# Minimax Mutual Information Approach for Independent Component Analysis

Deniz Erdogmus, Kenneth E. Hild II, Yadunandana N. Rao, Jose C. Principe

Computational NeuroEngineering Laboratory,

Electrical & Computer Engineering Department

University of Florida, Gainesville, FL 32611, USA

## Abstract

Minimum output mutual information is regarded as a natural criterion for independent component analysis (ICA) and is used as the performance measure in many ICA algorithms. Two common approaches in information theoretic ICA algorithms are *minimum mutual information* and *maximum output entropy* approaches. In the former approach, we substitute some form of probability density function (pdf) estimate into the mutual information expression, and in the latter we incorporate the source pdf assumption in the algorithm through the use of nonlinearities matched to the corresponding cumulative density functions (cdf). Alternative solutions to ICA utilize higher order cumulant-based optimization criteria, which are related to either one of these approaches through truncated series approximations for densities. In this paper, we propose a new ICA algorithm motivated by the Maximum Entropy Principle (for estimating signal distributions). The optimality criterion is the minimum output mutual information, where the estimated pdfs are from the exponential family, and are approximate solutions to a constrained entropy maximization problem. This approach yields an upper bound for the actual mutual information of the output signals, hence the name Minimax Mutual Information ICA algorithm. In addition, we demonstrate that for a specific selection of the constraint functions in the maximum entropy density estimation procedure, the algorithm relates strongly to ICA methods using higher order cumulants.

**Keywords:** Independent components analysis, maximum entropy principle, minimum mutual information.

## 1. Introduction

Independent component analysis (ICA) deals with the problem of finding a set of directions such that when the input random vector **x** is projected on these directions, the projected random variables are as independent as possible. As a natural measure of independence between random variables, mutual information is commonly used in this framework. Shannon's definition of mutual information between $n$ random variables $Y_1,\ldots,Y_n$, whose joint pdf is $f_{\mathbf{Y}}(\mathbf{y})$ and marginal pdfs are $f_1(y^1),\ldots,f_n(y^n)$, respectively, is given by (Cover and Thomas, 1991)

$$I(\mathbf{Y}) = \int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}) \log \frac{f_{\mathbf{Y}}(\mathbf{y})}{\prod_{o=1}^{n} f_o(y^o)} d\mathbf{y} \qquad (1)$$

where the vector $\mathbf{y}$ is composed of the entries $y^o$, $o = 1,\ldots,n$. The mutual information is related to the marginal entropies and the joint entropy of these random variables through (Cover and Thomas, 1991)

$$I(\mathbf{Y}) = \sum_{o=1}^{n} H(Y_o) - H(\mathbf{Y}) \tag{2}$$

where the marginal and joint entropies are defined as (Cover and Thomas, 1991)

$$H(Y_o) = -\int_{-\infty}^{\infty} f_o(y^o) \log f_o(y^o) dy^o \tag{3}$$

$$H(\mathbf{Y}) = -\int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}) \log f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \tag{4}$$

Minimization of output mutual information is “the canonical contrast for source separation” as Cardoso states (Cardoso and Souloumiac, 1993). Many researchers agree with this comment (Yang and Amari, 1997; Hyvarinen, 1999a; Almeida, 2000). However, three of the most well known methods for ICA, namely JADE (Cardoso and Souloumniac, 1993), Infomax (Bell and Sejnowski, 1995), and FastICA (Hyvarinen, 1999b), use the diagonalization of cumulant matrices, maximization of output entropy, and fourth order-cumulants, respectively. The difficulty encountered in information theoretic measures is the estimation of the underlying density of the output signals (or, in the case of Infomax, an accurate guess of the independent source densities). Algorithms that do not utilize robust estimations for the probability density functions (pdf) suffer from sensitivity to samples.

One commonly taken approach in designing information theoretic ICA algorithms is the use of some form of polynomial expansion to approximate the pdf of the signals. Some of the commonly used polynomial expansions include Gram-Charlier, Edgeworth, and Legendre, where the pdf estimation is performed by taking a truncated polynomial estimate of the signal pdfs evaluated in the vicinity of a maximum entropy density (Comon, 1994; Amari *et al.*, 1996; Hyvarinen, 1998; Hyvarinen, 1999a; Erdogmus *et al.*, 2001). Even the higher-order cumulant-based contrasts can be understood in this framework (Cardoso, 1999). Since the truncation of these infinite polynomials is necessary to keep computational requirements at a minimum, errors are naturally generated in these approaches. Another density estimation method used in the Minimum Mutual Information (MMI) context is Parzen windowing. Hild *et al.* combine the Parzen window pdf estimation method (Parzen, 1962) with Renyi's mutual information (Renyi, 1970) to derive the Mermaid algorithm, which uses the same topology proposed by Comon (Comon, 1994; Hild *et al.*, 2001). Alternative algorithms using Parzen windowing include the quadratic information divergence approach (Xu *et al.*, 1998) and Pham's sum-of-marginal-Shannon's-entropy approach (Pham, 1996). In addition, the use of orthonormal basis functions in pdf estimation could be pursued for ICA (Girolami, 2002). However, such pdf estimates might become invalid when truncated (i.e., have negative values and do not integrate to one).

Alternative techniques that do not use minimization of mutual information include second-order methods that achieve source separation through decorrelation of the outputs (Weinstein *et al.*, 1993; Wu and Principe, 1997; Parra and Spence, 2000; Pham, 2001), nonlinear principal component analysis (NPCA) approaches (Oja, 1999), maximization of higher auto-statistics (Simon *et al.*, 1998), cancellation of higher-order cross statistics (Comon, 1996; Cardoso, 1998; Hyvarinen, 1999a; Sun and Douglas, 2001), non-Gaussianity measures like the negentropy (Girolami, and Fyfe, 1997; Torkkola, 1999; Wu and Principe, 1999a), maximum likelihood techniques, which are parametric by definition, (Girolami, 1997; Wu and Principe, 1999b; Karvanen *et al.*, 2000), and finally maximum entropy methods (Amari, 1997; Torkkola, 1996; Principe and Xu, 1999).

In this paper, we will take the minimum output mutual information approach in order to come up with an efficient and robust ICA algorithm that is based on a density estimate stemming from Jaynes' maximum entropy principle. The commonly used whitening-rotation scheme, which is also described by Comon (Comon, 1994), will be assumed, where the orthonormal portion of the separation matrix (the rotation stage) will be parameterized using Givens angles (Golub and van Loan, 1993). In this framework, approximations to the maximum entropy pdf estimates that are "consistent to the largest extent with the available data and least committed with respect to unseen data" (Jaynes, 1957) will be used. Upon investigation, under the specific choice of polynomial moments as the constraint functions in the maximum entropy principle, the resulting criterion and the associated algorithm are found to be related to the kurtosis and other higher-order cumulant methods.

## 2. The Topology and the Cost Function

Suppose that the random vector **z** is generated by a linear mixture of the form **z=Hs**, where the components of **s** are independent. Assume that the mixture is square with size *n* and that the source vector **s** is zero-mean and has a covariance matrix of identity. In that case, it is well known that the original independent sources can be obtained from **z** through a two-stage process: spatial whitening to generate uncorrelated but not necessarily independent mixture **x** =**Wz**, and a coordinate system rotation in the *n* dimensional mixture space to determine the independent components, **y=Rx** (Comon, 1994; Cardoso, 1999; Hild *et al.*, 2001). The whitening matrix **W** is determined solely by the second order statistics of the mixture. Specifically, $\mathbf{W}=\Lambda^{-1/2}\,\mathbf{\Phi}^T$, where Λ is the diagonal eigenvalue matrix and **Φ** is the corresponding orthonormal eigenvector matrix of the mixture covariance matrix $\Sigma=E[\mathbf{z}\,\mathbf{z}^T]$, assuming that the observations are zero mean. The rotation matrix **R**, which is restricted to be orthonormal by definition, is optimized through the minimization of the mutual information between the output signals (Comon, 1994; Hild *et al.*, 2001). Considering (2), and the fact that the joint entropy is invariant under rotations; mutual information simplifies to the sum of marginal output entropies for this topology.

$$J(\mathbf{\Theta}) = \sum_{o=1}^{n} H(Y_o) \tag{5}$$

In (5), the vector **Θ** contains the Givens angles, which are used to parameterize the rotation matrix (Golub and van Loan, 1993). According to this, an *n*-dimensional rotation matrix is parameterized by *n*(*n*-1)/2 parameters, $\theta_{ij}$, where *i*=1,…,*n*-1 and *j*=*i*+1,…,*n*. The rotation matrix $\mathbf{R}^{ij}(\theta_{ij})$ is constructed by starting with an *n*x*n* identity matrix and replacing the entries (*i*,*i*)th, (*i*,*j*)th, (*j*,*i*)th, and (*j*,*j*)th with $\cos\theta_{ij}$, -$\sin\theta_{ij}$, $\sin\theta_{ij}$, and $\cos\theta_{ij}$, respectively. The total rotation matrix is then found as the product of these 2-dimensional rotations:

$$\mathbf{R}(\boldsymbol{\Theta}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \mathbf{R}^{ij}(\theta_{ij}) \tag{6}$$

The described whitening-rotation scheme using the Givens angle parameterization of the rotation matrix has been utilized in ICA algorithms by many researchers and efficient ways of handling the optimization of these parameters have been studied before. Especially, the Jacobi iteration approach for sweeping the Givens angles, thus splitting the high-dimensional optimization problem into a sequence of 1-dimensional problems, has found great interest (Comon, 1994).

## 3. The Maximum Entropy Principe

Jaynes' maximum entropy principle states that in order to determine the pdf estimate that best fits the known data without committing extensively to unseen data, one must maximize the entropy of the estimated distribution under some constraints. The reason for this is that the entropy of a pdf is related with the uncertainty of the associated random variable. In addition, the optimality properties of density estimates obtained using generalized maximum entropy principles has been discussed previously (Shore and Johnson, 1980; Kapur and Kesavan, 1992). The constrained entropy maximization problem is defined as follows:

$$\max_{p_{\overline{X}}(x)} H = -\int_{-\infty}^{\infty} p_{\overline{X}}(x) \log p_{\overline{X}}(x) dx$$

$$\text{subject to } E_{\overline{X}}[f_k(\overline{X})] = \alpha_k = E_X[f_k(X)] \quad k = 1, \ldots, m \tag{7}$$

It can be shown using calculus of variations that the solution to this problem is given by (Cover and Thomas, 1991)

$$p_{\overline{X}}(x) = C(\boldsymbol{\lambda}) \exp\left( \sum_{k=1}^{m} \lambda_k f_k(x) \right) \tag{8}$$

where $\boldsymbol{\lambda} = [\lambda_1 \quad \ldots \quad \lambda_m]^T$ is the Lagrange multiplier vector and $C(\boldsymbol{\lambda})$ is the normalization constant. The constants $\alpha_k$ are pre-specified or in the case of adaptive ICA, determined from the data. The Lagrange multipliers need to be solved simultaneously from the constraints. This, however, is not an easy task in the case of continuous random variables. Analytical results do not exist and numerical techniques are not practical for arbitrary constraint functions due to the infinite range of the definite integrals involved. In order to get around these problems, we will take a different approach. Consider now, for example, the $i^{th}$ constraint equation.

$$\alpha_i = E_{\overline{X}}[f_i(\overline{X})] = \int_{-\infty}^{\infty} f_i(x) p_{\overline{X}}(x) dx \tag{9}$$

Applying the integrating by parts method with the following definitions

$$\begin{aligned} u &= p_{\overline{X}}(x) & v &= \int f_i(x)dx \overset{\Delta}{=} F_i(x) \\ du &= \left( \sum_{k=1}^{m} \lambda_k f_k'(x) \right) p_{\overline{X}}(x) & dv &= f_i(x)dx \end{aligned} \tag{10}$$

where $f_k'(x)$ is the derivative of the constraint function, we obtain

$$\alpha_i = F_i(x) p_{\overline{X}}(x) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} F_i(x) \left( \sum_{k=1}^{m} \lambda_k f_k'(x) \right) p_{\overline{X}}(x) dx \tag{11}$$

If the functions $f_i(x)$ are selected such that their integrals $F_i(x)$ do not diverge faster than the decay rate of the exponential pdf $p_{\overline{X}}(x)$, then the first term on the right hand side of (11) goes to zero. For example, if the constraint functions were selected as the moments of the random variable, this condition would be satisfied, since $F_i(x)$ will become a polynomial and $p_{\overline{X}}(x)$ decays exponentially. This yields

$$\begin{aligned} \alpha_i &= -\sum_{k=1}^{m} \lambda_k \int_{-\infty}^{\infty} F_i(x) f_k'(x) p_{\overline{X}}(x) dx \\ &= -\sum_{k=1}^{m} \lambda_k E_{\overline{X}}\left[F_i(\overline{X}) f_k'(\overline{X})\right] \overset{\Delta}{=} -\sum_{k=1}^{m} \lambda_k \beta_{ik} \end{aligned} \tag{12}$$

Note that the coefficients $\beta_{ik}$ can be estimated using the samples of $X$ by approximating the expectation operators by sample means.[1] Finally, introducing the vector $\boldsymbol{\alpha} = [\alpha_1 \quad \ldots \quad \alpha_m]^T$ and the matrix $\boldsymbol{\beta} = [\beta_{ik}]$, the Lagrange multipliers are given by the following solution of the linear system of equations shown in (12).

$$\boldsymbol{\lambda} = -\boldsymbol{\beta}^{-1}\boldsymbol{\alpha} \tag{13}$$

[1] The expectation is over the maximum entropy distribution, but using the sample mean will approximate these values by equivalently taking the expectation over the actual data distribution. This estimation will become more accurate as the actual density of the samples approaches the corresponding maximum entropy distribution. However, the irrelevance of the accuracy of this approximation for the operation of the algorithm will become clear with the following discussion.

The approach presented above provides us a computationally simple way of finding the coefficients of the estimated pdf of the data directly from the samples, once $\alpha$ and $\beta$ are estimated using sample means. Besides being an elegant approach to find the Lagrange multipliers of the constrained entropy maximization problem using only a linear system of equations, the proposed approach has an additional advantage. Since in the evaluation of $\beta$ the sample mean estimates are utilized, the pdf obtained with the corresponding Lagrange multiplier values will satisfy additional consistency conditions with the samples besides the normally imposed constraints in the problem definition. These extra conditions satisfied by the determined pdf will be of the form $E\left[F_i(X)f_k'(X)\right] = (1/N)\sum_{j=1}^{N} F_i(x_j)f_k'(x_j)$ for $i$=1,…,$m$ and $k$=1,…,$m$. In order to understand this effect better, consider the choice $f_k(x) = x^k$ for the constraint functions. In that case, $\beta_{ik} = k\alpha_{i+k}/(i+1)$, since $F_i(x) = x^{i+1}/(i+1)$ and $f_k'(x) = kx^{k-1}$. Consequently, the first 2$m$ moments of the pdf estimate given by (8) are consistent with those of the samples. However, it should be noted that the pdf possesses the maximum entropy among all pdfs that have the same first $m$ moments.

## 4. Gradient Update for the Givens Angles

In order to find the optimal ICA solution according to criterion (5), using the entropy estimate described in the previous section, a gradient descent update rule for the Givens angles can be employed. The derivative of $H(Y_o)$ with respect to $\theta_{pq}$ is given in (14) (derivation in Appendix A).

$$\frac{\partial H(Y_o)}{\partial \theta_{pq}} = -\sum_{k=1}^{m} \lambda_k^o \frac{\partial \alpha_k^o}{\partial \theta_{pq}} \tag{14}$$

Here, $\boldsymbol{\lambda}^o$ is the parameter vector for the pdf of the $o$th output signal and $\alpha_k^o$ is the value of the $k$th constraint for the pdf of the $o$th output. The former is found by solving the linear equations in (13), and the latter is easily determined from the corresponding output samples using

$$\alpha_k^o = \frac{1}{N}\sum_{j=1}^{N} f_k(y_{o,j}) \tag{15}$$

where $y_{o,j}$ is the $j$th sample at the $o$th output for the current angles. Finally, the derivative of (15) with respect to the angle $\theta_{pq}$ is given by

$$\begin{aligned}\frac{\partial \alpha_k^o}{\partial \theta_{pq}} &= \frac{1}{N}\sum_{j=1}^{N} f_k'(y_{o,j}) \frac{\partial y_{o,j}}{\partial \theta_{pq}} = \frac{1}{N}\sum_{j=1}^{N} f_k'(y_{o,j}) \left(\frac{\partial y_{o,j}}{\partial \mathbf{R}_{o:}}\right)^T \left(\frac{\partial \mathbf{R}_{o:}}{\partial \theta_{pq}}\right)^T \\ &= \frac{1}{N}\sum_{j=1}^{N} f_k'(y_{o,j}) \mathbf{x}_j^T \left(\frac{\partial \mathbf{R}}{\partial \theta_{pq}}\right)_{o:}^T\end{aligned} \tag{16}$$

where the subscript in $\mathbf{R}_{o:}$ and $\left(\partial \mathbf{R}/\partial \theta_{pq}\right)_{o:}$ mean the $o$th row of the corresponding matrix. The derivative of the rotation matrix with respect to $\theta_{pq}$ is also calculated from

$$\begin{aligned}\frac{\partial \mathbf{R}}{\partial \theta_{pq}} = &\left(\prod_{i=1}^{p-1}\prod_{j=i+1}^{n} \mathbf{R}^{ij}(\theta_{ij})\right)\left(\prod_{j=p+1}^{q-1} \mathbf{R}^{pj}(\theta_{pj})\right)\frac{\partial \mathbf{R}^{pq}(\theta_{pq})}{\partial \theta_{pq}} \cdot \\ &\left(\prod_{j=q+1}^{n} \mathbf{R}^{pj}(\theta_{pj})\right)\left(\prod_{i=p+1}^{n-1}\prod_{j=i+1}^{n} \mathbf{R}^{ij}(\theta_{ij})\right)\end{aligned} \tag{17}$$

The over all update rule for the Givens angles is the sum of contributions from each output.

$$\mathbf{\Theta}_{t+1} = \mathbf{\Theta}_t - \eta \sum_{o=1}^{n} \frac{\partial H(Y_o)}{\partial \mathbf{\Theta}} \tag{18}$$

where $\eta$ is a small step size. The computational complexity could be reduced by alternatively following the Jacobi iteration approach (Golub and van Loan, 1993; Comon, 1994; Cardoso, 1994).

## 5. Discussion on the Algorithm

In the previous sections, we have described an ICA algorithm where the selected exponential density estimate for the outputs is motivated by Jaynes' maximum entropy principle. Due to existing difficulties in solving for the Lagrange multipliers analytically, we proposed an approximate numerical solution, which replaces the expectation operator over the maximum entropy distribution by a sample mean over the data distribution. This approximation causes the estimated cost function and the gradient update to deviate from theory. In this section, we will show that this deviation is not critical to the operation of the proposed algorithm. In addition, we will provide an argument on how to choose the constraint functions $f_k(.)$ in the formulation, as well as demonstrating how the criterion reduces to one based simply on higher order moments/cumulants for a specific choice of constraint functions.

Consider the gradient update given in (14) in vector form $\partial H(Y_o)/\partial\theta_{pq} = -(\boldsymbol{\lambda}^o)^T(\partial\boldsymbol{\alpha}^o/\partial\theta_{pq})$, and recall from (13) that $\boldsymbol{\lambda}^o = -(\boldsymbol{\beta}^o)^{-1}\boldsymbol{\alpha}^o$.

Thus, the gradient contribution from the $o^{\text{th}}$ output channel is $\partial H(Y_o)/\partial\theta_{pq} = \left(\boldsymbol{\alpha}^o\right)^T \left(\boldsymbol{\beta}^o\right)^{-T} \left(\partial\boldsymbol{\alpha}^o/\partial\theta_{pq}\right)$, where '-T' denotes inverse-transpose matrix operation. This can be observed to be the gradient of entropy with respect to the Givens angles where the entropy estimate is obtained using the exponential density estimate $p_o(\xi) = C(\boldsymbol{\lambda}^o)\exp\left(\sum_{k=1}^{m} \lambda_k^o f_k(\xi)\right)$ for the output signal $Y_o$. This entropy estimate is found to be $H(Y_o) = -\log C(\boldsymbol{\lambda}^o) - (\boldsymbol{\lambda}^o)^T \boldsymbol{\alpha}^o = -\log C(\boldsymbol{\lambda}^o) + \left(\boldsymbol{\alpha}^o\right)^T \left(\boldsymbol{\beta}^o\right)^{-T} \boldsymbol{\alpha}^o$. Therefore, the proposed algorithm is understood to be essentially minimizing the following cost function.

$$J(\boldsymbol{\theta}) = \sum_{o=1}^{n} H(Y_o) = \sum_{o=1}^{n}\left(-\log C(\boldsymbol{\lambda}^o) + \left(\boldsymbol{\alpha}^o\right)^T \left(\boldsymbol{\beta}^o\right)^{-T} \boldsymbol{\alpha}^o\right) \tag{19}$$

Our brief investigation on the effect of the selected constraint functions on the separation performance of the algorithm revealed that the simple moments of the form $f_k(x) = x^k$ yield significantly better solutions.[2] Therefore, this choice of constraints becomes particularly interesting for further analysis. One motivation for using these constraint functions is the asymptotic properties of the exponential

[2] We have performed Monte Carlo comparisons using the following heuristically selected alternative constraint functions: $f_k(x)=\arctan(kx)$, $f_k(x)=|x|^{1/k}$, $f_k(x)=\tan(kx)$, and $f_k(x)=e^{-k|x|}$. The cost functions associated with these alternatives exhibited numerous local minima, which hindered the performance of the gradient descent algorithm greatly. The performance surface corresponding to the moment constraints was found to be much smoother.

density estimates of the form $p(\xi) = \exp\left(\lambda_0 + \sum_{k=1}^{m} \lambda_k \xi^k\right)$. Consider continuous infinite support distributions that could be approximated by the following Taylor series expansion: $\log q(\xi) = \sum_{k=0}^{\infty}\left(\xi^k / k!\right)\left(\partial^k (\log q(\xi)) / \partial \xi^k \Big|_{q(\xi_*)=1}\right)$. It is known that if the Taylor series expansion exists and the infinite summation of $q(x)$ converges uniformly, then the exponential density of order $m$ converges uniformly as the order $m\to\infty$ (Barndorff-Nielsen, 1978; Crain, 1974). In addition, since the family of exponential distributions form a linear manifold with orthogonality properties in the space of natural parameters, the maximum entropy distribution is the orthogonal projection of the $\infty$-dimensional density to the $m$-dimensional linear manifold (Amari, 1985; Crain, 1974).

Besides the desirable asymptotic convergence properties of the exponential family of density estimates, the selection of moments as constraint functions result in gradient updates that are simply gradients of higher order-moments with respect to the Givens angles. Specifically, for this selection of constraints the gradient for the cost function in (19) becomes

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_{pq}} = \sum_{o=1}^{n} \frac{\partial H(Y_o)}{\partial \theta_{pq}} = \sum_{o=1}^{n} \left(\boldsymbol{\alpha}^o\right)^T \left(\boldsymbol{\beta}^o\right)^{-T} \left(\frac{\partial \boldsymbol{\alpha}^o}{\partial \theta_{pq}}\right)$$

$$= \sum_{o=1}^{N} \left( \begin{array}{l} \begin{bmatrix} \alpha_1^o & \cdots & \alpha_m^o \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & (m+1) \end{bmatrix} \begin{bmatrix} \alpha_2^o & \cdots & \alpha_{m+1}^o \\ \vdots & \ddots & \vdots \\ \alpha_{m+1}^o & \cdots & \alpha_{2m}^o \end{bmatrix}^{-1} \cdot \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1/m \end{bmatrix} \begin{bmatrix} \partial \alpha_1^o / \partial \theta_{pq} \\ \vdots \\ \partial \alpha_m^o / \partial \theta_{pq} \end{bmatrix} \end{array} \right) \qquad (20)$$

where $\alpha_k^o = \frac{1}{N}\sum_{j=1}^{N} y_{o,j}^k$ .

Consider the simple case, where $m$=4, constraint functions are the moments, and the source distributions (and the particular samples in the finite-sample case) are symmetric[3]. In this case, the odd moments vanish, and also due to the whitening, the second order moments are fixed to unity. Thus the gradient in (20) becomes

$$\begin{aligned}\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_{pq}} &= \sum_{o=1}^{N}\left(\begin{array}{l}\begin{bmatrix}0 & 1 & 0 & \alpha_4^o\end{bmatrix} diag(2,\ldots,5)\begin{bmatrix}1 & 0 & \alpha_4^o & 0\\ 0 & \alpha_4^o & 0 & \alpha_6^o\\ \alpha_4^o & 0 & \alpha_6^o & 0\\ 0 & \alpha_6^o & 0 & \alpha_8^o\end{bmatrix}^{-1}\cdot\\ \qquad diag(1,\ldots,1/m)\begin{bmatrix}0 & 0 & 0 & \partial\alpha_4^o/\partial\theta_{pq}\end{bmatrix}^T\end{array}\right)\\ &= \sum_{o=1}^{n}\frac{\left(5\left(\alpha_4^o\right)^2 - 3\alpha_6^o\right)}{4\left(\alpha_4^o\alpha_8^o - \left(\alpha_6^o\right)^2\right)}\frac{\partial\alpha_4^o}{\partial\theta_{pq}}\end{aligned} \qquad (21)$$

The directional contribution from each output to the gradient is along the gradient of the kurtosis of that output. The sign and the magnitude are controlled

[3] In the finite case, the odd-sample-moments need not become zero. If we extend the observation sample set by including –**x** samples (so that the total set consists of $\{\mathbf{x}_k, -\mathbf{x}_k\}$ $k$=1,…,$N$), the odd-sample-moments of the extended data set becomes zero. This corresponds to modifying the source distributions to become symmetric, yet all the measurements are mixtures of the *new* symmetric sources through the same mixing matrix **H**.

by the term $(5(\alpha_4^o)^2 - 3\alpha_6^o)/(\alpha_4^o\alpha_8^o - (\alpha_6^o)^2)$.[4] In order to demonstrate how the sign of this term detects sub- or super-Gaussianity to adjust the update direction accordingly, we present an evaluation of this quantity for the generalized Gaussian family in Fig. 1 (sample approximations with 30000 samples are utilized for each distribution). Since the cost function in (19) is minimized, negative-gradient direction is used so the updates using the gradient in (21) try to minimize the kurtosis for sub-Gaussian signals and maximize kurtosis for super-Gaussian signals (as expected). In general, the overall gradient includes the terms up to $\partial\alpha_m^o \big/ \partial\theta_{pq}$, therefore the update direction is not only determined by the gradients of the output kurtosis, but also the gradients of higher order moments.

[4] Notice that the constants 3 and 5 in this expression correspond to the fourth and sixth order moments of a zero-mean unit-variance Gaussian distribution. Consequently, if the output distribution approaches a unit-variance Gaussian, the numerator approaches zero.

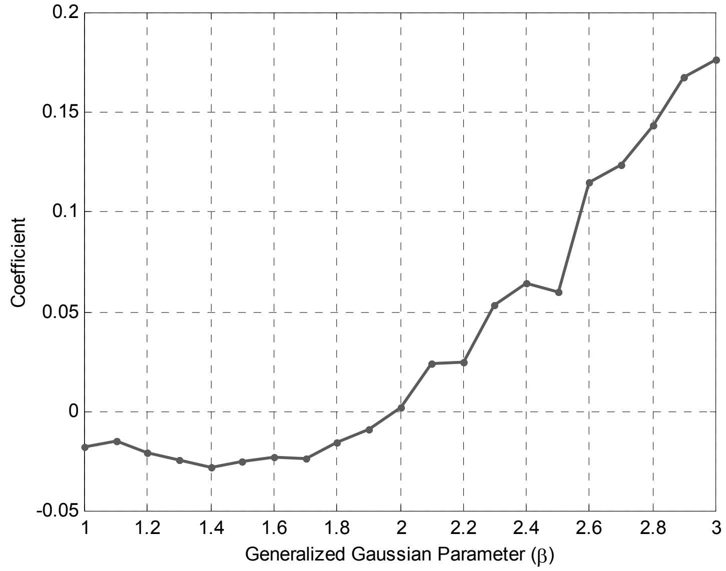

Figure 1. The coefficient $(5(\alpha_4^o)^2 - 3\alpha_6^o)/(\alpha_4^o\alpha_8^o - (\alpha_6^o)^2)$ evaluated for the generalized Gaussian family for a range of the parameter. The generalized Gaussian family includes Laplacian (β=1), Gaussian (β=2), and Uniform (β→∞) as special cases.

According to the analysis above, for the choice of moment constraints, Minimax ICA becomes an auto-cumulant method (identical to Fast ICA, in principle, when moments up to order 4 are considered). Other cumulant methods, for instance JADE (Cardoso, 1999), consider cross-cumulants of the outputs. If, at the density estimation stage we employ the maximum entropy principle to obtain an estimate of the joint output distribution, which can then be utilized to find the marginal output distributions, the resulting algorithm would also involve updates based on the cross-moments/cumulants of the outputs. This extension to cross-cumulants will be demonstrated and studied in a future paper.

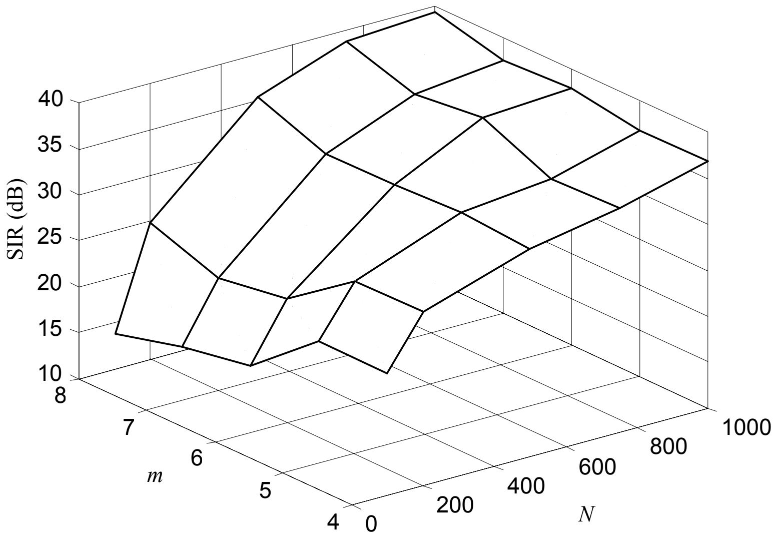

Figure 2. Average performance of the Minimax ICA algorithm using the sample moments as constraints evaluated for combinations of sample size and number of constraints.

## 6. Numerical Results

In this section, we will investigate the performance of the proposed Minimax ICA algorithm. Comparisons with other benchmark ICA algorithms will be provided. The performance measure that will be used throughout this section is the commonly used average signal-to-interference ratio (SIR)

$$SIR\ \ (dB) = \frac{1}{n}\sum_{o=1}^{n} 10\log_{10} \frac{\max\limits_{k}(\mathbf{O}_{ok}^{2})}{\mathbf{O}_{o:}\mathbf{O}_{o:}^{T} - \max\limits_{k}(\mathbf{O}_{ok}^{2})} \tag{22}$$

where **O** is the overall matrix after separation, i.e. **O**=**RWH** (Hild *et al.*, 2001).[5] In each row, the source corresponding to the maximum entry is assumed to be the

[5] The subscript '*o*:' means the $o^{th}$ row and '*ok*' indicates the corresponding entry.

main signal at that output. The averaging is done over the dB values of each row to eliminate the possibility of one very good row becoming dominant in the SIR calculation. Although for arbitrarily selected **O** matrices, the above SIR measure could have some problems (e.g., if two rows of **O** are identical with one dominant entry, SIR would be large, but source separation would not have been achieved, since two outputs would yield the same source signal), such occurrences are prevented by the selected topology, i.e. the whitening-rotation scheme. As long as the mixing matrix is invertible and all sources have non-zero power, the overall matrix will be full rank and this problem is avoided.

Our first case study investigates the effect of sample size and the number of constraints on the performance of the Minimax ICA algorithm. In this experiment, we use a 2x2 mixture, where the sources are zero-mean, independent Gaussian and uniformly distributed random variables. For each combination of sample size and number of constraints ($m$,$N$) selected from the sets {4,5,6,7,8} and {100,200,500,750,1000} respectively, we performed 100 Monte Carlo simulations. In each run, a mixing matrix **H** is selected randomly (each entry uniform in [-1,1]) and new independent and identically distributed (iid) samples are generated. The constraint functions are selected as the moments. The average SIR levels obtained for each combination are shown in Fig. 2. As expected, regardless of the number of constraints, increasing the number of training samples increases performance. In addition, the worst performance is obtained for the combination $m=8$ $N$=100. This is also expected since the estimation of higher order moments require a larger sample set for robust estimation. As a consequence

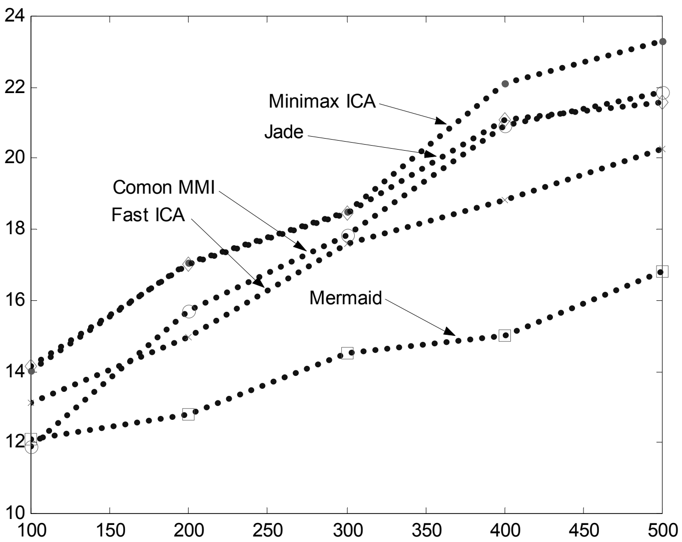

Figure 3. Average SIR (dB) obtained by Minimax ICA (dot), JADE (diamond), Comon MMI (circle), Fast ICA (cross), and Mermaid (square) versus the number of training samples.

the performance of the Minimax ICA algorithm suffers from sensitivity to samples. However, we note that the performance increases as the number of constraints (the order of moments) is increased (under the condition that the number of samples is sufficient to accurately estimate these higher order statistics). In conclusion, if the training set is small one must use a small number of constraints (lower order moments), and if the training set is large, one can increase the number of constraints (the order of moments) without compromising robustness.

In the second case study, we will conduct Monte Carlo simulations to compare the performance of five algorithms based on the minimum output mutual information contrast: Minimax ICA (with 4 moment constraints), JADE (Cardoso, 1999), Comon's minimum mutual information algorithm (MMI) (Comon, 1994), Fast ICA (Hyvarinen, 1999b), minimum Renyi's mutual information algorithm – Mermaid (Hild *et al.*, 2001). Although we initially aimed to include Yang and Amari's minimum mutual information method (Yang and Amari, 1997), the preliminary results obtained with it were not competitive with the five methods listed. Among the considered five approaches, Mermaid was the computationally most expensive one, whose requirements increase as $O(N^2)$ with the number of samples (Hild *et al.*, 2001) followed by Minimax ICA, whose complexity increases as $O(Nm)$ with the number of samples ($N$) and the number of moment constraints ($m$).

In each run, $N$ samples of a source vector composed of one Gaussian, one Laplacian (super-Gaussian), and one uniformly (sub-Gaussian) distributed entry were generated. A 3x3 random mixing matrix, whose entries are selected from the interval [-1,1], is also fixed. The mixed signals are then fed into the four algorithms. Evaluation of the overall performance is done by considering the average final SIR value obtained by each algorithm after convergence is achieved. The averaging is done over 100 runs and the results are summarized in Fig. 3. According to these batch-adaptation results, Minimax ICA and JADE perform identically for very small number of samples, outperforming the other three methods. However, as the number of samples increase, Minimax ICA takes more advantage of this and yields better separation solutions. The example presented here is a particularly difficult case since all three types of distributions (Gaussian, super-Gaussian, and sub-Gaussian) are represented in the sources. Experiments (not shown here) performed with all-same-type source distributions (with not more than one Gaussian source) showed that the performance of all algorithms increase significantly and the difference between performance becomes much less, especially for small training sets.

## 7. Conclusions

Jaynes' maximum entropy principle has found successful applications in many areas, starting with statistical physics and including the problem of spectral estimation. It basically states that one should use the probabilistic model that best describes the observed data, yet commits minimally to any possible unseen data. In order to achieve this, the density estimates are obtained through a constrained entropy maximization procedure. The constraints assure that the final density is consistent with the current data. On the other hand, entropy maximization guarantees the selection of the model with maximum uncertainty, in other words, the model that is least committed to unseen data. In this paper, we proposed a novel ICA algorithm that is based on Jaynes' maximum entropy principle in the pdf estimation step. The commonly used whitening-rotation topology is borrowed from the literature, whereas the criterion used, minimum output mutual information, is considered to be the natural information theoretic measure for ICA. We have shown that the Lagrange multipliers of the maximum entropy pdf can be easily estimated from the samples by solving a system of linear equations. In addition, the gradient of the output mutual information with respect to the rotation angles, which characterize the rotation matrix, turned out to have a simple expression in terms of these Lagrange multipliers.

Numerical case studies have shown that the sample moments form a useful set of constraint functions that result in a smooth cost function, free of local minima, and with accurate solutions. Although the specific alternative nonlinear constraint functions investigated in this paper resulted in surfaces with many local minima, for some applications the designer might have *a priori* knowledge about the form of the constraints that the data might satisfy. The selection of these functions provides some freedom to the designer in that respect.

Comparisons with benchmark algorithms like Comon's MMI, Fast ICA, and Jade in problems involving mixed-kurtosis sources (Gaussian, sub-Gaussian, and super-Gaussian) showed that the average performance of Minimax ICA equal to that of JADE for small sample sets and gets better with increasing number of samples. In simulations not reported in this paper, the authors have observed that in cases where all sources are uniformly distributed (or have densities with light tails), Fast ICA provides very good results that compete with Minimax ICA. On the other hand, since Minimax ICA specifically looks for the maximum entropy density estimates, in some extreme situations, the performance could degrade, especially if the actual densities do not belong to the exponential family arising from the maximum entropy assumption. Future investigation will focus on extensions of the Minimax ICA algorithm to cross-cumulants by incorporating the joint moments of the outputs into the estimation. In addition, we will determine the effect of using series expansions other than Taylor's (which leads to the moment constraints). This will allow the algorithm to extend beyond the exponential family of sources and the cumulant-based weight updates.

**Acknowledgments:** The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This work is partially supported by NSF grant ECS-0300340. Parts of this paper was presented at the ICA 2003 conference at Nara, Japan (Erdogmus *et al.*, 2003).

## Appendix A

In this appendix, we present the step-by-step derivation of the derivative of output marginal entropy with respect to one of the Givens angles.

$$
\begin{aligned}
\frac{\partial H(Y_o)}{\partial \theta_{pq}} &= -\frac{\partial}{\partial \theta_{pq}} \int_{-\infty}^{\infty} p_o(\xi) \log p_o(\xi) d\xi \\
&= -\int_{-\infty}^{\infty} \left[ \frac{\partial p_o(\xi)}{\partial \theta_{pq}} \log p_o(\xi) + p_o(\xi) \frac{\partial p_o(\xi) / \partial \theta_{pq}}{p_o(\xi)} \right] d\xi \\
&= -\int_{-\infty}^{\infty} \frac{\partial p_o(\xi)}{\partial \theta_{pq}} \log p_o(\xi) d\xi - \int_{-\infty}^{\infty} \frac{\partial p_o(\xi)}{\partial \theta_{pq}} d\xi \\
&= -\int_{-\infty}^{\infty} \frac{\partial p_o(\xi)}{\partial \theta_{pq}} \left( C^o(\boldsymbol{\lambda}^o) + \sum_{k=1}^{m} \lambda_k^o f_k(\xi) \right) d\xi - \frac{\partial}{\partial \theta_{pq}} \int_{-\infty}^{\infty} p_o(\xi) d\xi \\
&= -(1 + C^o(\boldsymbol{\lambda}^o)) \frac{\partial}{\partial \theta_{pq}} \int_{-\infty}^{\infty} p_o(\xi) d\xi - \sum_{k=1}^{m} \lambda_k^o \int_{-\infty}^{\infty} \frac{\partial p_o(\xi)}{\partial \theta_{pq}} f_k(\xi) d\xi \\
&= -\sum_{k=1}^{m} \lambda_k^o \frac{\partial}{\partial \theta_{pq}} \int_{-\infty}^{\infty} p_o(\xi) f_k(\xi) d\xi = -\sum_{k=1}^{m} \lambda_k^o \frac{\partial \alpha_k^o}{\partial \theta_{pq}}
\end{aligned}
$$

## References

Almeida, L.B. (2000). Linear and nonlinear ICA based on mutual information. *Proceedings of AS-SPCC'00, Lake Louise, Canada*, NJ: IEEE Press, 117-122.

Amari, S.I. (1985). *Differential–geometrical methods in statistics*. Berlin: Springer-Verlag.

Amari, S.I. (1997). Neural learning in structured parameter spaces – natural riemmanian gradient. *Proceedings of NIPS'97*, MA: MIT Press, 127-133.

Amari, S.I., Cichocki, A., and Yang, H.H. (1996). A new learning algorithm for blind signal separation. *Proceedings of NIPS'96*, MA: MIT Press, 757-763.

Barndorff-Nielsen, O.E. (1978). *Information and exponential families in statistical theory*. Chichester: Wiley.

Bell, T., and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, 1129-1159.

Cardoso, J.F. (1994). On the performance of orthogonal source separation algorithms. *Proceedings of EUSIPCO'94*, NJ: IEEE Press, 776-779.

Cardoso, J.F. (1998). Blind signal separation: statistical principles. *Proceedings of IEEE*, 86, NJ: IEEE Press, 2009-2025.

Cardoso, J.F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, 11, 157-192.

Cardoso, J.F., and Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. *IEE Proceedings F Radar and Signal Processing*, 140, 362-370.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36, 287-314.

Comon, P. (1996). Constrasts for multichannel blind deconvolution. *IEEE Signal Processing Letters*, 3, 209-211.

Cover, T.M., and Thomas, J.A. (1991). *Elements of Information theory*. New York: Wiley.

Crain, B.R. (1974). Estimation of distributions using orthogonal expansions. *Annals of Statistics*, 2, 454-463.

Erdogmus, D., Hild II, K.E., and Principe, J.C. (2001). Independent component analysis using Renyi's mutual information and Legendre density estimation. *Proceedings of IJCNN'01*, NJ: IEEE Press, 2762-2767.

Erdogmus, D., Hild II, K.E., Rao, Y.N., and Principe, J.C. (2003). Independent components analysis using Jaynes' maximum entropy principe. *Proceedings of ICA'03*, 385-390 (http://www.kecl.ntt.co.jp/icl/signal/ica2003/cdrom/index.htm).

Girolami, M., and Fyfe, C. (1997). Kurtosis extrema and identification of independent components: a neural network approach. *Proceedings of ICASSP'97*, NJ: IEEE Press, 3329-3332.

Girolami, M. (1997). Symmetric adaptive maximum likelihood estimation for noise cancellation and signal estimation. *Electronic Letters*, 33, 1437-1438.

Girolami, M. (2002). Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14, 669-688.

Golub, G., and Van Loan, C. (1993). *Matrix Computation*. Baltimore: John Hopkins University Press.

Hild II, K.E., Erdogmus, D., and Principe, J.C. (2001). Blind source separation using Renyi's mutual information. *IEEE Signal Processing Letters*, 8, 174-176.

Hyvarinen, A. (1998). New approximations of differential entropy for independent component analysis and projection pursuit. *Proceedings of NIPS'98*, MA: MIT Press, 273-279.

Hyvarinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, 2, 94-128.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10, 626-634.

Jaynes, E.T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, 106, 620-630.

Kapur, J., and Kesavan, H. (1992). *Entropy optimization principles and applications*. New York: Academic Press.

Karvanen, J., Eriksson, J., and Koivunen, V. (2000). Maximum likelihood estimation of ICA model for wide class of source distributions. *Proceedings of NNSP'00*, NJ: IEEE Press, 445-454.

Oja, E. (1999). The nonlinear PCA learning rule in independent component analysis. *Proceedings of ICA'99*, 143-148.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33, 1065-1076.

Parra, L., and Spence, C. (2000). Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech Audio Processing*, 46, 320-327.

Pham, D.T. (1996). Blind separation of instantaneous mixture sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44, 2768-2779.

Pham, D.T. (2001). Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion. *Signal Processing*, 81, 855-870.

Principe, J.C., and Xu, D. (1999). Information theoretic learning using Renyi's quadratic entropy. *Proceedings of ICA'99*, 407-412.

Renyi, A. (1970). *Probability Theory*. Amsterdam: North-Holland.

Simon, C., Loubaton, P., Vignat, C., Jutten, C., and d'Urso, G. (1998). Blind source separation of convolutive mixtures by maximizing of fourth-order cumulants: The non-IID case. *Proceedings of ACSSC'98*, NJ: IEEE Press, 1584-1588.

Shore, J.E., and Johnson, R.W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26, 26-37.

Sun, X., Douglas, S.C. (2001). Adaptive paraunitary filter banks for contrast-based multichannel blind deconvolution. *Proceedings of ICASSP'01*, NJ: IEEE Press, 2753-2756.

Torkkola, K. (1996). Blind separation of delayed sources based on information maximization. *Proceedings of NNSP'96*, NJ: IEEE Press, 1-10.

Torkkola, K. (1999). Blind separation for audio signals – are we there yet? *Proceedings of ICA'99*, 239-244.

Weinstein, E., Feder, M., and Oppenheim, A. (1993). Multi-channel signal separation by decorrelation. *IEEE Transactions on Speech Audio Processing*, 1, 405-413.

Wu, H.C., and Principe, J.C. (1997). A unifying criterion for blind source separation and decorrelation: Simultaneous diagonalization of correlation matrices. *Proceedings of NNSP'97*, NJ: IEEE Press, 465-505.

Wu, H.C., and Principe, J.C. (1999). A Gaussianity measure for blind source separation insensitive to the sign of kurtosis. *Proceedings of NNSP'99*, NJ: IEEE Press, 58-66.

Wu, H.C., Principe, J.C. (1999). Generalized anti-Hebbian learning for source separation. *Proceedings of ICASSP'99*, NJ: IEEE Press, 1073-1076.

Xu, D., Principe, J.C., Fisher, J., and Wu, H.C. (1998). A novel measure for independent component analysis. *Proceedings of ICASSP'98*, NJ: IEEE Press, 1161-1164.

Yang, H.H., and Amari, S.I. (1997). Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information. *Neural Computation*, 9, 1457-1482.

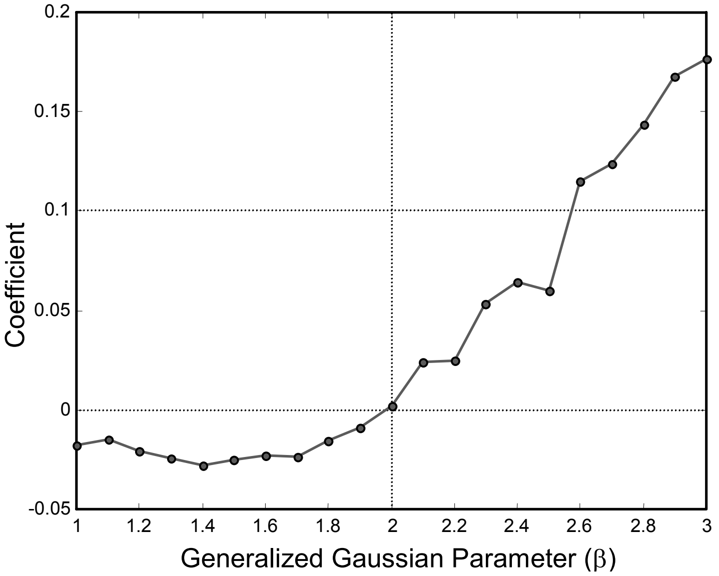

Figure 1 of MS 2709: Erdogmus

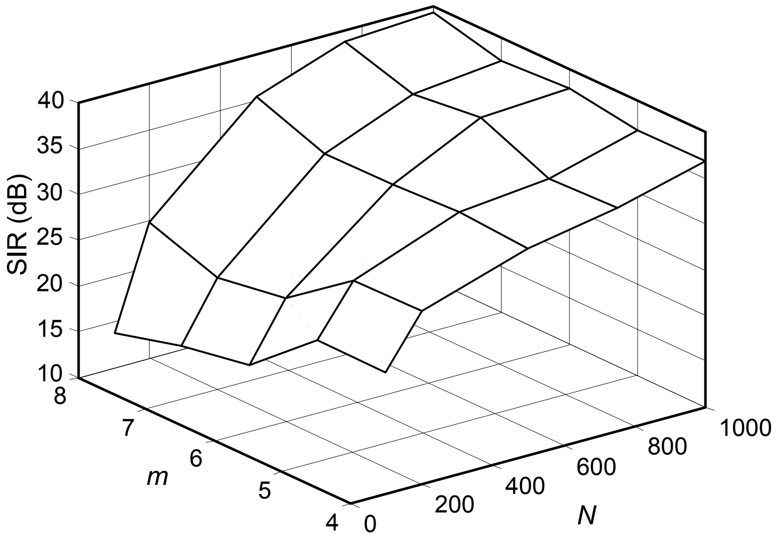

Figure 2 of MS 2709: Erdogmus

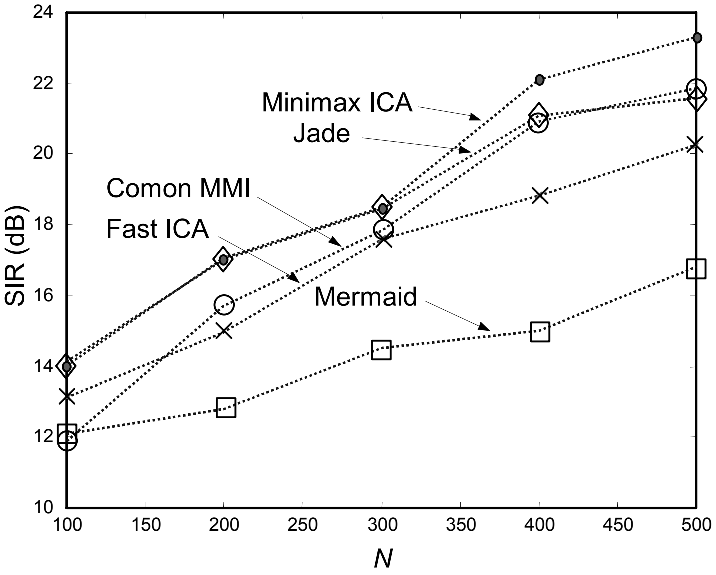

Figure 3 of MS 2709: Erdogmus