Robust Independent Component Analysis Using Quadratic Negentropy

Jaehyung Lee, Taesu Kim, and Soo-Young Lee

Department of Bio & Brain Engineering, KAIST, Republic of Korea
{jaehyung.lee, taesu.kim, sylee}@kaist.ac.kr

Abstract. We present a robust algorithm for independent component analysis that uses the sum of marginal quadratic negentropies as a dependence measure. It can handle arbitrary source density functions by using kernel density estimation, but is robust for a small number of samples by avoiding empirical expectation and directly calculating the integration of quadratic densities. In addition, our algorithm is scalable because the gradient of our contrast function can be calculated in O(LN) using the fast Gauss transform, where L is the number of sources and N is the number of samples. In our experiments, we evaluated the performance of our algorithm for various source distributions and compared it with other, well-known algorithms. The results show that the proposed algorithm consistently outperforms the others. Moreover, it is extremely robust to outliers and is particularly more effective when the number of observed samples is small and the number of mixed sources is large.

1 Introduction

In the last decade, Independent Component Analysis (ICA) has shown to be a great success in many applications, including sound separation, EEG signal analysis, and feature extraction. ICA shows quite a good performance for simple source distributions, if given assumptions hold well, but its performance is degraded for sources with skewed or complex density functions [1]. Several ICA methods are currently available for arbitrary distributions, but these methods have not yet shown practical performance when the number of sources is large and the number of observed samples is small, thus preventing their application to more challenging real-world applications, such as blind source separation for non-stationary mixing environments and frequency-domain BSS for convolutive mixtures [2].

The problem of ICA for arbitrary distributions mainly arises from the difficulty of estimating marginal entropies that usually appear in the contrast function derived from mutual information. Direct estimation of marginal entripies without parametric assumptions involves excessive computation, including numerical integration, and is sensitive to outliers because of the log terms. Several approximations are available, but these still rely on higher order statistical terms that are also sensitive to outliers. Different estimators of entropy [3] or dependence measure based on canonical correlations [1] have been suggested to

M.E. Davies et al. (Eds.): ICA 2007, LNCS 4666, pp. 227–235, 2007.

overcome this problem and have shown promising performance. In addition, there have been approaches using nonparametric mutual information via Renyi's entropy [4] for ICA [5]. However, this method requires sign correction by kurtosis because Renyi's entropy does not have a maximum at a Gaussian distribution [6].

In this paper, we define the concept of quadratic negentropy, replace the original negentropy with quadratic negentropy in the original definition of mutual information, and obtain a new contrast function for ICA. Using kernel density estimation along with quadratic negentropy can reduce the integration terms into sums of pairwise interactions between samples. The final contrast function can be calculated efficiently using the fast Gauss transform, guaranteeing scalability. The performance of our algorithm consistently outperforms the best existing algorithms for various source distributions and the existence of outliers, especially when the number of observed samples is small and the number of mixed sources is large.

This paper is organized as follows. In Section 2, we review the basic problem of ICA and the contrast function using negentropy. In Section 3, we define a new contrast function for ICA using quadratic negentropy along with kernel density estimation. We also apply the fast Gauss transform to reduce computation. In Section 4, we evaluate the performance of the derived algorithm on various source distributions, varying the number of sources and the number of samples, to compare the proposed algorithm with other, well-known algorithms, such as FastICA and KernelICA.

2 Background on ICA

In this section, we briefly review the basic problem of ICA and the contrast function using original negentropy.

2.1 The Basic Problem of ICA

Let $s_1, s_2, ..., s_L$ be L statistically independent source random variables that are linearly mixed by some unknown but fixed mixing coefficients to form m observed random variables $x_1, x_2, ..., x_L$. For example, source variables can be the voices of different people at a location and observation variables represent the recordings from several microphones at the location. This can be written in matrix form as

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

where $\mathbf{x} = (x_1, x_2, ..., x_L)^T$, $\mathbf{s} = (s_1, s_2, ..., s_L)^T$, and $\mathbf{A}$ is an L × L matrix. The basic problem of ICA is to determine $\mathbf{W}$, the inverse of mixing matrix $\mathbf{A}$, to recover the original sources from observations, by using N samples of observation $\mathbf{x}$ under the assumption that sources are independent of each other.

2.2 Contrast Function Using Negentropy

Mutual information between components of estimated source vectors is known to be a natural contrast function for ICA because it has a zero value when

the components are independent and a positive value otherwise. In addition, it is well known that mutual information can be represented using joint and marginal negentropies [7], as follows:

$$I(\mathbf{x}) = J(\mathbf{x}) - \sum_{i=1}^{N} J(x_i) + \frac{1}{2}\log\frac{\prod V_{ii}}{\det V} \tag{2}$$

where $\mathbf{x}$ is a vector random variable of dimension N, x_i is the i-th component of $\mathbf{x}$, V is the covariance matrix of $\mathbf{x}$, and $J(\mathbf{x})$ is the negentropy of a random variable $\mathbf{x}$, which can be represented using Kullback-Leibler divergence, as shown below. The proof is based on the fact that only the first and second order moment of Gaussian density are nonzero and that $\log p_\phi(\boldsymbol{\xi})$ is a polynomial of degree 2 [7].

$$J(\mathbf{x}) = D_{KL}(p_x||p_\phi) = \int p_x(\boldsymbol{\xi})\log\frac{p_x(\boldsymbol{\xi})}{p_\phi(\boldsymbol{\xi})}d\boldsymbol{\xi} \tag{3}$$

where $\boldsymbol{\phi}$ is a Gaussian random variable that has the same mean and variance with $\mathbf{x}$, and p_ϕ is the pdf of $\boldsymbol{\phi}$. As a result, it is nonnegative, invariant to invertible transforms and zero if $p_x \equiv p_\phi$.

If we assume $\mathbf{x}$ be whitened, then the last term of Eq. (2) becomes zero and only negentropy terms remain. Now, we define the contrast function of ICA using mutual information, as

$$C(\hat{\mathbf{W}}) = -I(\hat{\mathbf{s}}) = \sum_{i=1}^{L} J(\hat{s}_i) - J(\hat{\mathbf{s}}). \tag{4}$$

In Eq. (4), $\hat{\mathbf{s}} = \hat{\mathbf{W}}\mathbf{x}$ is the estimated sources using the current estimate of the unmixing matrix $\hat{\mathbf{W}}$, and $\hat{s}_i$ is the i-th component of $\hat{\mathbf{s}}$. We assume that the observation is whitened and thus can restrict the unmixing matrix to rotations only, thus making the first term constant and the third term zero in Eq. (2). The final contrast function of ICA using negentropy can be interpreted as the total nongaussianity of the estimated source components.

3 ICA Using Quadratic Negentropy

3.1 Contrast Function Using Quadratic Negentropy

We replace the KL divergence with the L_2 distance in Eq. (3) and obtain quadratic negentropy defined as

$$J_q(\mathbf{x}) = \int \left(p_x(\boldsymbol{\xi}) - p_\phi(\boldsymbol{\xi})\right)^2 d\boldsymbol{\xi}. \tag{5}$$

We can easily show that it is nonnegative, invariant under rotational transform, and zero if $p_x \equiv p_\phi$. Assuming $\mathbf{x}$ is whitened and using quadratic negentropy instead of the original negentropy in Eq. (4), we obtain

$$C_q(\hat{\mathbf{W}}) = -I_q(\hat{\mathbf{s}}) = \sum_{i=1}^{L} J_q(\hat{s}_i) - J_q(\hat{\mathbf{s}}). \tag{6}$$

In addition, $J_q(\hat{\mathbf{s}})$ is constant because the quadratic negentropy is invariant under a rotational transform. Ignoring the constant gives us

$$C_q(\hat{\mathbf{W}}) = \sum_{i=1}^{L} J_q(\hat{s}_i) = \sum_{i=1}^{L} \int (\hat{p}_{\hat{s}_i}(\xi) - \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2})^2 d\xi \tag{7}$$

where $\hat{p}_{\hat{s}_i}$ is the estimated marginal pdf of $\hat{s}_i$. Here $\hat{s}_i$ has zero mean and unit variance because $\hat{\mathbf{W}}$ is rotation and $\mathbf{x}$ is whitened. Thus p_ϕ in (5) becomes a standard Gaussian pdf.

To be a contrast function, Eq. (6) and (7) should have a global maximum when components are independent. We hope this can be proved for general source distributions, but currently we have proof only for Laplacian distributions and further work is needed.

3.2 Kernel Density Estimation

Using kernel density estimation, $\hat{p}_{\hat{s}_i}$ can be estimated as

$$\hat{p}_{\hat{s}_i}(y) = \frac{1}{N} \sum_{n=1}^{N} G(y - \hat{s}_i(n), \sigma^2) \tag{8}$$

where N is the number of observed samples, $\hat{s}_i(n)$ is the n-th observed sample of i-th estimated source, and $G(y, \sigma^2)$ is a Gaussian kernel defined as

$$G(y, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2}. \tag{9}$$

Interestingly, the calculation of integration involving quadratic terms of $\hat{p}_{\hat{s}_i}$ estimated as (8) can be simplified as pairwise interactions between samples [8]. Simplifying Eq. (7) using this yields

$$C_q(\hat{\mathbf{W}}) = \sum_{i=1}^{L} \left(\frac{1}{2\sqrt{\pi}} + \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} G(\hat{s}_i(n) - \hat{s}_i(m), 2\sigma^2) - \frac{2}{N} \sum_{n=1}^{N} G(\hat{s}_i(n), 1+\sigma^2) \right), \tag{10}$$

which is our final contrast function to maximize. Obtaining the partial derivative of $C_q(\hat{\mathbf{W}})$ with respect to w_{ij} yields

$$\begin{aligned} \frac{\partial C_q}{\partial w_{ij}} = \sum_{n=1}^{N} \Bigg(& \frac{2 \cdot G(\hat{s}_i(n), 1+\sigma^2) \hat{s}_i(n)}{N \cdot (1+\sigma^2)} \\ & - \sum_{m=1}^{N} \frac{G(\hat{s}_i(n) - \hat{s}_i(m), 2\sigma^2)(\hat{s}_i(n) - \hat{s}_i(m))}{N^2 \cdot \sigma^2} \Bigg) x_j(n) \end{aligned} \tag{11}$$

where symmetry with respect to m and n is utilized to simplify equation. Also note that $\hat{s}_i(n) = \sum_{j=1}^{L} w_{ij} x_j(n)$.

3.3 Efficient Computation Using Fast Gauss Transform

It takes $O(LN^2)$ time to directly compute the gradient given in Eq. (11). To reduce the computation we use the fast Gauss transform [9] that evaluates the following in $O(N+N')$ time, given 'source' points $\mathbf{x} = \{x_1, ..., x_N\}$ and 'target' points $\mathbf{y} = \{y_1, ..., y_{N'}\}$.

$$FGT(y_j, \mathbf{x}, \mathbf{q}, h) = \sum_{i=1}^{N} q_i e^{-(y_j - x_i)^2/h^2}, j = 1, ..., N' \tag{12}$$

where $\mathbf{q} = \{q_1, ..., q_N\}$ are weight coefficients and h is the bandwidth parameter.

Using Eq. (12), Eq. (10) can be rewritten as

$$C_q(\hat{\mathbf{W}}) = \sum_{i=1}^{L} \left(\frac{1}{2\sqrt{\pi}} + \frac{1}{N^2} \sum_{n=1}^{N} \frac{FGT(\hat{s}_i(n), \hat{\mathbf{s}}_\mathbf{i}, \mathbf{1}, \sqrt{2}\sigma)}{2\sqrt{\pi}\sigma} - \frac{2}{N} \sum_{n=1}^{N} G(\hat{s}_i(n), 1+\sigma^2) \right), \tag{13}$$

and the partial derivative in Eq. (10) can be rewritten as

$$\begin{aligned} \frac{\partial C_q}{\partial w_{ij}} = \sum_{n=1}^{N} \Bigg(& \frac{2 \cdot G(\hat{s}_i(n), 1+\sigma^2)\hat{s}_i(n)}{N \cdot (1+\sigma^2)} \\ & - \frac{FGT(\hat{s}_i(n), \hat{\mathbf{s}}_\mathbf{i}, \hat{\mathbf{s}}_\mathbf{i}, \sqrt{2}\sigma) - FGT(\hat{s}_i(n), \hat{\mathbf{s}}_\mathbf{i}, \mathbf{1}, \sqrt{2}\sigma)}{2\sqrt{\pi} \cdot N^2 \cdot \sigma^3} \Bigg) x_j(n) \end{aligned} \tag{14}$$

where $\hat{\mathbf{s}}_\mathbf{i} = \{\hat{s}_i(1), ..., \hat{s}_i(N)\}$ and $\mathbf{1}$ is an N-dimensional one vector.

Now, Eq. (13) and Eq. (14) can be computed in $O(LN)$ by performing the fast Gauss transform $2L$ times.

3.4 Steepest Descent on Stiefel Manifold

The set of orthogonal matrices is a special case of the Stiefel manifold and a gradient of a function can be computed based on the canonical metric of the Stiefel manifold [10]. Unconstrained optimization on the Stiefel manifold is more efficient than orthogonalizing the weight matrix per each iteration. In this paper, we used the steepest descent with a bracketed backtracking line search along geodesics.

3.5 Parameter Selection and Convergence Criterion

Our learning rule has one parameter: the bandwidth parameter σ of the kernel density estimation. We used $\sigma = 1.06 \times N^{-1/5}$ [11].

We calculated the value of the contrast function per each iteration to check convergence. If the difference between iterations becomes less than a given ratio $\tau = 10^{-8}$ of the contrast function, then it is regarded as convergence.

In general, ICA contrast functions have multiple local maxima. This is also true for our contrast function, and we needed a fixed number of restarts to find

a good local optimum. We restarted our algorithm four times with a convergence criterion $\tau = 10^{-6}$ and picked the best one as an initial estimate for final optimization.

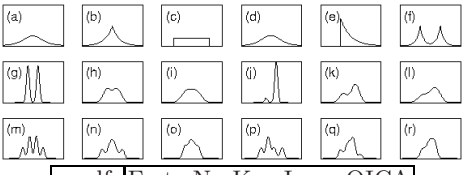
4 Experimental Results

We conducted an extensive set of simulation experiments using a variety of source distributions, sample numbers, and components. The 18 source distributions used in our experiment were adopted from the KernelICA paper [1]. They include subgaussian, supergaussian and nearly Gaussian source distributions and

Table 1. LEFT: The normalized Amari errors (×100) for mixtures of identical source distributions (top left) and random source distributions (bottom left). L: number of mixed components, N: number of samples, Fast: FastICA, Np: NpICA, Kgv: KernelICA-KGV, Imax: extended infomax ICA, QICA: our method. For identical sources, simulation is repeated 100 times for each of the 18 source distributions for $L = \{2, 4\}$, 50 times for $L = 8$, and 20 times for $L = 16$. For random sources, simulation is repeated 2000 times for $L = \{2, 4\}$, 1000 times for $L = 8$, and 400 times for $L = 16$. **RIGHT:** Amari errors for each source distributions for $L = 2$ and $N = 1000$.

L	N	Fast	Np	Kgv	Imax	QICA
	100	20.6	20.3	16.3	21.3	**15.7**
2	250	13.0	12.9	8.6	14.4	**7.7**
	1000	6.5	9.8	3.0	8.5	**2.9**
	100	28.6	23.0	28.4	23.1	**18.9**
4	250	16.8	13.9	19.2	14.5	**9.8**
	1000	6.9	6.5	7.2	8.7	**3.6**
	250	30.2	20.9	31.3	18.1	**15.9**
8	1000	10.6	7.8	20.6	8.2	**4.7**
	2000	6.4	4.7	14.4	6.2	**2.8**
	1000	26.2	17.3	30.4	**11.1**	12.4
16	2000	11.8	12.5	26.1	**6.6**	6.9
	4000	7.1	6.9	21.3	4.8	**4.3**

L	N	Fast	Np	Kgv	Imax	QICA
	100	18.0	13.6	13.4	19.0	**12.0**
2	250	11.3	7.2	6.3	13.1	**6.1**
	1000	5.6	2.8	**2.4**	6.7	2.5
	100	24.5	18.1	26.3	21.2	**14.9**
4	250	13.7	8.5	14.1	13.1	**6.9**
	1000	5.7	2.6	3.4	5.9	**2.5**
	250	25.4	14.8	30.0	16.0	**10.1**
8	1000	6.3	2.9	13.4	6.0	**2.7**
	2000	4.0	**1.7**	5.6	4.1	1.8
	1000	12.5	8.5	27.9	7.9	**4.1**
16	2000	4.3	2.6	27.0	4.3	**2.3**
	4000	2.9	**1.2**	20.3	2.9	2.0



pdfs	Fast	Np	Kgv	Imax	QICA
a	4.7	5.6	3.0	**2.1**	2.7
b	5.5	4.1	3.0	2.7	**2.4**
c	2.3	3.1	**1.6**	3.0	2.1
d	7.2	8.8	**5.7**	6.4	6.4
e	5.7	**0.9**	1.3	3.3	1.6
f	4.7	26.9	**1.5**	1.6	**1.5**
g	1.7	30.0	1.3	**1.1**	1.3
h	5.8	5.7	4.5	**3.4**	3.6
i	9.4	14.9	9.5	**6.9**	7.3
j	7.0	29.7	**1.4**	11.4	**1.4**
k	5.8	3.3	2.8	4.9	**2.7**
l	12.1	**4.8**	5.5	8.2	**4.8**
m	3.5	14.9	**1.4**	4.3	**1.4**
n	5.7	10.7	**1.8**	22.3	1.9
o	4.4	**3.1**	3.6	4.2	3.9
p	3.8	**1.1**	1.5	8.0	1.6
q	21.8	4.3	**2.1**	53.2	2.5
r	6.0	3.5	**2.9**	5.1	3.5
mean	6.5	9.8	3.0	8.5	**2.9**
std	4.5	9.7	2.1	12.2	**1.7**

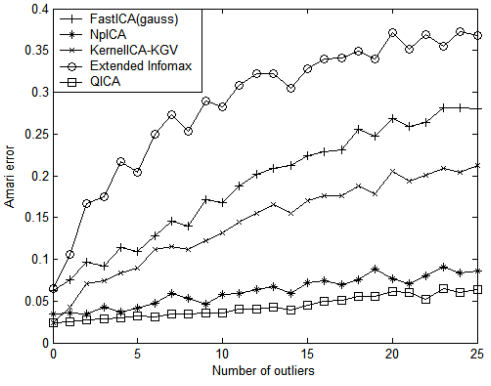


Fig. 1. Robustness to outliers for $L = 2$, $N = 1000$. Up to 25 observations are corrupted by adding +5 or -5. The experiment is repeated 1000 times with random source distributions.

unimodal, multimodal, symmetric, and skewed sources. We varied the number of samples from 100 to 4000 and the number of components from 2 to 16.

Comparisons were made with four existing ICA algorithms: the FastICA algorithm [12], the KernelICA-KGV algorithm [1], the extended infomax algorithm [13] using tanh nonlinearity, and the NpICA algorithm [14]. Software programs were downloaded from corresponding authors' websites and were used with default parameters, except for the extended infomax algorithm, which is our own implementation. Note that KernelICA-KGV also has four restarts as a default to obtain initial estimates. The performance was measured using the Amari error [15], which is invariant to permutation and scaling, lies between 0 and L−1 and is zero for perfect demixing. We normalized the Amari error by dividing it by L−1, where L is the number of independent components.

We summarized our results in Table 1. Consistent performance improvement over existing algorithms was observed. The improvement was significant if the number of components was large and the number of observations was small. However, the performance gain became smaller as the number of observations increased. Amari errors for each source pdf are also shown separately for two-components and 1000 observations. The proposed method showed the smallest standard deviation among the five methods. All of the methods, except for KernelICA-KGV and the proposed method had problems with specific pdfs.

Another interesting result was the high performance of the extended infomax algorithm for a large number of components. For $L = 16$, it showed the best performance among the five methods. But further experiments with outliers discouraged its practical use.

Fig. 1 shows the result of the outlier experiment. We randomly chose up to 25 observations and added the value +5 or -5 to a single component in the observation, which was the same as the one in the KernelICA paper. The results show that our method is extremely robust to outliers.

5 Conclusions

We have proposed a robust algorithm for independent component analysis that uses the sum of marginal quadratic negentropies as a dependence measure. The proposed algorithm can handle arbitrary source distributions and is scalable with respect to the number of components and observations. Experimental results have shown that the proposed algorithm consistently outperforms others. In addition, it is extremely robust to outliers and more effective when the number of observed samples is small and the number of mixed sources is large.

The proposed contrast function is not guaranteed to have the same maximum with the original one. Empirically, however, our method shows good performance and can be applied to cases where a limited number of observations is available.

Acknowledgment

This research was supported as a Brain Neuroinformatics Research Program by Korean Ministry of Commerce, Industries, and Energy.

References

1. Bach, F.R., Jordan, M.I: Kernel independent component analysis. Journal of Machine Learning Research 3, 1–48 (2002)
2. Araki, S., Makino, S., Nishikawa, T., Saruwatari, H.: Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech. In: Proc. ICASSP. vol. 5, pp. 2737–2740 (2001)
3. Learned-Miller, E.G.: Ica using spacings estimates of entropy. Journal of Machine Learning Research 4, 1271–1295 (2003)
4. Torkkola, K.: Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research 3, 1415–1438 (2003)
5. Hild II, K.E., Erdogmus, D., Principe, J.C.: Blind source separation using renyi's mutual information. IEEE Signal Processing Letters 8(6), 174–176 (2001)
6. Hild II, K.E., Erdogmus, D., Principe, J.C.: An analysis of entropy estimators for blind source separation. Signal Processing 86, 182–194 (2006)
7. Comon, P.: Independent component analysis, a new concept? Signal Processing 36, 287–314 (1994)
8. Principe, J.C., Fisher III, J.W., Xu, D.: Information theoretic learning. In: Haykin, S. (ed.) Unsupervised Adaptive Filtering, Wiley, New York (2000)
9. Greengard, L., Strain, J.: The fast gauss transform. SIAM Journal on Scientific and Statistical Computing 12(1), 79–94 (1991)
10. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications 20(2), 303–353 (1998)
11. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, Sydney (1986)
12. Hyvarinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483–1492 (1997)

13. Lee, T.-W., Girolami, M., Sejnowski, T.J.: Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. Neural Computation 11, 417–441 (1999)
14. Boscolo, R., Pan, H., Roychowdhury, V.P.: Independent component analysis based on nonparametric density estimation. IEEE Transactions on Neural Networks 15(1), 55–65 (2004)
15. Amari, S., Cichocki, A., Yang, H.: A new learning algorithm for blind source separation. In: Advances in Neural Information Processing 8 (Proc. NIPS'95), pp. 757–763 (1996)