SOURCE DISTRIBUTION ADAPTIVE MAXIMUM LIKELIHOOD ESTIMATION OF ICA MODEL

Jan Eriksson, Juha Karvanen*, and Visa Koivunen*

Signal Processing Laboratory
Helsinki University of Technology
P.O. Box 3000, FIN-02015 HUT, Finland
{jan.eriksson,juha.karvanen,visa.koivunen}@hut.fi

ABSTRACT

In this paper a new approach for performing Independent Component Analysis (ICA) is introduced. The Extended Generalized Lambda Distribution (EGLD) is employed for modeling source distributions. The major benefit of the EGLD is that it also takes into account the skewness of the distributions. We briefly review maximum likelihood approach in ICA and study how the parameters of EGLD may be estimated. The score function of EGLD based ICA is presented and algorithms for its maximization are proposed. The simulation examples illustrate that the proposed method reliably separates the sources in situations where some widely used contrast functions may perform poorly.

1. INTRODUCTION

Independent Component Analysis (ICA) provides a powerful tool for various signal processing, data analysis and communications applications. Theoretical measures of independence used in ICA require information on the probability distributions of the sources. In blind methods, however, no such information is assumed to be available.

Many practical algorithms developed for estimating the ICA model employ only the second and fourth moments. In engineering applications there are many distributions of great importance that are skewed. Furthermore, these non-Gaussian distributions may have Gaussian fourth moment i.e. kurtosis near zero. Consequently, ICA model estimation based on kurtosis only may give misleading results. In this paper we propose a source distribution adaptive approach to maximum likelihood estimation of ICA model. The estimated sources are modeled using the Extended Generalized Lambda Distribution (EGLD) model. The modeling approach provides a useful connection between practical estimator and theoretical measure of independence. The EGLD model covers an extensive range of skewness and kurtosis values that characterize a wide class of distributions of interest in engineering and data analysis. Many widely used skewed distributions are included in this class. The score function of the EGLD is used as an ICA contrast function that we maximize. Natural gradient and fixed-point algorithms employing EGLD model are proposed for maximization. Simulation results are presented illustrating reliable performance of an EGLD based algorithm in separating sources with commonly used distributions in signal processing and communications.

This paper is organized as follows. Section 2 describes briefly the ICA model and maximum likelihood estimation of the model. In section 3, the EGLD model is described and some commonly used distributions covered by the model are pointed out. Fitting the EGLD model is considered as well. A method for estimating the ICA model using EGLD is introduced in Section 4. Examples illustrating the reliable performance of the method are given in Section 5. Finally, section 6 concludes the paper.

2. ICA MODEL

2.1. Data model

We consider the classical 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

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

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 $\mathbf{s}$.

2.2. Maximum Likelihood Approach

Many different proposed approaches to ICA are equivalent [9] in the sense that they lead to the same iterative algorithm. We describe here a variant of the MLE approach, see e.g. [3]. The main idea is to minimize the Kullback-Leibler (KL) divergence

$$D\big(\mathbf{y}\,||\,\mathbf{s}\big) = \int f(\mathbf{y}) \log \frac{f(\mathbf{y})}{g(\mathbf{s})} d\mathbf{x} \tag{3}$$

between the distribution of the output $\mathbf{y}$ and the distribution of the source $\mathbf{s}$ with respect to the matrix $\mathbf{W}$ and with respect to the source distribution. The basic property of KL divergence is that for any vector $\mathbf{s}$ with independent entries,

$$D\big(\mathbf{y}\,||\,\mathbf{s}\big) = D\big(\mathbf{y}\,||\,\hat{\mathbf{y}}\big) + D\big(\hat{\mathbf{y}}\,||\,\mathbf{s}\big), \tag{4}$$

where $\hat{\mathbf{y}}$ denotes the vector with independent entries with each entry distributed as the corresponding marginal of $\mathbf{y}$. Since the first term of (4) does not depend from $\mathbf{s}$, the divergence is minimized in $\mathbf{s}$ by minimizing its second term. This is simply done by taking $\mathbf{s} = \hat{\mathbf{y}}$, i.e. we should model the source distribution as the corresponding marginal distribution. Having done that, it is left to minimize the mutual information

$$D\big(\mathbf{y}\,||\,\hat{\mathbf{y}}\big) \tag{5}$$

with respect to $\mathbf{W}$. This may be done, for example, by using the the gradient based methods [1, 3], or Hyvärinen's fixed-point algorithm [5, 6]. It is shown in [10] that the relative gradient for the likelihood given the source densities is

$$E\{\varphi(\mathbf{y})\mathbf{y}^T - I\}, \tag{6}$$

where $\varphi(\mathbf{y})$ is the vector of the score functions $\varphi_i = -\frac{d}{dy_i}\log g(y_i)$ related to each source. Therefore, if we can reliably estimate the probability densities of the marginals $\hat{\mathbf{y}}$, we have an efficient algorithm for ICA. In this paper, we propose a method that uses the Extended Generalized Lambda Distribution as the model.

3. THE EXTENDED GENERALIZED LAMBDA DISTRIBUTION

The Extended Generalized Lambda Distribution (EGLD) is a large family of distributions covering the whole space of the third, α_3, and the fourth, α_4, moments cited in literature as actually occuring in practice [8]. The lambda distribution was presented by Tukey [14] in 1960. The concept was generalized in 70's [12, 13, 11]. Its main use has been in fitting a distribution to the empirical data, and in the computer generation of different distributions. The latest extension of the family by Karian and Dudewicz in 1996 [8] is a combination of Generalized Lambda Distribution (GLD) and Generalized Beta Distribution (GBD). The space of (α_3, α_4) values, which is covered by the EGLD distribution family, includes the values for all the most important distribution including normal, uniform, gamma and beta distributions. This is illustrated in Figure 1.

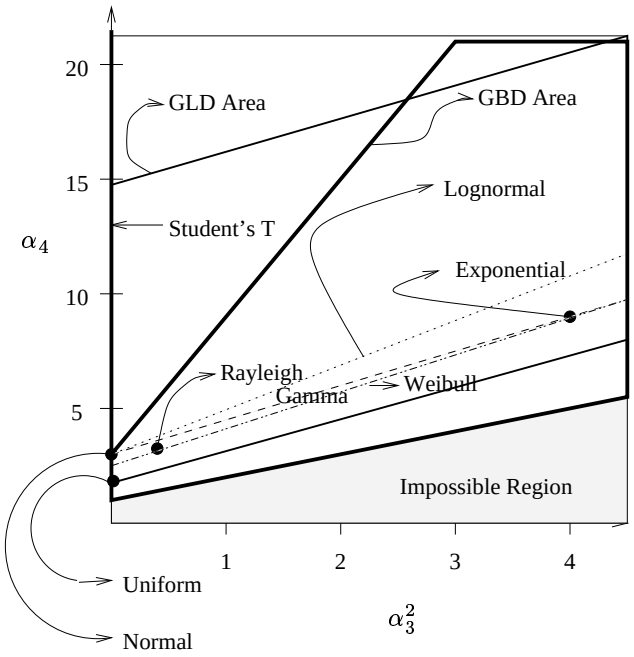


Figure 1: Characterization of some standardized distributions by their third and fourth moments. EGLD covers the area above the shaded region, which is not valid for any distribution. Skewness and kurtosis of many distributions occurring in engineering applications are pointed out

The Generalized Lambda Distribution is defined by the inverse distribution function

$$F^{-1}(p) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, \tag{7}$$

where $0 \le p \le 1$ and λ_1, λ_2, λ_3 and λ_4 are the parameters of the distribution. Karian and Dudewicz [8]

showed that GLD is a valid distribution if and only if

$$\frac{\lambda_2}{\lambda_3 p^{\lambda_3-1} + \lambda_4(1-p)^{\lambda_4-1}} \geq 0. \quad (8)$$

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 GLD employing the inverse distribution function. Estimation of GLD parameters using method of moments is proposed in [8]. The four first sample moments are calculated from data

$$\hat{\alpha}_1 = \bar{x} = \sum_{i=1}^{n} x_i/n \quad (9)$$

$$\hat{\alpha}_2 = \hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n \quad (10)$$

$$\hat{\alpha}_3 = \sum_{i=1}^{n} (x_i - \bar{x})^3/(n\hat{\sigma}^3) \quad (11)$$

$$\hat{\alpha}_4 = \sum_{i=1}^{n} (x_i - \bar{x})^4/(n\hat{\sigma}^4) \quad (12)$$

The relationship between parameters λ_1, λ_2, λ_3 and λ_4 and moments α_1, α_2, α_3 and α_4 is established by four non-linear equations that can be solved numerically. However, due to the intricacy of the computational process, the parameters λ_1, λ_2, λ_3 and λ_4 are tabulated in [7] as functions of α_3 and α_4 for standardized data where $\alpha_1 = 0$ and $\alpha_2 = 1$.

The other part, the Generalized Beta Distribution, is characterized by the density function

$$f(x) = C\beta_2^{-(\beta_3+\beta_4+1)}(x-\beta_1)^{\beta_3}(\beta_1+\beta_2-x)^{\beta_4} \quad (13)$$

on the interval $[\beta_1, \beta_1+\beta_2]$ and zero elsewhere. C is a constant.

The area that the GBD covers is given in terms of moments

$$1 + \alpha_3^2 < \alpha_4 < 3 + 2\alpha_3^2. \quad (14)$$

When the EGLD is fitted to data the choice between GLD and GBD is made based on the values of $\hat{\alpha_3}$ and $\hat{\alpha_4}$. In the area where GLD and GBD overlap either model may be used. Experimentally, it seems that models with parameter values near the theoretical boundaries of GLD or GBD should be avoided.

Although the calculation of estimators for parameters of GBD is not so complicated as in the case of GLD, tables are provided [7] for parameters β_3 and β_4 as functions of sample moments $\hat{\alpha}_3$ and $\hat{\alpha}_4$. However, the parameters can be obtained directly by solving the moment equations

$$\hat{\alpha}_3 = \frac{2(\beta_4-\beta_3)\sqrt{C_3}}{C_4\sqrt{(\beta_3+1)(\beta_4+1)}} \quad (15)$$

and

$$\hat{\alpha}_4 = \frac{3C_3(\beta_3\beta_4C_2 + 3\beta_3^2 + 5\beta_3 + 3\beta_4^2 + 5\beta_4 + 4)}{C_4C_5(\beta_3+1)(\beta_4+1)}, \quad (16)$$

where $C_i = \beta_3 + \beta_4 + i$ for $i = 1, \ldots, 5$. Then the parameters β_1 and β_2 are given by

$$\beta_2 = (\beta_3+\beta_4+2)\sqrt{\frac{(\beta_3+\beta_4+3)\hat{\alpha}_2}{(\beta_3+1)(\beta_4+1)}} \quad (17)$$

and

$$\beta_1 = \hat{\alpha}_1 - \frac{\beta_2(\beta_3+1)}{\beta_3+\beta_4+2}. \quad (18)$$

4. ICA USING EGLD MODEL

The underlying source distributions are estimated through the marginal distributions by fitting them to EGLD family using the method of moments as described in the previous section. 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 (7). If we write

$$p = F(y), \quad (19)$$

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

$$\varphi(p) = -\frac{\lambda_2 p^{\lambda_3-2}(\lambda_3-1)\lambda_3}{(p^{\lambda_3-1}\lambda_3 + (1-p)^{\lambda_4-1}\lambda_4)^2} + \frac{\lambda_2(1-p)^{\lambda_4-2}(\lambda_4-1)\lambda_4}{(p^{\lambda_3-1}\lambda_3 + (1-p)^{\lambda_4-1}\lambda_4)^2} \quad (20)$$

In some algorithms the derivative of the score function

$$\begin{aligned}\varphi'(p) &= \frac{2{\lambda_2}^2\left(p^{\lambda_3-2}(\lambda_3-1)\lambda_3 - (1-p)^{\lambda_4-2}(\lambda_4-1)\lambda_4\right)^2}{\left(p^{\lambda_3-1}\lambda_3 + (1-p)^{\lambda_4-1}\lambda_4\right)^4} \\ &- \frac{{\lambda_2}^2 p^{\lambda_3-3}(\lambda_3-2)(\lambda_3-1)\lambda_3}{\left(p^{\lambda_3-1}\lambda_3 + (1-p)^{\lambda_4-1}\lambda_4\right)^3} \\ &- \frac{{\lambda_2}^2(1-p)^{\lambda_4-3}(\lambda_4-2)(\lambda_4-1)\lambda_4}{\left(p^{\lambda_3-1}\lambda_3 + (1-p)^{\lambda_4-1}\lambda_4\right)^3}\end{aligned} \quad (21)$$

is also needed.

The value of score function for observation y is computed by numerically solving for the value of p from (7) and then applying formula (20). The corresponding formula for the GBD is obtained by the straightforward differentiation directly from the density function (13), and is given by

$$\varphi(y) = -\frac{\beta_3}{y-\beta_1} + \frac{\beta_4}{\beta_1+\beta_2-y}, \tag{22}$$

For the derivative of score function we obtain

$$\varphi'(y) = \frac{\beta_3}{(y-\beta_1)^2} + \frac{\beta_4}{(-y+\beta_1+\beta_2)^2}. \tag{23}$$

The actual algorithm optimizing the derived criterion could be any suitable ICA algorithm where maximum likelihood contrasts are utilized. In our experiments we used natural gradient [2] or relative gradient [3] algorithm

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \eta \left(I - \varphi(\mathbf{y})\mathbf{y}^T\right) \mathbf{W}_k, \tag{24}$$

where η is the learning rate, and fixed point algorithm [5, 6]

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \mathbf{D}\left(E\{\varphi(\mathbf{y})\mathbf{y}^T\} - \mathrm{diag}(E\{\varphi(y_i)y_i\})\right)\mathbf{W}_k, \tag{25}$$

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

A procedure for the EGLD-ICA may be given as follows:
Repeat until convergence [1]

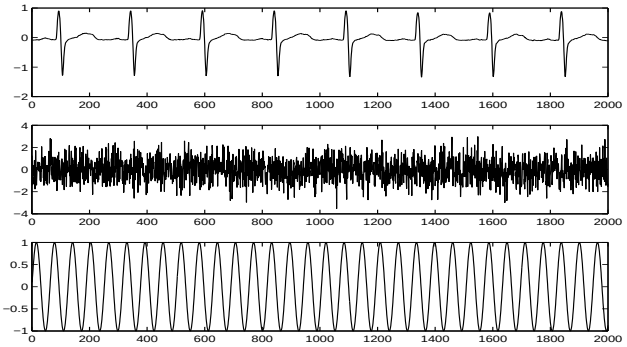
1. Calculate the third and fourth sample moments α_3 and α_4 for current data $\mathbf{y}_k = \mathbf{W}_k\mathbf{x}$ and select the GLD if $\alpha_4 > 2.2 + 2 * \alpha_3^2$ and else the GBD.
2. Estimate parameters for GLD or GBD by method of moments and calculate scores $\varphi(\mathbf{y}_k)$.
3. Calculate the demixing matrix $\mathbf{W}_{k+1}$ using algorithm (24) or (25).

5. SIMULATION EXAMPLES

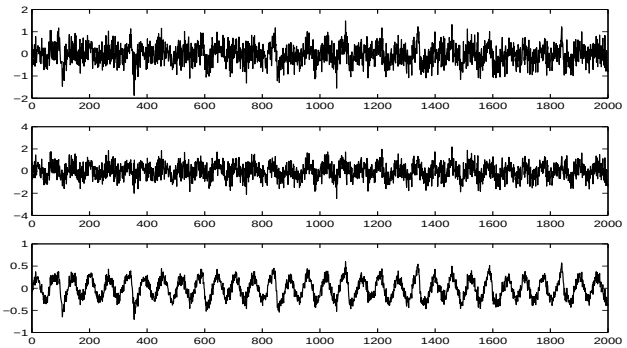
In order to illustrate the performance of the proposed algorithm, we consider first an example with a mixture of three sources: a sine wave (sub-Gaussian), a synthetic ECG signal (super-Gaussian), and a random Gaussian sequence with zero mean and unit variance. The sample size is 10000, and the random mixing matrix

$$\mathbf{A} = \begin{pmatrix} 0.6044 & 0.3829 & 0.1827 \\ 0.0778 & 0.5902 & 0.4741 \\ 0.2332 & 0.0816 & 0.2597 \end{pmatrix}.$$

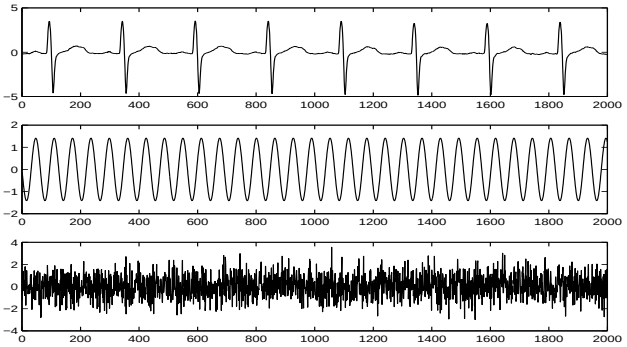
The separation result is shown in Figure 2. It can be seen that the sources are well separated.



(a) Source signals



(b) Mixed signals



(c) Separation by the EGLD-ICA

Figure 2: Separation of a sine wave, an ECG signal, and a Gaussian noise signal with EGLD-ICA.

[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[4] with $\varepsilon = 0.0001$.

Method	Source signals Lognormal	Rayleigh	Normal	GLD
EGLD	48.13	31.96	34.84	26.95
Pow3	25.55	14.39	7.34	1.92
Gauss	11.84	29.25	4.89	4.98
Tanh	13.55	23.42	4.63	4.90

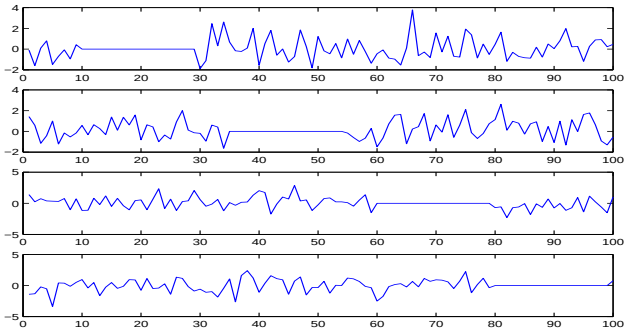
Table 1: The Signal to Interference Ratio (SIR(dB)=$-10\log_{10}$(MSE)) error performance of EGLD-ICA and different contrast functions of FastICA (symmetric approach). The sources and separated signals were normalized to zero mean and unit variance.

The EGLD-ICA algorithm can separate signals with kurtosis equal to that of Gaussian distribution. To illustrate this, four signals of sample length 10000 is generated: Lognormal(0.1,0.15), Rayleigh(1), Normal(0,1), and GLD(0.2370,0.1983,0.1672,0.1065). The GLD distribution has the theoretical moments $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = 0.2$, and $\alpha_4 = 3$. Lognormal and Rayleigh distributions are commonly used for modeling the fading communication channels. The sources are mixed using the (randomly generated) matrix

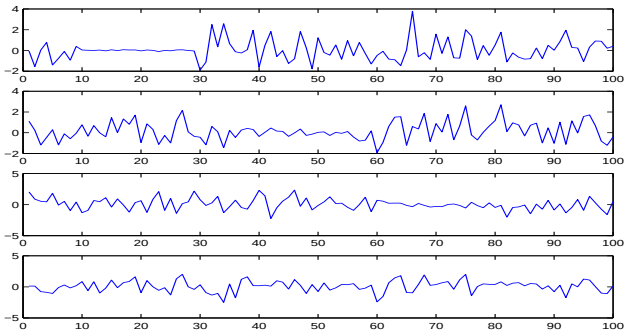
$$\mathbf{A} = \begin{pmatrix} 0.7396 & 0.9084 & 0.2994 & 0.3089 \\ 0.4898 & 0.2980 & 0.5771 & 0.4108 \\ 0.1096 & 0.7808 & 0.8361 & 0.4669 \\ 0.4199 & 0.8799 & 0.2706 & 0.7467 \end{pmatrix}. \quad (26)$$

The EGLD-ICA is compared to FastICA algorithm [4] with different contrast functions with the symmetric approach. The deflation approach of the FastICA seems to give similar results. The comparison is done using the Signal to Interference Ratios (SIR(dB)= $-10\log_{10}$(MSE)) between the zero mean, the unit variance normalized signals. The sources and the sign adjusted signals are matched by taking the signal with the minimum SIR value to be the separation estimate. The minimum SIR values are shown in Table 1. The first hundred observations of the EGLD-ICA and kurtosis separated signals are plotted in Figure 3. As it can be seen the EGLD-ICA performed well for all the sources. The FastICA algorithm with the fixed contrast functions is unable to separate the normal signal from the GLD signal. This also reflects to the SIR values of the other two signals.

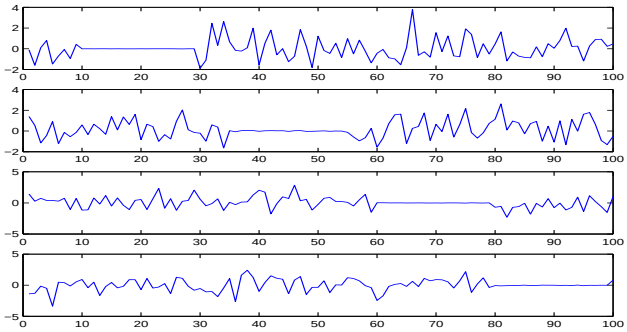
The EGLD-ICA is also tried for the same sources with different sample sizes. The same mixing matrix (26) is used to generate the mixtures. The SIR values averaged over 250 realizations are presented in Figure 4. The results indicate that in this case about 1000 observations are needed in order to achieve good separation. Naturally this result depends on source distri-



(a) Original



(b) Separation by kurtosis (FastICA:Pow3)



(c) Separation by the EGLD-ICA

Figure 3: Separation results of the EGLD-ICA and the kurtosis contrast are compared in the case where the sources are Lognormal(0.1,0.15), Rayleigh(1), Normal(0,1), and GLD(0.2370,0.1983,0.1672,0.1065). The number of samples was 10000 but only the first 100 observations are shown. In order to visualize the quality of separation a short sequence of zeros was added to every source signal. It can be seen that the kurtosis based method fails to separate the GLD and the normal signals, but the EGLD-ICA reliably separates all the sources.

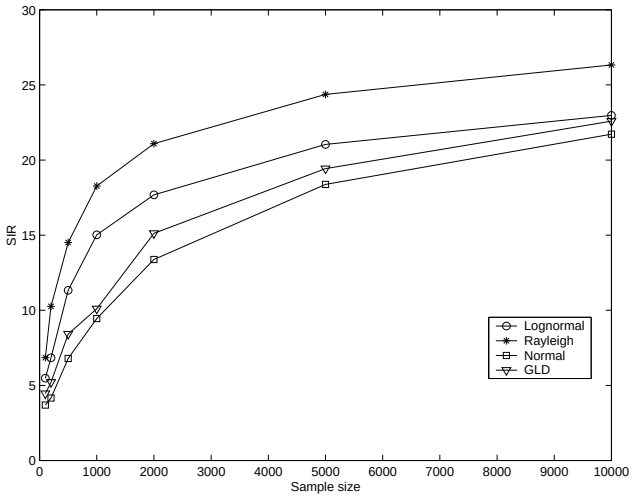


Figure 4: The SIR error performance of the EGLD-ICA as the function of the sample size.

butions, and an estimator used for EGLD parameters.

6. CONCLUSION

We introduced the EGLD-ICA algorithm for ICA. It employs the Maximum Likelihood Principle. Source distribution adaptive contrast function is derived using the Extended Generalized Lambda Distribution as the model. We have presented the theoretical background for the use of EGLD-ICA, and shown that it can separate a wide class of source signals including sub- and super-Gaussian, and skewed distributions with zero kurtosis.

7. REFERENCES

[1] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[2] S.-I. Amari and A. Cichocki. Adaptive blind signal processing – neural network approaches. *Proceedings of the IEEE*, 86(10):2026–2048, 1998.

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

[4] A. Hyvärinen. Code with references available at `http://www.cis.hut.fi/projects/ica/fastica/`.

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

[6] A. Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 10(1):1–5, 1999.

[7] 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.

[8] 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. *Comm. Stat.: Simulation and Computation*, 25(3):611–642, 1996.

[9] T.-W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski. A unifying information theoretic framework for independent component analysis. *International Journal on Mathematical and Computer Modelling*, (in press).

[10] D. T. Pham and P. Garat. Blind signal separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. Signal Processing*, 45(7):1712–1725, 1997.

[11] J. S. Ramberg, E. J. Dudewicz, P. R. Tadikamalla, and E. F. Mykytka. A probability distribution and its uses in fitting data. *Technometrics*, 21:201–204, 1979.

[12] J. S. Ramberg and B. W. Schmeiser. An approximate method for generating asymmetric random variables. *Comm. ACM*, 15:987–990, 1972.

[13] J. S. Ramberg and B. W. Schmeiser. An approximate method for generating asymmetric random variables. *Comm. ACM*, 17:78–82, 1974.

[14] J. W. Tukey. The practical relationship between the common transformations of percentages of counts and of amounts. Technical Report 36, Statistical Techniques Research Group, Princeton University, 1960.