2}^2\left(p^{\lambda_3-2}(\lambda_3-1)\lambda_3 - (1-p)^{\lambda_4-2}(\lambda_4-1)\lambda_4\right)^2}{\left(p^{\lambda_3-1}\lambda_3 + (1-p)^{\lambda_4-1}\lambda_4\right)^4} \\ & - \frac{{\lambda_2}^2 p^{\lambda_3-3}(\lambda_3-2)(\lambda_3-1)\lambda_3 - {\lambda_2}^2(1-p)^{\lambda_4-3}(\lambda_4-2)(\lambda_4-1)\lambda_4}{\left(p^{\lambda_3-1}\lambda_3 + (1-p)^{\lambda_4-1}\lambda_4\right)^3} \end{aligned} \tag{9}$$

is also needed.

The other part, the GBD, is characterized by the density function

$$f(x) = C\beta_2^{-(\beta_3+\beta_4+1)}(x-\beta_1)^{\beta_3}(\beta_1+\beta_2-x)^{\beta_4} \tag{10}$$

on the interval $[\beta_1, \beta_1 + \beta_2]$ and zero elsewhere, where β_1 β_2 β_3 and β_4 are the parameters of the distribution and C is a constant. The score function for the GBD is obtained by the straightforward differentiation directly from (10), and is given by

$$\varphi_{GBD}(y) = -\frac{\beta_3}{y-\beta_1} + \frac{\beta_4}{\beta_1+\beta_2-y}, \tag{11}$$

For the derivative of score function we obtain

$$\varphi'_{GBD}(y) = \frac{\beta_3}{(y-\beta_1)^2} + \frac{\beta_4}{(-y+\beta_1+\beta_2)^2}. \tag{12}$$

The maximization of the likelihood criterion may be carried out by gradient methods. The choice of the algorithm is generally independent on the choice models for source distributions. Fast convergence is achieved, for example when fixed point algorithm [11, 12]

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \mathbf{D}\left(E\{\varphi(\mathbf{y})\mathbf{y}^T\} - \mathrm{diag}(E\{\varphi(y_i)y_i\})\right)\mathbf{W}_k, \tag{13}$$

where $D = \mathrm{diag}\left(1/(E\{\varphi(y_i)y_i\} - E\{\varphi'(y_i)\}\right)$ is employed.

A procedure for the source distribution adaptive BSS method based on maximum likelihood approach and the third and fourth sample moments may be given as follows:
For whitened data repeat until convergence [1]

1. Calculate the third and fourth sample moments α_3 and α_4 for current data $\mathbf{y}_k = \mathbf{W}_k\mathbf{x}$ and select suitable model (e.g. in the case of the EGLD make the choice between the GLD and the GBD).

2. Estimate parameters for the model by using the method of moments and calculate the scores $\varphi(\mathbf{y}_k)$.

3. Calculate the separating matrix $\mathbf{W}_{k+1}$ using a gradient algorithm [1, 7] or fixed point (13) algorithm.

The stability of optimization algorithms in source distribution adaptive maximum likelihood approach remains an open problem. The complexity of the EGLD and the Pearson system implies that analytical results are hard to obtain. However, stability problems are seldom encountered in practice. It is possible to improve the speed and stability of the optimization algorithm using combination of adaptive score functions and fixed non-linearities. For Pearson-ICA this kind of heuristic improvement was already introduced in [15] where tanh non-linearity was used for clearly sub-Gaussian or clearly super-Gaussian sources. The idea is use fixed non-linearities when kurtosis is the dominating property and adaptive score functions in vicinity of Gaussian distribution. The tanh non-linearity can be employed also in EGLD-ICA for clearly sub-Gaussian sources but according to our experiments even better results are achieved when all sub-Gaussian distributions are fitted to the nearest GLD with the same skewness. This version of EGLD-ICA where only the GLD part of the EGLD is exploited is used in the simulations of this paper.

MEASURING THE QUALITY OF SEPARATION

In order to compare the performance of ICA algorithms we need quantitative measures that describe the quality of separation. The intuitive definition of quality is here that the extracted signals are as close to the original signals as possible. Hence, the quality of separation can be evaluated only in situation where the original source signals are known. The results are averaged over a large number of realizations. Simulation studies are a practical way study the performance on the ICA algorithms.

The quality of the separation may be evaluated by comparing estimated matrices to known true matrices or by comparing estimated source signals

[1] The convergence criterion can be any suitable for the gradient algorithms and the fixed-point algorithm respectively. In our experiments, we used a criterion similar to the symmetric FastICA[10] with $\varepsilon = 0.0001$.

to true source signals. The comparison of matrices may be done using matrix norms and measures for matrix perturbations [18]. In ICA literature Performance Index [3] is used as permutations invariant matrix measure. An alternative way to do the comparison is calculate the angles between row vectors of the true separating matrix and row vectors of the estimated separating matrix. Since it is possible to solve the separating matrix only up to some permutation of rows we have to match the true row vectors with the obtained row vectors so that the angles are as close to zero as possible. Due to the indeterminacy of signs of obtained row vectors the angles are transformed to the interval $[0, \frac{\pi}{2}]$.

In the signal comparison the difference between original signals and extracted signals is computed. Here again, the scaling and permutation differences must be eliminated prior to comparison. For signal comparison we used Signal to Inference Ratio (SIR) (SIR(dB)= $-10\log_{10}(\mathrm{MSE})$) where MSE stands for Mean Square Error MSE= $E\left\{(s(t)-y(t))^2\right\}$. To eliminate scaling differences both original signals and extracted signals are standardized or whitened to have zero mean and unit variance before the calculation of the SIR. After that source signals are matched to the extracted signals so that the resulting MSE values are as small as possible. This may be time consuming operation if the number of signals is large. Even if the separation has failed the matching gives a permutation that is the least poor in the sense of the SIR values.

SIMULATION EXAMPLE

In our simulation example we demonstrate how the performance of the standard algorithms significantly decreases when source distributions are skewed. In [15] and [9] it was demonstrated that Pearson-ICA and EGLD-ICA can separate skewed zero-kurtosis sources that generally are unachievable for kurtosis based method. Now we show that skewness information may be essential also in situation where sources are sub-Gaussian or super-Gaussian.

In our simulation we try to separate six Rayleigh distributed sources. The Rayleigh distribution is commonly used for modeling the fading communication channels. The theoretical cumulants of Rayleigh(1) distribution are mean $= \sqrt{\frac{\pi}{2}} \approx 1.253$, variance $= 2 - \frac{\pi}{2} \approx 0.429$, skewness $= \left((-3+\pi)\sqrt{\frac{\pi}{2}}\right)/\left(2-\frac{\pi}{2}\right)^{\frac{3}{2}} \approx 0.631$ and kurtosis $= \left(32 - 3\pi^2\right)/(4-\pi)^2 \approx 3.245$. Thus, Rayleigh distribution is regarded as super-Gaussian.

The algorithms in comparison are Pearson-ICA, EGLD-ICA, JADE [5], Extended infomax [17] and FastICA [10]. In FastICA there are three different contrast functions are used with symmetric approach. In simulation the length of signals is 5000 and 1001 realizations are generated. For each realization, a random full-rank mixing matrix is generated.

Boxplots are used to describe the distributions of measures of quality for each method. Boxplot is graphical presentation tool for sample distributions and it is widely utilized in applied statistics. In a boxplot the box defines

Algorithm	Median	Maximum
Pearson-ICA	0.68	5.09
EGLD-ICA	1.03	3.30
FastICA(Pow3)	10.04	22.36
FastICA(Tanh)	21.17	30.46
FastICA(Gauss)	18.61	30.12
JADE	8.19	22.60
Extended Infomax	19.48	29.31

Table 1: The median and maximum Performance Indexes for the tested algorithms in simulation where 1001 realizations of six Rayleigh distributed source signals of length 5000 were mixed. In the case of six signals the values of Performance Index belongs to the interval $[0, 60]$, zero value corresponding to the perfect separation. The maximums of Performance Index tell about the reliability of the Pearson-ICA and the EGLD-ICA: even the maximum values are smaller than medians of the other algorithms.

the quintile range (from 25% percentile to 75% percentile). The line inside the box is the median. The 'whiskers' are lines extending from each end of the box to show the extent of the rest of the data. The length of a whisker is defined as 1.5 times the length of the quintile range. Nevertheless, the whiskers are always bound by sample minimum and maximum. The possible outliers, outside the whiskers area, are marked by crosses.

The results are shown in Figures 1 and 2 and in Table 1. In Figure 1 boxplots of SIR-values are presented. It is immediately seen that the Pearson and the EGLD-ICA outperform the other methods. The median SIR-values of the Pearson and the EGLD-ICA are 30.40 dB and 26.99 dB, respectively. The median values for FastICA(Pow3) and JADE are 9.21 dB and 9.89 dB. The median SIR-values of FastICA(Tanh), FastICA(Gauss) and Extended Infomax are under 4 dB. In Figure 2 boxplots of the angles between row vectors of the true separating matrix and row vectors of the estimated separating matrix are presented for row vectors corresponding the first source signal. The angles belongs to the interval $[0, \frac{\pi}{2}]$, zero angle corresponding to the perfect separation. Conclusions are similar to Figure 1.

In this simulation the median performance of the Pearson-ICA is slightly better than the median performance of the EGLD-ICA. This is reasonable since Rayleigh distribution is actually a special case of Gamma distribution and thus belongs to Pearson system. In other words, the score functions of the Pearson-ICA can adapt perfectly to the score function of Rayleigh distribution. However, in this simulation the EGLD-ICA seems to be more stable: the minimum SIR-value of the EGLD-ICA is 16.70 dB whereas the minimum SIR-value of the Pearson-ICA is 9.91 dB.

In Table 1 Performance Indexes are presented. In the case of six signals the values of Performance Index belongs to the interval $[0, 60]$, zero value corresponding to the perfect separation. Median Performance Indexes summarize the overall performance. Small maximum values of Performance Index indicate reliability of the algorithms. It is seen that even the maximum values for the Pearson-ICA and EGLD-ICA are smaller than the median values of

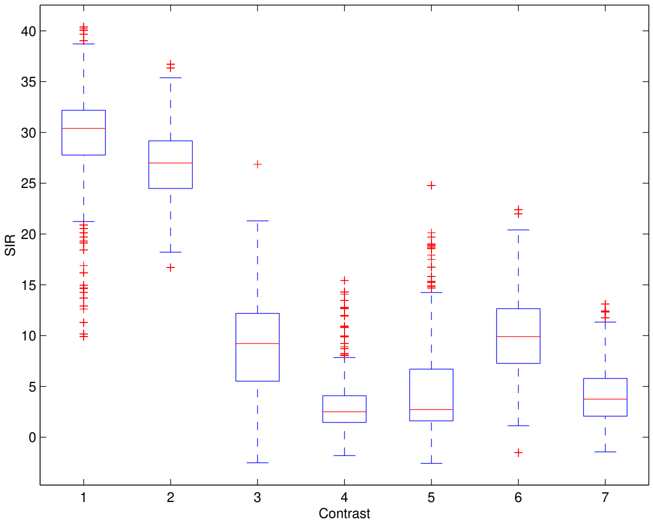

Figure 1: Boxplot of the SIR values of Pearson-ICA (boxplot number 1), EGLD-ICA (2), FastICA (contrasts 'Pow3' (3), 'Tanh' (4) and 'Gauss' (5)) , JADE (6) and Extended infomax (7) algorithm. Six Rayleigh distributed source signals of length 5000 were mixed. The number of realizations was 1001. The presented SIR-values are the SIR-values between the first source signal and its estimate; the SIR-values of the other source signals are similar. The Pearson-ICA and the EGLD-ICA, which exploit also skewness, perform very well: median SIR-values are 30.40 dB and 26.99 dB, respectively. The median values for FastICA(Pow3) and JADE are 9.22 dB and 9.89 dB. The median SIR-values of FastICA(Tanh), FastICA(Gauss) and Extended Infomax are under 4 dB.

other methods.

It is worth of noticing that in this simulation all sources are theoretically separable by using only the fourth order information. Although the sources are asymmetrically distributed, the fourth cumulants differ from zero. Nevertheless, this simulation example shows that in practice the use of skewness information may significantly enhance the quality of separation.

CONCLUSION

Two maximum likelihood methods for blind separation are proposed. The methods are based on adaptive estimation of source distributions using the Pearson system and the Extended Generalized Lambda Distribution. Quantitative measures for the quality of separation are used in simulation to compare performance of the proposed methods. Rayleigh distributed source signals are used in simulation examples. The results show that the performance of the Pearson-ICA and the EGLD-ICA is significantly better than the performance of some other widely used ICA methods when the skewness is the dominant property of source distributions.

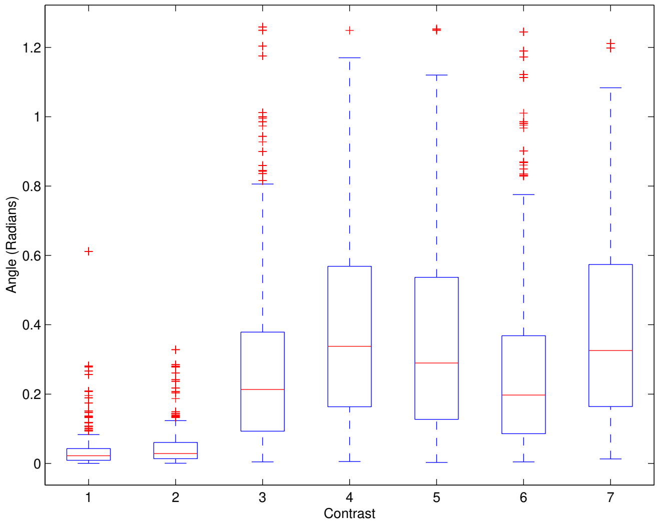

Figure 2: Boxplot of the angles between row vectors of the true separating matrix and row vectors of the estimated separating matrix error performance of Pearson-ICA (boxplot number 1), EGLD-ICA (2), FastICA (contrasts 'Pow3' (3), 'Tanh' (4) and 'Gauss' (5)) , JADE (6) and Extended infomax (7) algorithm. Six Rayleigh distributed source signals of length 5000 were mixed. The number of realizations was 1001. The presented angles are radians between the first source signal and its estimate; zero angle corresponds to perfect separation.

REFERENCES

[1] S.-I. Amari, "Natural Gradient Works Efficiently in Learning," **Neural Computation**, vol. 10, pp. 251–276, 1998.

[2] S.-I. Amari and A. Cichocki, "Adaptive Blind Signal Processing – Neural Network Approaches," **Proceedings of the IEEE**, vol. 86, no. 10, pp. 2026–2048, 1998.

[3] S.-I. Amari, A. Cichocki and H. Yang, "A New Learning Algorithm for Blind Signal Separation," in **Advances in Neural Information Processing Systems**, Cambridge MA: MIT Press, vol. 8, pp. 757–763, 1996.

[4] J. Cao and N. Murata, "A Stable and Robust ICA Algorithm Based on t-distribution and Generalized Gaussian Distribution Models," **Neural Networks for Signal Processing IX, 1999. The IEEE 1999 Proceedings.**, pp. 283–292, 1999.

[5] J. F. Cardoso, "JADE Matlab code with references available at http://sig.enst.fr:80/~cardoso/," .

[6] J. F. Cardoso, "Infomax and Maximum Likelihood for Blind Signal Processing," **IEEE Signal Processing Letters**, vol. 4, no. 4, April 1997.

[7] J. F. Cardoso, "Blind Signal Separation: Statistical Principles," **Proceedings of the IEEE**, vol. 86, no. 10, pp. 2009–2025, 1998.

[8] S. Choi, A. Cichocki and S. Amari, "Flexible Independent Component Analysis," in **Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop**, 1998, pp.

83–92.

[9] J. Eriksson, J. Karvanen and V. Koivunen, “Source Distribution Adaptive Maximum Likelihood Estimation of ICA Model.” in **ICA2000. Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation**, 2000, pp. 227–232.

[10] A. Hyvärinen, “FastICA Matlab code with references available at http://www.cis.hut.fi/projects/ica/fastica/,” .

[11] A. Hyvärinen, “Fast and Robust Fixed-Point Algorithms for Independent Component Analysis,” **IEEE Transactions on Neural Networks**, vol. 10, no. 3, pp. 626–634, 1999.

[12] A. Hyvärinen, “The Fixed-Point Algorithm and Maximum Likelihood Estimation for Independent Component Analysis,” **Neural Processing Letters**, vol. 10, no. 1, pp. 1–5, 1999.

[13] Z. A. Karian and E. J. Dudewicz, “The Extended Generalized Lambda Distribution (EGLD) System for Fitting Distributions to Data with Moments, II: Tables,” **American Journal of Mathematical and Management Sciences**, 1996.

[14] Z. A. Karian, E. J. Dudewicz and P. McDonald, “The Extended Generalized Lambda Distribution System for Fitting Distributions to Data: History, Completion of Theory, Tables, Applications, the "Final Word" on Moment Fits,” **Communications in Statistics: Simulation and Computation**, vol. 25, no. 3, pp. 611–642, 1996.

[15] J. Karvanen, J. Eriksson and V. Koivunen, “Pearson System Based Method for Blind Separation,” in **ICA2000. Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation**, 2000, pp. 585–590.

[16] T.-W. Lee, **Independent Component Analysis: Theory and Applications**, Boston: Kluwer Academic Publishers, 1998.

[17] S. Makeig, T. Bell, T.-W. Lee and et al., “Infomax Matlab code with references available at http://www.cnl.salk.edu/˜scott/ica-download-form.html,” .

[18] G. W. Steward and J.-G. Sun, **Matrix Perturbation Theory**, Boston: Academic Press, 1990.

[19] A. Stuart and J. K. Ord, **Kendall's Advanced Theory of Statistics: Distribution Theory**, vol. 1, Edward Arnold, sixth edn., 1994.