LETTER Communicated by Anthony Bell

High-Order Contrasts for Independent Component Analysis

Jean-François Cardoso
Ecole Nationale Supérieure des Télécommunications, 75634 Paris Cedex 13, France

This article considers high-order measures of independence for the independent component analysis problem and discusses the class of Jacobi algorithms for their optimization. Several implementations are discussed. We compare the proposed approaches with gradient-based techniques from the algorithmic point of view and also on a set of biomedical data.

1 Introduction

Given an $n \times 1$ random vector X, independent component analysis (ICA) consists of finding a basis of $\mathbf{R}^n$ on which the coefficients of X are as independent as possible (in some appropriate sense). The change of basis can be represented by an $n \times n$ matrix B and the new coefficients given by the entries of vector $Y = BX$. When the observation vector X is modeled as a linear superposition of source signals, matrix B is understood as a separating matrix, and vector $Y = BX$ is a vector of source signals. Two key issues of ICA are the definition of a measure of independence and the design of algorithms to find the change of basis (or separating matrix) B optimizing this measure.

Many recent contributions to the ICA problem in the neural network literature describe stochastic gradient algorithms involving as an essential device in their learning rule a nonlinear activation function. Other ideas for ICA, most of them found in the signal processing literature, exploit the algebraic structure of high-order moments of the observations. They are often regarded as being unreliable, inaccurate, slowly convergent, and utterly sensitive to outliers. As a matter of fact, it is fairly easy to devise an ICA method displaying all these flaws and working on only carefully generated synthetic data sets. This may be the reason that cumulant-based algebraic methods are largely ignored by the researchers of the neural network community involved in ICA. This article tries to correct this view by showing how high-order correlations can be efficiently exploited to reveal independent components.

This article describes several ICA algorithms that may be called Jacobi algorithms because they seek to maximize measures of independence by a technique akin to the Jacobi method of diagonalization. These measures of independence are based on fourth-order correlations between the entries of Y. As a benefit, these algorithms evades the curse of gradient descent: they can move in macroscopic steps through the parameter space. They also have other benefits and drawbacks, which are discussed in the article and summarized in a final section. Before outlining the content of this article, we briefly review some gradient-based ICA methods and the notion of contrast function.

1.1 Gradient Techniques for ICA.

Many online solutions for ICA that have been proposed recently have the merit of a simple implementation. Among these adaptive procedures, a specific class can be singled out: algorithms based on a multiplicative update of an estimate $B(t)$ of B. These algorithms update a separating matrix $B(t)$ on reception of a new sample $x(t)$ according to the learning rule

$$y(t) = B(t)x(t), \qquad B(t+1) = \big(I - \mu_t H(y(t))\big)\, B(t), \tag{1.1}$$

where I denotes the $n \times n$ identity matrix, $\{\mu_t\}$ is a scalar sequence of positive learning steps, and H: $\mathbf{R}^n \to \mathbf{R}^{n\times n}$ is a vector-to-matrix function. The stationary points of such algorithms are characterized by the condition that the update has zero mean, that is, by the condition,

$$\mathrm{E}H(Y) = 0. \tag{1.2}$$

The online scheme, in equation 1.1, can be (and often is) implemented in an off-line manner. Using T samples $X(1), \ldots, X(T)$, one goes through the following iterations where the field H is averaged over all the data points:

1. *Initialization*. Set $y(t) = x(t)$ for $t = 1, \ldots, T$.
2. *Estimate the average field*. $\mathcal{H} = \frac{1}{T}\sum_{t=1}^{T} H(y(t))$.
3. *Update*. If $\mathcal{H}$ is small enough, stop; else update each data point $y(t)$ by $y(t) \leftarrow (I - \mu\mathcal{H})y(t)$ and go to 2.

The algorithm stops for a (arbitrarily) small value of the average field: it solves the estimating equation,

$$\frac{1}{T}\sum_{t=1}^{T} H(y(t)) = 0, \tag{1.3}$$

which is the sample counterpart of the stationarity condition in equation 1.2.

Both the online and off-line schemes are gradient algorithms: the mapping $H(\cdot)$ can be obtained as the gradient (the relative gradient [Cardoso & Laheld, 1996] or Amari's natural gradient [1996]) of some contrast function, that is, a real-valued measure of how far the distribution Y is from some ideal distribution, typically a distribution of independent components. In particular, the gradient of the infomax—maximum likelihood (ML) contrast yields a function $H(\cdot)$ in the form

$$H(y) = \psi(y)y^\dagger - I, \tag{1.4}$$

where $\psi(y)$ is an $n \times 1$ vector of component-wise nonlinear functions with $\psi_i(\cdot)$ taken to be minus the log derivative of the density of the i component (see Amari, Cichocki, & Yang, 1996, for the online version and Pham & Garat, 1997, for a batch technique).

1.2 The Orthogonal Approach to ICA.

In the search for independent components, one may decide, as in principal component analysis (PCA), to request exact decorrelation (second-order independence) of the components: matrix B should be such that $Y = BX$ is "spatially white," that is, its covariance matrix is the identity matrix. The algorithms described in this article take this design option, which we call the orthogonal approach.

It must be stressed that components that are as independent as possible according to some measure of independence are not necessarily uncorrelated because exact independence cannot be achieved in most practical applications. Thus, if decorrelation is desired, it must be enforced explicitly; the algorithms described below optimize under the whiteness constraint approximations of the mutual information and of other contrast functions (possibly designed to take advantage of the whiteness constraint).

One practical reason for considering the orthogonal approach is that off-line contrast optimization may be simplified by a two-step procedure as follows. First, a whitening (or "sphering") matrix W is computed and applied to the data. Since the new data are spatially white and one is also looking for a white vector Y, the latter can be obtained only by an orthonormal transformation V of the whitened data because only orthonormal transforms can preserve the whiteness. Thus, in such a scheme, the separating matrix B is found as a product $B = VW$. This approach leads to interesting implementations because the whitening matrix can be obtained straightforwardly as any matrix square root of the inverse covariance matrix of X and the optimization of a contrast function with respect to an orthonormal matrix can also be implemented efficiently by the Jacobi technique described in section 4.

The orthonormal approach to ICA need not be implemented as a two-stage Jacobi-based procedure; it also exists as a one-stage gradient algorithm (see also Cardoso & Laheld, 1996). Assume that the relative/natural gradient of some contrast function leads to a particular function $H(\cdot)$ for the update rule, equation 1.1, with stationary points given by equation 1.2. Then the stationary points for the optimization of the same contrast function with respect to orthonormal transformations are characterized by $\mathrm{E}H(Y) - H(Y)^\dagger = 0$ where the superscript † denotes transposition. On the other hand, for zero-mean variables, the whiteness constraint is $\mathrm{E}YY^\dagger = I$, which we can also write as $\mathrm{E}YY^\dagger - I = 0$. Because $\mathrm{E}YY^\dagger - I$ is a symmetric matrix matrix while $\mathrm{E}H(Y) - H(Y)^\dagger$ is a skew-symmetric matrix, the whiteness condition and the stationarity condition can be combined in a single one by just adding them. The resulting condition is $\mathrm{E}\{YY^\dagger - I + H(Y) - H(Y)^\dagger\} = 0$. When it holds true, both the symmetric part and the skew-symmetric part cancel; the former expresses that Y is white, the latter that the contrast function is stationary with respect to all orthonormal transformations.

Thus, if the algorithm in equation 1.1 optimizes a given contrast function with H given by equation 1.4, then the same algorithm optimizes the same contrast function under the whiteness constraint with H given by

$$H(y) = yy^\dagger - I + \psi(y)y^\dagger - y\psi(y)^\dagger. \tag{1.5}$$

It is thus simple to implement orthogonal versions of gradient algorithms once a regular version is available.

1.3 Data-Based Versus Statistic-Based Techniques.

Comon (1994) compares the data-based option and the statistic-based option for computing off-line an ICA of a batch $x(1), \ldots, x(T)$ of T samples; this article will also introduce a mixed strategy (see section 4.3). In the data-based option, successive linear transformations are applied to the data set until some criterion of independence is maximized. This is the iterative technique outlined above. Note that it is not necessary to update explicitly a separating matrix B in this scheme (although one may decide to do so in a particular implementation); the data themselves are updated until the average field $\frac{1}{T}\sum_{t=1}^{T} H(y(t))$ is small enough; the transform B is implicitly contained in the set of transformed data.

Another option is to summarize the data set into a smaller set of statistics computed once and for all from the data set; the algorithm then estimates a separating matrix as a function of these statistics without accessing the data. This option may be followed in cumulant-based algebraic techniques where the statistics are cumulants of X.

1.4 Outline of the Article.

In section 2, the ICA problem is recast in the framework of (blind) identification, showing how entropic contrasts readily stem from the maximum likelihood (ML) principle. In section 3, high-order approximations to the entropic contrasts are given, and their algebraic structure is emphasized. Section 4 describes different flavors of Jacobi algorithms optimizing fourth-order contrast functions. A comparison between Jacobi techniques and a gradient-based algorithm is given in section 5 based on a real data set of electroencephalogram (EEG) recordings.

2 Contrast Functions and Maximum Likelihood Identification

Implicitly or explicitly, ICA tries to fit a model for the distribution of X that is a model of independent components: $X = AS$, where A is an invertible $n \times n$ matrix and S is an $n \times 1$ vector with independent entries. Estimating the parameter A from samples of X yields a separating matrix $B = A^{-1}$. Even if the model $X = AS$ is not expected to hold exactly for many real data sets, one can still use it to derive contrast functions. This section exhibits the contrast functions associated with the estimation of A by the ML principle (a more detailed exposition can be found in Cardoso, 1998). Blind separation based on ML was first considered by Gaeta and Lacoume (1990) (but the authors used cumulant approximations as those described in section 3), Pham and Garat (1997), and Amari et al. (1996).

2.1 Likelihood. Assume that the probability distribution of each entry S_i of S has a density $r_i(\cdot)$.[1] Then, the distribution $\mathcal{P}_S$ of the random vector S has a density $r(\cdot)$ in the form $r(s) = \prod_{i=1}^n r_i(s_i)$, and the density of X for a given mixture A and a given probability density $r(\cdot)$ is:

$$p(x; A, r) = |\det A|^{-1} r(A^{-1}x), \tag{2.1}$$

so that the (normalized) log-likelihood $L_T(A, r)$ of T independent samples $x(1), \ldots, x(T)$ of X is

$$\begin{aligned} L_T(A, r) &\overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \log p(x(t); A, r) \\ &= \frac{1}{T}\sum_{t=1}^{T} \log r(A^{-1}x(t)) - \log|\det A|. \end{aligned} \tag{2.2}$$

Depending on the assumptions made about the densities $r_1, \ldots, r_n$, several contrast functions can be derived from this log-likelihood.

2.2 Likelihood Contrast. Under mild assumptions, the normalized log-likelihood $L_T(A, r)$, which is a sample average, converges for large T to its ensemble average by law of large numbers:

$$\begin{aligned} L_T(A, r) &= \frac{1}{T}\sum_{t=1}^{T} \log r(A^{-1}x(t)) - \log|\det A| \\ &\longrightarrow_{T\to\infty} \mathrm{E}\log r(A^{-1}x) - \log|\det A|, \end{aligned} \tag{2.3}$$

which simple manipulations (Cardoso, 1997) show to be equal to $-\mathbf{H}(\mathcal{P}_\mathbf{X}) - \mathbf{K}(\mathcal{P}_\mathbf{Y}|\mathcal{P}_\mathbf{S})$. Here and in the following, $\mathbf{H}(\cdot)$ and $\mathbf{K}(\cdot|\cdot)$, respectively, denote the differential entropy and the Kullback-Leibler divergence. Since $\mathbf{H}(\mathcal{P}_\mathbf{X})$ does not depend on the model parameters, the limit for large T of $-L_T(A, r)$ is, up to a constant, equal to

$$\phi^{\mathrm{ML}}(Y) \overset{\text{def}}{=} \mathbf{K}(\mathcal{P}_Y|\mathcal{P}_S). \tag{2.4}$$

Therefore, the principle of ML coincides with the minimization of a specific contrast function, which is nothing but the (Kullback) divergence $\mathbf{K}(\mathcal{P}_Y|\mathcal{P}_S)$ between the distribution $\mathcal{P}_Y$ of the output and a model distribution $\mathcal{P}_S$.

The classic entropic contrasts follow from this observation, depending on two options: **(1)** trying or not to estimate $\mathcal{P}_S$ from the data and **(2)** forcing or not the components to be uncorrelated.

2.3 Infomax. The technically simplest statistical assumption about $\mathcal{P}_S$ is to select fixed densities $r_1, \ldots, r_n$ for each component, possibly on the basis of prior knowledge. Then $\mathcal{P}_S$ is a fixed distributional assumption, and the minimization of $\phi^{\mathrm{ML}}(Y)$ is performed only over $\mathcal{P}_Y$ via $Y = BX$. This can be rephrased: Choose B such that $Y = BX$ is as close as possible in distribution to the hypothesized model distribution $\mathcal{P}_S$, the closeness in distribution being measured in the Kullback divergence. This is also the contrast function derived from the infomax principle by Bell and Sejnowski (1995). The connection between infomax and ML was noted in Cardoso (1997), MacKay (1996), and Pearlmutter and Parra (1996).

2.4 Mutual Information. The theoretically simplest statistical assumption about $\mathcal{P}_S$ is to assume no model at all. In this case, the Kullback mismatch $\mathbf{K}(\mathcal{P}_Y|\mathcal{P}_S)$ should be minimized not only by optimizing over B to change the distribution of $Y = BX$ but also with respect to $\mathcal{P}_S$. For each fixed B, that is, for each fixed distribution $\mathcal{P}_Y$, the result of this minimization is theoretically very simple: the minimum is reached when $\mathcal{P}_S = \bar{\mathcal{P}}_Y$, which denotes the distribution of independent components with each marginal distribution equal to the corresponding marginal distribution of Y. This stems from the property that

$$\mathbf{K}(\mathcal{P}_Y|\mathcal{P}_S) = \mathbf{K}(\mathcal{P}_Y|\bar{\mathcal{P}}_Y) + \mathbf{K}(\bar{\mathcal{P}}_Y|\mathcal{P}_S) \tag{2.5}$$

for any distribution $\mathcal{P}_S$ with independent components (Cover & Thomas, 1991). Therefore, the minimum in $\mathcal{P}_S$ of $\mathbf{K}(\mathcal{P}_Y|\mathcal{P}_S)$ is reached by taking $\mathcal{P}_S = \bar{\mathcal{P}}_Y$ since this choice ensures $\mathbf{K}(\bar{\mathcal{P}}_Y|\mathcal{P}_S) = 0$. The value of ϕ^{ML} at this point then is

$$\phi^{\mathrm{MI}}(Y) \overset{\text{def}}{=} \min_{\mathcal{P}_S} \mathbf{K}(\mathcal{P}_Y|\mathcal{P}_S) = \mathbf{K}(\mathcal{P}_Y|\bar{\mathcal{P}}_Y). \tag{2.6}$$

[1] All densities considered in this article are with respect to the Lebesgue measure on $\mathbf{R}$ or $\mathbf{R}^n$.

We use the index MI since this quantity is well known as the mutual information between the entries of Y. It was first proposed by Comon (1994), and it can be seen from the above as deriving from the ML principle when optimization is with respect to both the unknown system A and the distribution of S. This connection was also noted in Obradovic and Deco (1997), and the relation between infomax and mutual information is also discussed in Nadal and Parga (1994).

2.5 Minimum Marginal Entropy. An orthogonal contrast $\phi(Y)$ is, by definition, to be optimized under the constraint that Y is spatially white: orthogonal contrasts enforce decorrelation, that is, an exact "second-order" independence. Any regular contrast can be used under the whiteness constraint, but by taking the whiteness constraint into account, the contrast may be given a simpler expression. This is the case of some cumulant-based contrasts described in section 3. It is also the case of $\phi^{\mathrm{MI}}(Y)$ because the mutual information can also be expressed as $\phi^{\mathrm{MI}}(Y) = \sum_{i=1}^{n} \mathbf{H}(\mathcal{P}_{\mathbf{Y_i}}) - \mathbf{H}(\mathcal{P}_{\mathbf{Y}})$; since the entropy $\mathbf{H}(\mathcal{P}_{\mathbf{Y}})$ is constant under orthonormal transforms, it is equivalent to consider

$$\phi^{\mathrm{ME}}(Y) = \sum_{i=1}^{n} \mathbf{H}(\mathcal{P}_{\mathbf{Y_i}}) \tag{2.7}$$

to be optimized under the whiteness constraint $\mathrm{E}YY^{\dagger} = I$. This contrast could be called orthogonal mutual information, or the marginal entropy contrast. The minimum entropy idea holds more generally under any volume-preserving transform (Obradovic & Deco, 1997).

2.6 Empirical Contrast Functions. Among all the above contrasts, only ϕ^{ML} or its orthogonal version are easily optimized by a gradient technique because the relative gradient of ϕ^{ML} simply is the matrix $\mathrm{E}H(Y)$ with $H(\cdot)$ defined in equation 1.4. Therefore, the relative gradient algorithm, equation 1.1, can be employed using either this function $H(\cdot)$ or its symmetrized form, equation 1.5, if one chooses to enforce decorrelation. However, this contrast is based on a prior guess $\mathcal{P}_S$ about the distribution of the components. If the guess is too far off, algorithm will fail to discover independent components that might be present in the data. Unfortunately, evaluating the gradient of contrasts based on mutual information or minimum marginal entropy is more difficult because it does not reduce to the expectation of a simple function of Y; for instance, Pham (1996) minimizes explicitly the mutual information, but the algorithm involves a kernel estimation of the marginal distributions of Y. An intermediate approach is to consider a parametric estimation of these distributions as in Moulines, Cardoso, and Gassiat (1997) or Pearlmutter and Parra (1996), for instance. Therefore, all these contrasts require that the distributions of components be known, or approximated or estimated. As we shall see next, this is also what the cumulant approximations to contrast functions are implicitly doing.

3 Cumulants

This section presents higher-order approximations to entropic contrasts, some known and some novel. To keep the exposition simple, it is restricted to symmetric distributions (for which odd-order cumulants are identically zero) and to cumulants of orders 2 and 4. Recall that for random variables $X_1, \ldots, X_4$, second-order cumulants are $\mathrm{Cum}(X_1, X_2) \overset{\mathrm{def}}{=} \mathrm{E}\bar{X}_1\bar{X}_2$ where $\bar{X}_i \overset{\mathrm{def}}{=} X_i - \mathrm{E}X_i$ and the fourth-order cumulants are

$$\begin{aligned} &\mathrm{Cum}(X_1, X_2, X_3, X_4) = \\ &\quad \mathrm{E}\bar{X}_1\bar{X}_2\bar{X}_3\bar{X}_4 - \mathrm{E}\bar{X}_1\bar{X}_2\mathrm{E}\bar{X}_3\bar{X}_4 - \mathrm{E}\bar{X}_1\bar{X}_3\mathrm{E}\bar{X}_2\bar{X}_4 - \mathrm{E}\bar{X}_1\bar{X}_4\mathrm{E}\bar{X}_2\bar{X}_3. \end{aligned} \tag{3.1}$$

The variance and the kurtosis of a real random variable X are defined as

$$\begin{aligned} \sigma^2(X) &\overset{\mathrm{def}}{=} \mathrm{Cum}(X, X) = \mathrm{E}\bar{X}^2, \\ k(X) &\overset{\mathrm{def}}{=} \mathrm{Cum}(X, X, X, X) = \mathrm{E}\bar{X}^4 - 3\mathrm{E}^2\bar{X}^2, \end{aligned} \tag{3.2}$$

that is, they are the second- and fourth-order autocumulants. A cumulant involving at least two different variables is called a cross-cumulant.

3.1 Cumulant-Based Approximations to Entropic Contrasts. Cumulants are useful in many ways. In this section, they show up because the probability density of a scalar random variable U close to the standard normal $n(u) = (2\pi)^{-1/2} \exp -u^2/2$ can be approximated as

$$p(u) \approx n(u) \left(1 + \frac{\sigma^2(U) - 1}{2} h_2(u) + \frac{k(U)}{4!} h_4(u) \right), \tag{3.3}$$

where $h_2(u) = u^2 - 1$ and $h_4(u) = u^4 - 6u^2 + 3$, respectively, are the second- and fourth-order Hermite polynomials. This expression is obtained by retaining the leading terms in an Edgeworth expansion (McCullagh, 1987). If U and V are two real random variables with distributions close to the standard normal, one can, at least formally, use expansion 3.3 to derive an approximation to $\mathbf{K}(\mathcal{P}_U|\mathcal{P}_V)$. This is

$$\mathbf{K}(\mathcal{P}_U|\mathcal{P}_V) \approx \frac{1}{4}(\sigma^2(U) - \sigma^2(V))^2 + \frac{1}{48}(k(U) - k(V))^2, \tag{3.4}$$

which shows how the pair (σ^2, k) of cumulants of order 2 and 4 play in some sense the role of a local coordinate system around $n(u)$ with the quadratic form 3.4 playing the role of a local metric. This result generalizes to multivariates, in which case we denote for conciseness $R^U_{ij} = \mathrm{Cum}(U_i, U_j)$ and $Q^U_{ijkl} = \mathrm{Cum}(U_i, U_j, U_k, U_l)$ and similarly for another random n-vector V with entries $V_1, \ldots, V_n$. We give without proof the following approximation:

$$\mathbf{K}(\mathcal{P}_U|\mathcal{P}_V) \approx \mathbf{K}_{24}(\mathcal{P}_U|\mathcal{P}_V) \overset{\text{def}}{=} \frac{1}{4}\sum_{ij}\left(R^U_{ij} - R^V_{ij}\right)^2 + \frac{1}{48}\sum_{ijkl}\left(Q^U_{ijkl} - Q^V_{ijkl}\right)^2. \tag{3.5}$$

Expression 3.5 turns out to be the simplest possible multivariate generalization of equation 3.4 (the two terms in equation 3.5 are a double sum over all the n^2 pairs of indices and a quadruples over all the n^4 quadruples of indices). Since the entropic contrasts listed above have all been derived from the Kullback divergence, cumulant approximations to all these contrasts can be obtained by replacing the Kullback mismatch $\mathbf{K}(\mathcal{P}_U|\mathcal{P}_V)$ by a cruder measure: its approximation is a cumulant mismatch by equation 3.5.

3.1.1 Approximation to the Likelihood Contrast. The infomax-ML contrast $\phi^{\mathrm{ML}}(Y) = \mathbf{K}(\mathcal{P}_Y|\mathcal{P}_S)$ for ICA (see equation 2.4) is readily approximated by using expression 3.5. The assumption $\mathcal{P}_S$ on the distribution of S is now replaced by an assumption about the cumulants of S. This amounts to very little: all the cross-cumulants of S being 0 thanks to the assumption of independent sources, it is needed only to specify the autocumulants $\sigma^2(S_i)$ and $k(S_i)$. The cumulant approximation (see equation 3.5) to the infomax-ML contrast becomes:

$$\phi^{\mathrm{ML}}(Y) \approx \mathbf{K}_{24}(\mathcal{P}_Y|\mathcal{P}_S) = \frac{1}{4}\sum_{ij}\left(R^Y_{ij} - \sigma^2(S_i)\delta_{ij}\right)^2 + \frac{1}{48}\sum_{ijkl}\left(Q^Y_{ijkl} - k(S_i)\delta_{ijkl}\right)^2, \tag{3.6}$$

where the Kronecker symbol δ equals 1 with identical indices and 0 otherwise.

3.1.2 Approximation to the Mutual Information Contrast. The mutual information contrast $\phi^{\mathrm{MI}}(Y)$ was obtained by minimizing $\mathbf{K}(\mathcal{P}_Y|\mathcal{P}_S)$ over all the distributions $\mathcal{P}_S$ with independent components. In the cumulant approximation, this is trivially done: the free parameters for $\mathcal{P}_S$ are $\sigma^2(S_i)$ and $k(S_i)$. Each of these scalars enters in only one term of the sums in equation 3.6 so that the minimization is achieved for $\sigma^2(S_i) = R^Y_{ii}$ and $k(S_i) = Q^Y_{iiii}$. In other words, the construction of the best approximating distribution with

independent marginals $\bar{\mathcal{P}}_Y$, which appears in equation 2.5, boils down, in the cumulant approximation, to the estimation of the variance and kurtosis of each entry of Y. Fitting both $\sigma^2(S_i)$ and $k(S_i)$ to R^Y_{ii} and Q^Y_{iii}, respectively, has the effect of exactly cancelling the diagonal terms in equation 3.6, leaving only

$$\phi^{\mathrm{MI}}(Y) \approx \phi^{\mathrm{MI}}_{24}(Y) \overset{\text{def}}{=} \frac{1}{4}\sum_{ij\neq ii}\left(R^Y_{ij}\right)^2 + \frac{1}{48}\sum_{ijkl\neq iiii}\left(Q^Y_{ijkl}\right)^2, \tag{3.7}$$

which is our cumulant approximation to the mutual information contrast in equation 2.6. The first term is understood as the sum over all the pairs of distinct indices; the second term is a sum over all quadruples of indices that are not all identical. It contains only off-diagonal terms, that is, cross-cumulants. Since cross-cumulants of independent variables identically vanish, it is not surprising to see the mutual information approximated by a sum of squared cross-cumulants.

3.1.3 Approximation to the Orthogonal Likelihood Contrast. The cumulant approximation to the orthogonal likelihood is fairly simple. The orthogonal approach consists of first enforcing the whiteness of Y that is $R^Y_{ij} = \delta_{ij}$ or $R^Y = I$. In other words, it consists of normalizing the components by assuming that $\sigma^2(S_i) = 1$ and making sure the second-order mismatch is zero. This is equivalent to replacing the weight $\frac{1}{4}$ in equation 3.6 by an infinite weight, hence reducing the problem to the minimization (under the whiteness constraint) of the fourth-order mismatch, or the second (quadruple) sum in equation 3.6. Thus, the orthogonal likelihood contrast is approximated by

$$\phi^{\mathrm{OML}}_{24}(Y) \overset{\text{def}}{=} \frac{1}{48}\sum_{ijkl}\left(Q^Y_{ijkl} - k(S_i)\delta_{ijkl}\right)^2. \tag{3.8}$$

This contrast has an interesting alternate expression. Developing the squares gives

$$\phi^{\mathrm{OML}}_{24}(Y) = \frac{1}{48}\sum_{ijkl}(Q^Y_{ijkl})^2 + \frac{1}{48}\sum_{ijkl}k^2(S_i)\delta^2_{ijkl} - \frac{2}{48}\sum_{ijkl}k(S_i)\delta_{ijkl}Q^Y_{ijkl}.$$

The first sum above is constant under the whiteness constraint (this is readily checked using equation 3.13 for an orthonormal transform), and the second sum does not depend on Y; finally the last sum contains only diagonal nonzero terms. It follows that:

$$\begin{aligned}\phi^{\mathrm{OML}}_{24}(Y) &\overset{\mathrm{c}}{=} -\frac{1}{24}\sum_i k(S_i)Q^Y_{iiii}\\ &= -\frac{1}{24}\sum_i k(S_i)k(Y_i) \overset{\mathrm{c}}{=} -\frac{1}{24}\sum_i k(S_i)\mathrm{E}\bar{Y}_i^4,\end{aligned} \tag{3.9}$$

where $\cdot \overset{\mathrm{c}}{=} \cdot$ denotes an equality up to a constant. An interpretation of the second equality is that the contrast is minimized by maximizing the scalar product between the vector $[k(Y_1), \dots, k(Y_n)]$ of the kurtosis of the components and the corresponding vector of hypothesized kurtosis $[k(S_1), \dots, k(S_n)]$. The last equality stems from the definition in equation 3.2 of the kurtosis and the constancy of $\mathrm{E}\bar{Y}_i^2$ under the whiteness constraint. This last form is remarkable because it shows that for zero-mean observations, $\phi_{24}^{\mathrm{OML}}(Y) \overset{\mathrm{c}}{=} \mathrm{E}l(Y)$, where $l(Y) = -\frac{1}{24}\sum_i k(S_i)Y_i^4$, so the contrast is just the expectation of a simple function of Y. We can expect simple techniques for its maximization.

3.1.4 Approximation to the Minimum Marginal Entropy Contrast. Under the whiteness constraint, the first sum in the approximation, equation 3.7, is zero (this *is* the whiteness constraint) so that the approximation to mutual information $\phi^{\mathrm{MI}}(Y)$ reduces to the last term:

$$\phi^{\mathrm{ME}}(Y) \approx \phi_{24}^{\mathrm{ME}}(Y) \overset{\mathrm{def}}{=} \frac{1}{48} \sum_{ijkl \neq iiii} \left(Q_{ijkl}^{Y}\right)^2 \overset{\mathrm{c}}{=} -\frac{1}{48} \sum_{i} \left(Q_{iiii}^{Y}\right)^2. \tag{3.10}$$

Again, the last equality up to constant follows from the constancy of $\sum_{ijkl} (Q_{ijkl}^{Y})^2$ under the whiteness constraint. These approximations had already been obtained by Comon (1994) from an Edgeworth expansion. They say something simple: Edgeworth expansions suggest testing the independence between the entries of Y by summing up all the squared cross-cumulants.

In the course of this article, we will find two similar contrast functions. The JADE contrast,

$$\phi^{\mathrm{JADE}}(Y) \overset{\mathrm{def}}{=} \sum_{ijkl \neq iikl} \left(Q_{ijkl}^{Y}\right)^2, \tag{3.11}$$

also is a sum of squared cross-cumulants (the notation indicates a sum is over all the quadruples $(ijkl)$ of indices with $i \neq j$). Its interest is to be also a criterion of joint diagonality of cumulants matrices. The SHIBBS criterion,

$$\phi^{\mathrm{SH}}(Y) \overset{\mathrm{def}}{=} \sum_{ijkl \neq iikk} \left(Q_{ijkl}^{Y}\right)^2, \tag{3.12}$$

is also introduced in section 4.3 as governing a similar but less memory-demanding algorithm. It also involves only cross-cumulants: those with indices $(ijkl)$ such that $i \neq j$ or $k \neq l$.

3.2 Cumulants and Algebraic Structures. Previous sections reviewed the use of cumulants in designing contrast functions. Another thread of ideas using cumulants stems from the method of moments. Such an approach is called for by the multilinearity of the cumulants. Under a linear transform $Y = BX$, which also reads $Y_i = \sum_p b_{ip} X_p$, the cumulants of order 4 (for instance) transform as:

$$\mathrm{Cum}(Y_i, Y_j, Y_k, Y_l) = \sum_{pqrs} b_{ip} b_{jq} b_{kr} b_{ls} \mathrm{Cum}(X_p, X_q, X_r, X_s), \tag{3.13}$$

which can easily be exploited for our purposes since the ICA model is linear. Using this fact and the assumption of independence by which $\mathrm{Cum}(S_p, S_q, S_r, S_s) = k(S_p)\delta(p, q, r, s)$, we readily obtain the simple algebraic structure of the cumulants of $X = AS$ when S has independent entries,

$$\mathrm{Cum}(X_i, X_j, X_k, X_l) = \sum_{u=1}^{n} k(S_u) a_{iu} a_{ju} a_{ku} a_{lu}, \tag{3.14}$$

where a_{ij} denotes the (ij)th entry of matrix A. When estimates $\widehat{\mathrm{Cum}}(X_i, X_j, X_k, X_l)$ are available, one may try to solve equation 3.4 in the coefficients a_{ij} of A. This is tantamount to cumulant matching on the empirical cumulants of X. Because of the strong algebraic structure of equation 3.14, one may try to devise fourth-order factorizations akin to the familiar second-order singular value decomposition (SVD) or eigenvalue decomposition (EVD) (see Cardoso, 1992; Comon, 1997; De Lathauwer, De Moor, & Vandewalle, 1996). However, these approaches are generally not equivalent to the optimization of a contrast function, resulting in estimates that are generally not equivariant (Cardoso, 1995). This point is illustrated below; we introduce cumulant matrices whose simple structure offers straightforward identification techniques, but we stress, as one of their important drawbacks, their lack of equivariance. However, we conclude by showing how the algebraic point of view and the statistical (equivariant) point of view can be reconciled.

3.2.1 Cumulant Matrices. The algebraic nature of cumulants is tensorial (McCullagh, 1987), but since we will concern ourselves mainly with second- and fourth-order statistics, a matrix-based notation suffices for the purpose of our exposition; we only introduce the notion of cumulant matrix defined as follows. Given a random $n \times 1$ vector X and any $n \times n$ matrix M, we define the associated cumulant matrix $Q_X(M)$ as the $n \times n$ matrix defined component-wise by

$$[Q^X(M)]_{ij} \overset{\mathrm{def}}{=} \sum_{k,l=1}^{n} \mathrm{Cum}(X_i, X_j, X_k, X_l) M_{kl}. \tag{3.15}$$

If X is centered, the definition in equation 3.1 shows that

$$Q^X(M) = \mathrm{E}\{(X^\dagger M X)\, X X^\dagger\} - R^X \mathrm{tr}(M R^X) - R^X M R^X - R^X M^\dagger R^X, \tag{3.16}$$

where $\mathrm{tr}(\cdot)$ denotes the trace and R^X denotes the covariance matrix of X, that is, $[R^X]_{ij} = \mathrm{Cum}(X_i, X_j)$. Equation 3.16 could have been chosen as an index-free definition of cumulant matrices. It shows that a given cumulant matrix can be computed or estimated at a cost similar to the estimation cost of a covariance matrix; there is no need to compute the whole set of fourth-order cumulants to obtain the value of $Q^X(M)$ for a particular value of M. Actually, estimating a particular cumulant matrix is one way of collecting part of the fourth-order information in X; collecting the whole fourth-order information requires the estimation of $O(n^4)$ fourth-order cumulants.

The structure of a cumulant matrix $Q^X(M)$ in the ICA model is easily deduced from equation 3.14:

$$\begin{aligned} &Q^X(M) = A\Delta(M)A^\dagger \\ &\Delta(M) = \mathrm{Diag}\left(\mathrm{k(S_1)}\,\mathrm{a}_1^\dagger \mathrm{Ma}_1, \ldots, \mathrm{k(S_n)}\,\mathrm{a}_\mathrm{n}^\dagger \mathrm{Ma}_\mathrm{n}\right), \end{aligned} \tag{3.17}$$

where $\mathbf{a_i}$ denotes the ith column of A, that is, $A = [\mathbf{a_1}, \ldots, \mathbf{a_n}]$. In this factorization, the (generally unknown) kurtosis enter only in the diagonal matrix $\Delta(M)$, a fact implicitly exploited by the algebraic techniques described below.

3.3 Blind Identification Using Algebraic Structures.

In section 3.1, contrast functions were derived from the ML principle assuming the model $X = AS$. In this section, we proceed similarly: we consider cumulant-based blind identification of A assuming $X = AS$ from which the structures 3.14 and 3.17 result.

Recall that the orthogonal approach can be implemented by first sphering explicitly vector X. Let W be a whitening, and denote $Z \overset{\text{def}}{=} WX$ the sphered vector. Without loss of generality, the model can be normalized by assuming that the entries of S have unit variance so that S is spatially white. Since $Z = WX = WAS$ is also white by construction, the matrix $U \overset{\text{def}}{=} WA$ must be orthonormal: $UU^\dagger = I$. Therefore sphering yields the model $Z = US$ with U orthonormal. Of course, this is still a model of independent components so that, similar to equation 3.17, we have for any matrix M the structure of the corresponding cumulant matrix of Z,

$$\begin{aligned} &Q^Z(M) = U\tilde{\Delta}(M)U^\dagger \\ &\tilde{\Delta}(M) = \mathrm{Diag}\left(\mathrm{k(S_1)}\,\mathrm{u}_1^\dagger \mathrm{Mu}_1, \ldots, \mathrm{k(S_n)}\,\mathrm{u}_\mathrm{n}^\dagger \mathrm{Mu}_\mathrm{n}\right), \end{aligned} \tag{3.18}$$

where $\mathbf{u_i}$ denotes the ith column of U. practical orthogonal statistic-based technique, one would first estimate a whitening matrix $\widehat{W}$, estimate some cumulants of $Z = \widehat{W}X$, compute an orthonormal estimate $\widehat{U}$ of U using these cumulants, and finally obtain an estimate $\widehat{A}$ of A as $\widehat{A} = \widehat{W}^{-1}\widehat{U}$ or obtain a separating matrix as $B = \widehat{U}^{-1}\widehat{W} = \widehat{U}^\dagger\widehat{W}$.

3.3.1 Nonequivariant Blind Identification Procedures.

We first present two blind identification procedures that exploit in a straightforward manner the structure 3.17; we explain why, in spite of attractive computational simplicity, they are not well behaved (not equivariant) and how they can be fixed for equivariance.

The first idea is not based on an orthogonal approach. Let M_1 and M_2 be two arbitrary $n \times n$ matrices, and define $Q_1 \overset{\text{def}}{=} Q^X(M_1)$ and $Q_2 \overset{\text{def}}{=} Q^X(M_2)$. According to equation 3.17, if $X = AS$ we have $Q_1 = A\Delta_1 A^\dagger$ and $Q_2 = A\Delta_2 A^\dagger$ with Δ_1 and Δ_2 two diagonal matrices. Thus, $G \overset{\text{def}}{=} Q_1 Q_2^{-1} = (A\Delta_1 A^\dagger)(A\Delta_2 A^\dagger)^{-1} = A\Delta A^{-1}$, where Δ is the diagonal matrix $\Delta_1\Delta_2^{-1}$. It follows that $GA = A\Delta$, meaning that the columns of A are the eigenvectors of G (possibly up to scale factors).

An extremely simple algorithm for blind identification of A follows: Select two arbitrary matrices M_1 and M_2; compute sample estimates $\hat{Q}_1$ and $\hat{Q}_2$ using equation 3.16; find the columns of A as the eigenvectors of $\hat{Q}_1\hat{Q}_2^{-1}$. There is at least one problem with this idea: we have assumed invertible matrices throughout the derivation, and this may lead to instability. However, this specific problem may be fixed by sphering, as examined next.

Consider now the orthogonal approach as outlined above. Let M be some arbitrary matrix M, and note that equation 3.18 is an eigendecomposition: the columns of U are the eigenvectors of $Q^Z(M)$, which are orthonormal indeed because $Q^Z(M)$ is symmetric. Thus, in the orthogonal approach, another immediate algorithm for blind identification is to estimate U as an (orthonormal) diagonalizer of an estimate of $Q^Z(M)$. Thanks to sphering, problems associated with matrix inversion disappear, but a deeper problem associated with these simple algebraic ideas remains and must be addressed. Recall that the eigenvectors are uniquely determined[2] if and only if the eigenvalues are all distinct. Therefore, we need to make sure that the eigenvalues of $Q^Z(M)$ are all distinct in order to preserve blind identifiability based on $Q^Z(M)$. According to equation 3.18, these eigenvalues depend on the (sphered) system, which is unknown. Thus, it is not possible to determine a priori if a given matrix M corresponds to distinct eigenvalues of $Q^Z(M)$. Of course, if M is randomly chosen, then the eigenvalues are distinct with probability 1, but we need more than this in practice because the algorithms use only sample estimates of the cumulant matrices. A small error in the sample estimate of $Q^Z(M)$ can induce a large deviation of the eigenvectors if the eigenvalues are not well enough separated. Again, this is impossible to guarantee a priori because an appropriate selection of M requires prior knowledge about the unknown mixture.

In summary, the diagonalization of a single cumulant matrix is computa-

[2] In fact, determined only up to permutations and signs that do not matter in an ICA context.

tionally attractive and can be proved to be almost surely consistent, but it is not satisfactory because the nondegeneracy of the spectrum cannot be controlled. As a result, the estimation accuracy from a finite number of samples depends on the unknown system and is therefore unpredictable in practice; this lack of equivariance is hardly acceptable. One may also criticize these approaches on the ground that they rely on only a small part of the fourth-order information (summarized in an $n \times n$ cumulant matrix) rather than trying to exploit more cumulants (there are $O(n^4)$ fourth-order independent cumulant statistics). We examine next how these two problems can be alleviated by jointly processing several cumulant matrices.

3.3.2 Recovering Equivariance. Let $\mathcal{M} = \{M_1, \ldots, M_P\}$ be a set of P matrices of size $n \times n$ and denote $Q_i \overset{\text{def}}{=} Q^Z(M_i)$ for $1 \le i \le P$ the associated cumulant matrices for the sphered data $Z = US$. Again, as above, for all i we have $Q_i = U\Delta_i U^\dagger$ with Δ_i a diagonal matrix given by equation 3.18. As a measure of nondiagonality of a matrix F, define Off(F) as the sum of the squares of the nondiagonal elements:

$$\mathrm{Off(F)} \overset{\text{def}}{=} \sum_{\mathrm{i \neq j}} \left(\mathrm{f_{ij}}\right)^2 . \tag{3.19}$$

We have in particular $\mathrm{Off(U^\dagger Q_i U)} = \mathrm{Off(\Delta_i)} = 0$ since $Q_i = U\Delta_i U^\dagger$ and $U^\dagger U = I$. For any matrix set $\mathcal{M}$ and any orthonormal matrix V, we define the following nonnegative joint diagonality criterion,

$$\mathcal{D}_{\mathcal{M}}(V) \overset{\text{def}}{=} \sum_{M_i \in \mathcal{M}} \mathrm{Off(V^\dagger Q^Z(M_i) V)}, \tag{3.20}$$

which measures how close to diagonality an orthonormal matrix V can simultaneously bring the cumulants matrices generated by $\mathcal{M}$.

To each matrix set $\mathcal{M}$ is associated a blind identification algorithm as follows: **(1)** fin sphering matrix W to whiten in the data X into $Z = WX$; **(2)** estimate the cumulant matrices $Q^Z(M)$ for all $M \in \mathcal{M}$ by a sample version of equation 3.16; **(3)** minimize the joint diagonality criterion, equation 3.20, that is, make the cumulant matrices as diagonal as possible by an orthonormal transform V; **(4)** estimate A as $A = VW^{-1}$ or its inverse as $B = V^\dagger W$ or the component vector as $Y = V^\dagger Z = V^\dagger WX$.

Such an approach seems to be able to alleviate the drawbacks mentioned above. Finding the orthonormal transform as the minimizer of a set of cumulant matrices goes in the right direction because it involves a larger number of fourth-order statistics and because it decreases the likelihood of degenerate spectra. This argument can be made rigorous by considering a maximal set of cumulant matrices. By definition, this is a set obtained whenever $\mathcal{M}$ is an orthonormal basis for the linear space of $n \times n$ matrices. Such a basis contains n^2 matrices so that the corresponding cumulant matrices total $n^2 \times n^2 = n^4$ entries, that is, as many as the whole fourth-order cumulant set. For any such maximal set (Cardoso & Souloumiac, 1993):

$$\mathcal{D}_{\mathcal{M}}(V) = \phi^{\mathrm{JADE}}(Y) \quad \text{with} \quad Y = V^\dagger Z, \tag{3.21}$$

where $\phi^{\mathrm{JADE}}(Y)$ is the contrast function defined at equation 3.11. The joint diagonalization of a maximal set guarantees blind identifiability of A if $k(S_i) = 0$ for at most one entry S_i of S (Cardoso & Souloumiac, 1993). This is a necessary condition for any algorithm using only second- and fourth-order statistics (Comon, 1994).

A key point is made by relationship 3.21. We managed to turn an algebraic property (diagonality) of the cumulants of the (sphered) observations into a contrast function—a functional of the distribution of the output $Y = V^\dagger Z$. This fact guarantees that the resulting estimates are equivariant (Cardoso, 1995).

The price to pay with this technique for reconciling the algebraic approach with the naturally equivariant contrast-based approach is twofold: it entails the computation of a large (actually, maximal) set of cumulant matrices and the joint diagonalization of $P = n^2$ matrices, which is at least as costly as P times the diagonalization of a single matrix. However, the overall computational burden may be similar (see examples in section 5) to the cost of adaptive algorithms. This is because the cumulant matrices need to be estimated once for a given data set and because it exists as a reasonably efficient joint diagonalization algorithm (see section 4) that is *not* based on gradient-style optimization; it thus preserves the possibility of exploiting the underlying algebraic nature of the contrast function, equation 3.11. Several tricks for increasing efficiency are also discussed in section 4.

4 Jacobi Algorithms

This section describes algorithms for ICA sharing a common feature: a Jacobi optimization of an orthogonal contrast function as opposed to optimization by gradient-like algorithms. The principle of Jacobi optimization is applied to a data-based algorithm, a statistic-based algorithm, and a mixed approach.

The Jacobi method is an iterative technique of optimization over the set of orthonormal matrices. The orthonormal transform is obtained as a sequence of plane rotations. Each plane rotation is a rotation applied to a pair of coordinates (hence the name: the rotation operates in a two-dimensional plane). If Y is an $n \times 1$ vector, the (i, j)th plane rotation by an angle θ_{ij} changes the coordinates i and j of Y according to

$$\begin{bmatrix} Y_i \\ Y_j \end{bmatrix} \leftarrow \begin{bmatrix} \cos(\theta_{ij}) & \sin(\theta_{ij}) \\ -\sin(\theta_{ij}) & \cos(\theta_{ij}) \end{bmatrix} \begin{bmatrix} Y_i \\ Y_j \end{bmatrix}, \tag{4.1}$$

while leaving the other coordinates unchanged. A sweep is one pass through all the $n(n-1)/2$ possible pairs of distinct indices. This idea is classic in numerical analysis (Golub & Van Loan, 1989); it can be considered in a wider context for the optimization of any function of an orthonormal matrix. Comon introduced the Jacobi technique for ICA (see Comon, 1994 for a data-based algorithm and an earlier reference in it for the Jacobi update of high-order cumulant tensors). Such a data-based Jacobi algorithm for ICA works through a sequence of Jacobi sweeps on the sphered data until a given orthogonal contrast $\phi(Y)$ is optimized. This can be summarized as:

1. *Initialization.* Compute a whitening matrix W and set $Y = WX$.
2. *One sweep.* For all $n(n-1)/2$ pairs, that is for $1 \le i < j \le n$, do:
   a. Compute the Givens angle θ_{ij}, optimizing $\phi(Y)$ when the pair (Y_i, Y_j) is rotated.
   b. If $\theta_{ij} < \theta_{\min}$, do rotate the pair (Y_i, Y_j) according to equation 4.1.
3. If no pair has been rotated in previous sweep, end. Otherwise go to 2 for another sweep.

Thus, the Jacobi approach considers a sequence of two-dimensional ICA problems. Of course, the updating step 2b on a pair (i, j) partially undoes the effect of previous optimizations on pairs containing either i or j. For this reason, it is necessary to go through several sweeps before optimization is completed. However, Jacobi algorithms are often very efficient and converge in a small number of sweeps (see the examples in section 5), and a key point is that each plane rotation depends on a single parameter, the Givens angle θ_{ij}, reducing the optimization subproblem at each step to a one-dimensional optimization problem.

An important benefit of basing ICA on fourth-order contrasts becomes apparent: because fourth-order contrasts are polynomial in the parameters, the Givens angles can often be found in close form.

In the above scheme, $\theta_{\min}$ is a small angle, which controls the accuracy of the optimization. In numerical analysis, it is determined according to machine precision. For a statistical problem as ICA, $\theta_{\min}$ should be selected in such a way that rotations by a smaller angle are not statistically significant. In our experiments, we take $\theta_{\min}$ to scale as $1/\sqrt{T}$, typically: $\theta_{\min} = \frac{10^{-2}}{\sqrt{T}}$. This scaling can be related to the existence of a performance bound in the orthogonal approach to ICA(Cardoso, 1994). This value does not seem to be critical, however, because we have found Jacobi algorithms to be very fast at finishing.

In the remainder of this section, we describe three possible implementations of these ideas. Each one corresponds to a different type of contrast function and to different options about updating. Section 4.1 describes a data-based algorithm optimizing $\phi^{\mathrm{OML}}(Y)$; section 4.2 describes a statistic-based algorithm optimizing $\phi^{\mathrm{JADE}}(Y)$; section 4.3 presents a mixed approach optimizing $\phi^{\mathrm{SH}}(Y)$; finally, section 4.4 discusses the relationships between these contrast functions.

4.1 A Data-Based Jacobi Algorithm: MaxKurt. We start by a Jacobi technique for optimizing the approximation, equation 3.9, to the orthogonal likelihood. For the sake of exposition, we consider a simplified version of $\phi^{\mathrm{OML}}(Y)$ obtained by setting $k(S_1) = k(S_2) = \ldots = k(S_n) = k$, in which case the minimization of contrast function, equation 3.9, is equivalent to the minimization of

$$\phi^{\mathrm{MK}}(Y) \overset{\mathrm{def}}{=} -k \sum_i Q^Y_{iiii}. \tag{4.2}$$

This criterion is also studied by Moreau and Macchi (1996), who propose a two-stage adaptive procedure for its optimization; it also serves as a starting point for introducing the one-stage adaptive algorithm of Cardoso and Laheld (1996).

Denote $G_{ij}(\theta)$ the plane rotation matrix that rotates the pair (i, j) by an angle θ as in step 2b above. Then simple trigonometry yields:

$$\phi^{\mathrm{MK}}(G_{ij}(\theta)Y) = \mu_{ij} - k\lambda_{ij}\cos(4(\theta - \Omega_{ij})), \tag{4.3}$$

where μ_{ij} does not depend on θ and λ_{ij} is nonnegative. The principal determination of angle Ω_{ij} is characterized by

$$\Omega_{ij} = \frac{1}{4}\arctan\left(4Q^Y_{iiij} - 4Q^Y_{ijjj}\ ,\ \ Q^Y_{iiii} + Q^Y_{jjjj} - 6Q^Y_{iijj}\right), \tag{4.4}$$

where $\arctan(y, x)$ denotes the angle $\alpha \in (-\pi, \pi]$ such that $\cos(\alpha) = \frac{x}{\sqrt{x^2+y^2}}$ and $\sin(\alpha) = \frac{y}{\sqrt{x^2+y^2}}$. If Y is a zero-mean sphered vector, expression 4.3 further simplifies to

$$\Omega_{ij} = \frac{1}{4}\arctan\left(4\mathrm{E}\left(Y_i^3 Y_j - Y_i Y_j^3\right)\ ,\ \ \mathrm{E}\left((Y_i^2 - Y_j^2)^2 - 4Y_i^2 Y_j^2\right)\right). \tag{4.5}$$

The computations are given in the appendix. It is now immediate to minimize $\phi^{\mathrm{MK}}(Y)$ for each pair of components and for either choice of the sign of k. If one looks for components with positive kurtosis (often called supergaussian), the minimization of $\phi^{\mathrm{MK}}(Y)$ is identical to the maximization of the sum of the kurtosis of the components since we have $k > 0$ in this case. The Givens angle simply is $\theta = \Omega_{ij}$ since this choice makes the cosine in equation 4.2 equal to its maximum value.

We refer to the Jacobi algorithm outlined above as MaxKurt. A Matlab implementation is listed in the appendix, whose simplicity is consistent with the data-based approach. Note, however, that it is also possible to use the same computations in a statistic-based algorithm. Rather than rotating the data themselves at each step by equation 4.1, one instead updates the set of all fourth-order cumulants according to the transformation law, equation 3.13, with the Givens angle for each pair still given by equation 4.3. In this case, the memory requirement is $O(n^4)$ for storing all the cumulants as opposed to nT for storing the data set.

The case $k < 0$ where, looking for light-tailed components, one should minimize the sum of the kurtosis is similar. This approach could be extended to kurtosis of mixed signs but the contrast function then has less symmetry. This is not included in this article.

4.1.1 Stability. What is the effect of the approximation of equal kurtosis made to derive the simple contrast $\phi^{\mathrm{MK}}(Y)$? When $X = AS$ with S of independent components, we can at least use the stability result of Cardoso and Laheld (1996), which applies directly to this contrast. Define the normalized kurtosis as $\kappa_i = \sigma_i^{-4} k(S_i)$. Then $B = A^{-1}$ is a stable point of the algorithm with $k > 0$ if $\kappa_i + \kappa_j > 0$ for all pairs $1 \leq i < j \leq n$. The same condition also holds with all signs reversed for components with negative kurtosis.

4.2 Statistic-Based Algorithm: JADE. This section outlines the JADE algorithm (Cardoso & Souloumiac, 1993), which is specifically a statistic-based technique. We do not need to go into much detail because the general technique follows directly from the considerations of section 3.3. The JADE algorithm can be summarized as:

1. *Initialization*. Estimate a whitening matrix $\hat{W}$ and set $Z = \hat{W}X$.
2. *Form statistics*. Estimate a maximal set $\{\hat{Q}_i^Z\}$ of cumulant matrices.
3. *Optimize an orthogonal contrast*. Find the rotation matrix $\hat{V}$ such that the cumulant matrices are as diagonal as possible, that is, solve $\hat{V} = \arg\min \sum_i \mathrm{Off}(\mathrm{V}^\dagger \hat{\mathrm{Q}}_{\mathrm{i}}^{\mathrm{Z}} \mathrm{V})$.
4. *Separate*. Estimate A as $\hat{A} = \hat{V}\hat{W}^{-1}$ and/or estimate the components as $\hat{S} = \hat{A}^{-1}X = \hat{V}^\dagger Z$.

This is a Jacobi algorithm because the joint diagonalizer at step 3 is found by a Jacobi technique. However, the plane rotations are applied not to the data (which are summarized in the cumulant matrices) but to the cumulant matrices themselves; the algorithm updates not data but matrix-valued statistics of the data. As with MaxKurt, the Givens angle at each step can be computed in closed form even in the case of possibly complex matrices (Cardoso & Souloumiac, 1993). The explicit expression for the Givens angles is not particularly enlightening and is not reported here. (The interested reader is referred to Cardoso & Souloumiac, 1993, and may request a Matlab implementation from the author.)

A key issue is the selection of the cumulant matrices to be involved in the estimation. As explained in section 3.2, the joint diagonalization criterion $\sum_i \mathrm{Off}(\mathrm{V}^\dagger \hat{\mathrm{Q}}_{\mathrm{i}}^{\mathrm{Z}} \mathrm{V})$ is made identical to the contrast function, equation 3.11, by using a maximal set of cumulant matrices. This is a bit surprising but very fortunate. We do not know of any other way for a priori selecting cumulant matrices that would offer such a property (but see the next section). In any case, it guarantees equivariant estimates because the algorithm, although operating on statistics of the sphered data, also optimizes implicitly a function of $Y = V^\dagger Z$ only.

Before proceeding, we note that true cumulant matrices can be exactly jointly diagonalized when the model holds, but this is no longer the case when we process real data. First, only sample statistics are available; second, the model $X = AS$ with independent entries in S cannot be expected to hold accurately in general. This is another reason that it is important to select cumulant matrices such that $\sum_i \mathrm{Off}(\mathrm{V}^\dagger \hat{\mathrm{Q}}_{\mathrm{i}}^{\mathrm{Z}} \mathrm{V})$ is contrast function. In this case, the impossibility of an exact joint diagonalization corresponds to the impossibility of finding $Y = BX$ with independent entries. Making a maximal set of cumulant matrices as diagonal as possible coincides with making the entries of Y as independent as possible as measured by (the sample version of) criterion 3.11.

There are several options for estimating a maximal set of cumulant matrices. Recall that such a set is defined as $\{Q^Z(M_i) | i = 1, n^2\}$ where $\{M_i | i = 1, n^2\}$ is any basis for the n^2-dimensional linear space of $n \times n$ matrices. A canonical basis for this space is $\{e_p e_q^\dagger | 1 \leq p, q \leq n\}$, where e_p is a column vector with a 1 in pth position and 0's elsewhere. It is readily checked that

$$[Q^Z(e_p e_q^\dagger)]_{ij} = \mathrm{Cum}(Z_i, Z_j, Z_p, Z_q). \tag{4.6}$$

In other words, the entries of the cumulant matrices for the canonical basis are just the cumulants of Z. A better choice is to consider a symmetric/skew-symmetric basis. Denote M^{pq} an $n \times n$ matrix defined as follows: $M^{pq} = e_p e_p^\dagger$ if $p = q$, $M^{pq} = 2^{-1/2}(e_p e_q^\dagger + e_q e_p^\dagger)$ if $p < q$ and $M^{pq} = 2^{-1/2}(e_p e_q^\dagger - e_q e_p^\dagger)$ if $p > q$. This is an orthonormal basis of $\mathbf{R}^{n \times n}$. We note that because of the symmetries of the cumulants $Q^Z(e_p e_q^\dagger) = Q^Z(e_q e_p^\dagger)$ so that $Q^Z(M^{pq}) = 2^{-1/2} Q^Z(e_p e_q^\dagger)$ if $p < q$ and $Q^Z(M^{pq}) = 0$ if $p > q$. It follows that the cumulant matrices $Q^Z(M^{pq})$ for $p > q$ do not even need to be computed. Being identically zero, they do not enter in the joint diagonalization criterion. It is therefore sufficient to estimate and to diagonalize $n + n(n-1)/2$ (symmetric) cumulant matrices.

There is another idea to reduce the size of the statistics needed to represent exhaustively the fourth-order information. It is, however, applicable only when the model $X = AS$ holds. In this case, the cumulant matrices do have the structure shown at equation 3.18, and their sample estimates are close to it for large enough T. Then the linear mapping $M \rightarrow Q^Z(M)$ has rank n

(more precisely, its rank is equal to the number of components with nonzero kurtosis) because there are n linear degrees of freedom for matrices in the form $U\Delta U^\dagger$, namely, the n diagonal entries of Δ. From this fact and from the symmetries of the cumulants, it follows that it exists n eigenmatrices $E_1, \ldots, E_n$, which are orthonormal, and satisfies $Q^Z(E_i) = \mu_i E_i$ where the scalar μ_i is the corresponding eigenvector. These matrices $E_1, \ldots, E_n$ span the range of the mapping $M \to Q^Z(M)$, and any matrix M orthogonal to them is in the kernel, that is, $Q^Z(M) = 0$. This shows that all the information contained in Q^Z can be summarized by the n eigenmatrices associated with the n n eigenvalues. By inserting $M = u_i u_i^\dagger$ in the expressions 3.18 and using the orthonormality of the columns of U (that is, $u_i^\dagger u_j = \delta_{ij}$), it is readily checked that a set of eigenmatrices is $\{E_i = u_i u_i^\dagger\}$.

The JADE algorithm was originally introduced as performing ICA by a joint approximate diagonalization of eigenmatrices in Cardoso and Souloumiac (1993), where we advocated the joint diagonalization of only the n most significant eigenmatrices of Q^Z as a device to reduce the computational load (even though the eigenmatrices are obtained at the extra cost of the eigendecomposition of an $n^2 \times n^2$ array containing all the fourth-order cumulants). The number of statistics is reduced from n^4 cumulants or $n(n+1)/2$ symmetric cumulant matrices of size $n \times n$ to a set of n eigenmatrices of size $n \times n$. Such a reduction is achieved at no statistical loss (at least for large T) only when the model holds. Therefore, we do not recommend reduction to eigenmatrices when processing data sets for which it is not clear a priori whether the model $X = AS$ actually holds to good accuracy. We still refer to JADE as ocess of jointly diagonalizing a maximal set of cumulant matrices, even when it is not further reduced to the n most significant eigenmatrices. It should also be pointed out that the device of truncating the full cumulant set by reduction to the most significant matrices is expected to destroy the equivariance property when the model does not hold. The next section shows how these problems can be overcome in a technique borrowing from both the data-based approach and the statistic-based approach.

4.3 A Mixed Approach: SHIBBS.

In the JADE algorithm, a maximal set of cumulant matrices is computed as a way to ensure equivariance from the joint diagonalization of a fixed set of cumulant matrices. As a benefit, cumulants are computed only once in a single pass through the data set, and the Jacobi updates are performed on these statistics rather than on the whole data set. This is a good thing for data sets with a large number T of samples. On the other hand, estimating a maximal set requires $O(n^4 T)$ operations, and its storage requires $O(n^4)$ memory positions. These figures can become prohibitive when looking for a large number of components. In contrast, gradient-based techniques have to store and update nT samples. This section describes a technique standing between the two extreme positions represented by the all-statistic approach and the all-data approach.

Recall that an algorithm is equivariant as soon as its operation can be expressed only in terms of the extracted components Y (Cardoso, 1995). This suggests the following technique:

1. *Initialization*. Select a fixed set $\mathcal{M} = \{M_1, \ldots, M_P\}$ of $n \times n$ matrices. Estimate a whitening matrix $\hat{W}$ and set $Y = \hat{W}X$.

2. *Estimate a rotation*. Estimate the set $\{\hat{Q}^Y(M_p) | 1 \le p \le P\}$ of P cumulant matrices and find a joint diagonalizer V of it.

3. *Update*. If V is close enough to the identity transform, stop. Otherwise, rotate the data: $Y \leftarrow V^\dagger Y$ and go to 2 .

Such an algorithm is equivariant thanks to the reestimation of the cumulants of Y after updating. It is in some sense data based since the updating in step 3 is on the data themselves. However, the rotation matrix to be applied to the data is computed in step 2 as in a statistic-based procedure.

What would be a good choice for the set $\mathcal{M}$? The set of n matrices $\mathcal{M} = \{e_1 e_1^\dagger, \ldots, e_n e_n^\dagger\}$ seems a natural choice: it is an order of magnitude smaller than the maximal set, which contains $O(n^2)$ matrices. The kth cumulant matrix in such a set is $Q^Y(e_k e_k^\dagger)$, and its (i, j)th entry is $\mathrm{Cum}(Y_i, Y_j, Y_k, Y_k)$, which is just an $n \times n$ square block of cumulants of Y. We call the set of n cumulant matrices obtained in this way when k is shifted from 1 to n the set of SHIfted Blocks for Blind Separation (SHIBBS), and we use the same name for the ICA algorithm that determines the rotation V by an iterative joint diagonalization of the SHIBBS set.

Strikingly enough, the small SHIBBS set guarantees a performance identical to JADE *when the model holds* for the following reason. Consider the final step of the algorithm where Y is close to S if it holds that $X = AS$ with S of independent components. Then the cumulant matrices $Q^Y(e_p e_q^\dagger)$ are zero for $p \neq q$ because all the cross-cumulants of Y are zero. Therefore, the only nonzero cumulant matrices used in the maximal set of JADE are those corresponding to $e_p = e_q$, *i.e.* p those included in SHIBBS. Thus the SHIBBS set actually tends to the set of “significant eigen-matrices” exhibited in the previous section. In this sense, SHIBBS implements the original program of JADE—the joint diagonalization of the significant eigen-matrices—but it does so without going through the estimation of the whole cumulant set and through the computation of its eigen-matrices.

Does the SHIBBS algorithm correspond to the optimization of a contrast function? We cannot resort to the equivalence of JADE and SHIBBS because it is established only when the model holds and we are looking for a statement independent of this later fact. Examination of the joint diagonality criterion for the SHIBBS set suggests that the SHIBBS technique solves the problem of optimizing the contrast function $\phi^{\mathrm{SH}}(Y)$ defined in equation 3.12. As a matter of fact, the condition for a given Y to be a fixed point of the SHIBBS algorithm is that for any pair $1 \le i < j \le n$:

$$\sum_k \mathrm{Cum}(Y_i, Y_j, Y_k, Y_k)\,\big(\mathrm{Cum}(Y_i, Y_i, Y_k, Y_k) - \mathrm{Cum}(Y_j, Y_j, Y_k, Y_k)\big) = 0, \tag{4.7}$$

and we can prove that this is also the stationarity condition of $\phi^{\mathrm{SH}}(Y)$. We do not include the proofs of these statements, which are purely technical.

4.4 Comparing Fourth-Order Orthogonal Contrasts. We have considered two approximations, $\phi^{\mathrm{JADE}}(Y)$ and $\phi^{\mathrm{SH}}(Y)$, to the minimum marginal entropy/mutual information contrast $\phi^{\mathrm{MI}}(I)$, which are based on fourth-order cumulants and can be optimized by Jacobi technique. The approximation $\phi_{24}^{\mathrm{ME}}(Y)$ proposed by Comon also belongs to this category. One may wonder about the relative statistical merits of these three approximations. The contrast $\phi_{24}^{\mathrm{ME}}(Y)$ stems from an Edgeworth expansion for approximating $\phi^{\mathrm{ME}}(Y)$, which in turn has been shown to derive from the ML principle (see section 2). Since ML estimation offers (asymptotic) optimality properties, one may be tempted to conclude to the superiority of $\phi_{24}^{\mathrm{ME}}(Y)$. However, this is not the case, as discussed now.

First, when the ICA model holds, it can be shown that even though $\phi_{24}^{\mathrm{ME}}(Y)$ and $\phi^{\mathrm{JADE}}(Y)$ are different criteria, they have the same asymptotic performance when applied to sample statistics (Souloumiac & Cardoso, 1991). This is also true of $\phi^{\mathrm{SH}}(Y)$ since we have seen that it is equivalent to JADE in this case (a more rigorous proof is possible, based on equation 4.7, but is not included).

Second, when the ICA model does not hold, the notion of identification accuracy does not make sense anymore, but one would certainly favor an orthogonal contrast reaching its minimum at a point as close as possible to the point where the "true" mutual information $\phi^{\mathrm{ME}}(Y)$ is minimized. However, it seems difficult to find a simple contrast (such as those considered here) that would be a good approximation to $\phi^{\mathrm{ME}}(Y)$ for any wide class of distributions of X. Note that the ML argument in favor of $\phi_{24}^{\mathrm{ME}}(Y)$ is based on an Edgeworth expansion that is valid for "almost gaussian" distributions—those distributions that make ICA very difficult and of dubious significance: In practice, ICA should be restricted to data sets where the components show a significant amount of nongaussianity, in which case the Edgeworth expansions cannot be expected to be accurate.

There is another way than Edgeworth expansion for arriving at $\phi_{24}^{\mathrm{ME}}(Y)$. Consider cumulant matching: the matching of the cumulants of Y to the corresponding cumulants of a hypothetical vector S with independent components. The orthogonal contrast functions $\phi^{\mathrm{JADE}}(Y)$, $\phi^{\mathrm{SH}}(Y)$, and $\phi_{24}^{\mathrm{ME}}(Y)$ can be seen as matching criteria because they penalize the deviation of the cross-cumulants of Y from zero (which is the value of cross-cumulants of a vector S with independent components, indeed), and they do so under the constraint that Y is white—that is, by enforcing an exact match of the second-order cumulants of Y.

It is possible to devise an asymptotically optimal matching criterion by taking into account the variability of the sample estimates of the cumulants. Such a computation is reported in Cardoso et al. (1996) for the matching of all second- and fourth-order cumulants of complex-valued signals, but a similar computation is possible for real-valued problems. It shows that the optimal weighting of the cross-cumulants depends on the distributions of the components so that the "flat weighting" of all the cross-cumulants, as in equation 3.10, is not the best one in general. However, in the limit of "almost gaussian" signals, the optimal weights tend to values corresponding precisely to the contrast $\mathbf{K}_{24}(Y|S)$ defined in equation 3.5. This is not unexpected and confirms that the crude cumulant expansion used in deriving equation 3.5 is sensible, though not optimal for significantly nongaussian components.

It seems from the definitions of $\phi^{\mathrm{JADE}}(Y)$, $\phi^{\mathrm{SH}}(Y)$, and $\phi_{24}^{\mathrm{ME}}(Y)$ that these different contrasts involve different types of cumulants. This is, however, an illusion because the compact definitions given above do not take into account the symmetries of the cumulants: the same cross-cumulant may be counted several times in each of these contrasts. For instance, the definition of JADE excludes the cross-cumulant $\mathrm{Cum}(Y_1, Y_1, Y_2, Y_3)$ but includes the cross-cumulant $\mathrm{Cum}(Y_1, Y_2, Y_1, Y_3)$, which is identical. Thus, in order to determine if any bit of fourth-order information is ignored by any particular contrast, a nonredundant description should be given. All the possible cross-cumulants come in four different patterns of indices: (*ijkl*), (*iikl*), (*iijj*), and (*ijjj*). Nonredundant expressions in terms of these patterns are in the form:

$$\phi[Y] = C_a \sum_{i<j<k<l} e_{ijkl} + C_b \sum_{i<k<l} (e_{iikl} + e_{kkil} + e_{llik}) + C_c \sum_{i<j} e_{iijj} + C_d \sum_{i<j} \left(e_{ijjj} + e_{jjji}\right),$$

where $e_{ijkl} \overset{\mathrm{def}}{=} \mathrm{Cum}(Y_i, Y_j, Y_k, Y_l)^2$ and the C_i's are numerical constants. It remains to count how many times a unique cumulant appears in the redundant definitions of the three approximations to mutual information considered so far. We give only the result of this uninspiring task in Table 1, which shows that *all* the cross-cumulants are actually included in the three contrasts, which therefore differ only by the different scalar weights given to each particular type. It means that the three contrasts essentially do the same thing. In particular, when the number n of components is large enough, the number of cross-cumulants of type [*ijkl*] (all indices distinct) grows as $O(n^4)$, while the number of other types grows as $O(n^3)$ at most. Therefore, the [*ijkl*] type outnumbers all the other types for large n: one may conjecture the equivalence of the three contrasts in this limit. Unfortunately, it seems

Table 1: Number of Times a Cross-Cumulant of a Given Type Appears in a Given Contrast.

Constants	C_a	C_b	C_c	C_d
Pattern	*ijkl*	*iikl*	*iijj*	*ijjj*
Comon ICA	24	12	6	4
JADE	24	10	4	2
SHIBBS	24	12	4	4

difficult to draw more conclusions. For instance, we have mentioned the asymptotic equivalence between Comon's contrast and the JADE contrast for any n, but it does not reveal itself directly in the weight table.

5 A Comparison on Biomedical Data

The performance of the algorithms presented above is illustrated using the averaged event-related potential (ERP) data recorded and processed by Makeig and coworkers. A detailed account of their analysis is in Makeig, Bell, Jung, and Sejnowski (1997). For our comparison, we use the data set and the "logistic ICA" algorithm provided with version 3.1 of Makeig's ICA toolbox.[3] The data set contains 624 data points of averaged ERP sampled from 14 EEG electrodes. The implementation of the logistic ICA provided in the toolbox is somewhat intermediate between equation 1.1 and its off-line counterpart: $H(Y)$ is averaged through subblocks of the data set. The nonlinear function is taken to be $\psi(y) = \frac{2}{1+e^{-y}} - 1 = \tanh\frac{y}{2}$. This is minus the log-derivative $\psi(y) = -\frac{r'(y)}{r(y)}$ of the density $r(y) = \beta\frac{1}{\cosh(y/2)}$ (β is a normalization constant). Therefore, this method maximizes over A the likelihood of model $X = AS$ under the assumptions that S has independent components with densities equal to $\beta\frac{1}{\cosh(y/2)}$.

Figure 1 shows the components Y^{JADE} produced by JADE (first column) and the components Y^{LICA} produced by the logistic ICA included in Makeig's toolbox, which was run with all the default options; the third column shows the difference between the components at the same scale. This direct comparison is made possible with the following postprocessing: the components Y^{LICA} were normalized to have unit variance and were sorted by increasing values of kurtosis. The components Y^{JADE} have unit variance by construction; they were sorted and their signs were changed to match Y^{LICA}. Figure 1 shows that Y^{JADE} and Y^{LICA} essentially agree on 9 of 14 components.

[3] Available from http://www.cnl.salk.edu/~scott/.

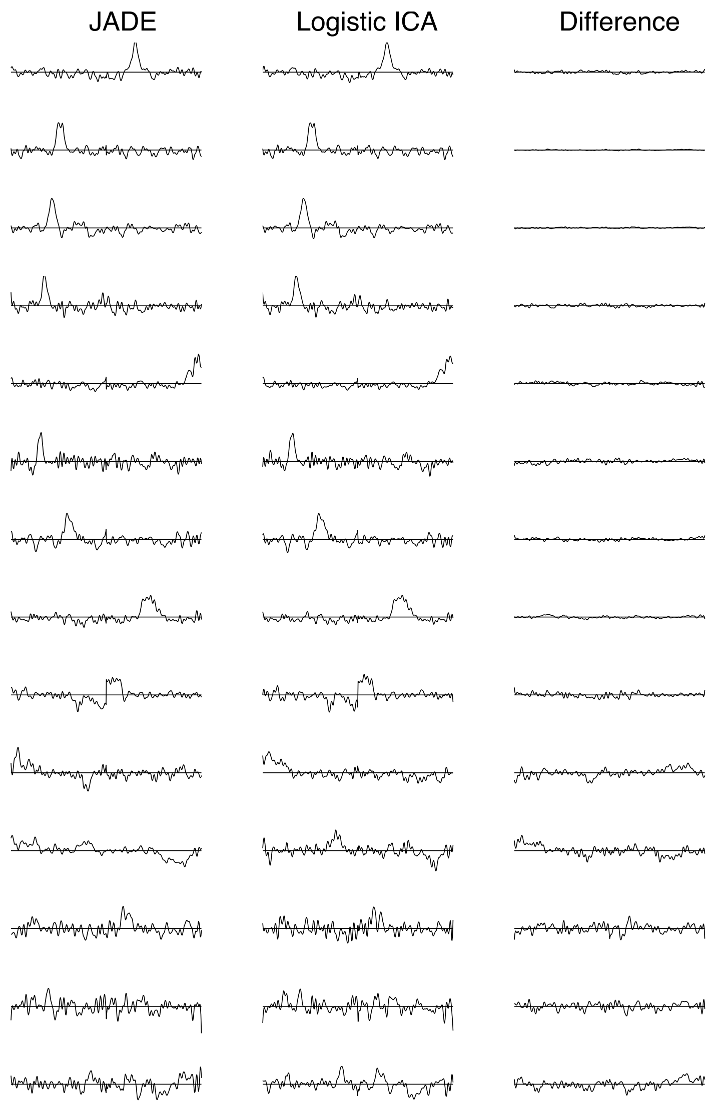

Figure 1: The source signals estimated by JADE and the logistic ICA and their differences.

Another illustration of this fact is given by the first row of Figure 2. The left panel shows the magnitude $|C_{ij}|$ of the entries of the transfer matrix C such that $C\,Y^{\mathrm{LICA}} = Y^{\mathrm{JADE}}$. This matrix was computed after the postprocessing of the components described in the previous paragraph: it should be the identity matrix if the two methods agreed, even only up to scales, signs, and permutations. The figure shows a strong diagonal structure in the northeast block while the disagreement between the two methods is apparent in the gray zone of the southwest block. The right panel shows the kurtosis $k(Y_i^{\mathrm{JADE}})$ plotted against the kurtosis $k(Y_i^{\mathrm{LICA}})$. A key observation is that the two methods do agree about the most kurtic components; these also are the components where the time structure is the most visible. In other words, the two methods essentially agree wherever an human eye finds the most visible structures. Figure 2 also shows the results of SHIBBS and MaxKurt. The transfer matrix C for MaxKurt is seen to be more diagonal than the transfer matrix for JADE, while the transfer for SHIBBS is less diagonal. Thus, the logistic ICA and MaxKurt agree more on this data set. Another figure (not included) shows that JADE and SHIBBS are in very close agreement over all components.

These results are very encouraging because they show that various ICA algorithms agree wherever they find structure on this particular data set. This is very much in support of the ICA approach to the processing of signals for which it is not clear that the model holds. It leaves open the question of interpreting the disagreement between the various contrast functions in the swamp of the low kurtosis domain.

It turns out that the disagreement between the methods on this data set is, in our view, an illusion. Consider the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the covariance matrix R^X of the observations. They are plotted on a dB scale (this is $10 \log_{10} \lambda_i$) in Figure 3. The two least significant eigenvalues stand rather clearly below the strongest ones with a gap of 5.5 dB. We take this as an indication that one should look for 12 linear components in this data set rather than 14, as in the previous experiments. The result is rather striking: by running JADE and the logistic ICA on the first 12 principal components, an excellent agreement is found over all the 12 extracted components, as seen on Figure 4. This observation also holds for MaxKurt and SHIBBS as shown by Figure 5.

Table 2 lists the number of floating-point operations (as returned by Matlab) and the CPU time required to run the four algorithms on a SPARC 2 workstation. The MaxKurt technique was clearly the fastest here; however, it was applicable only because we were looking for components with positive kurtosis. The same is true for the version of logistic ICA considered in this experiment. It is not true of JADE or SHIBBS, which are consistent as soon as at most one source has a vanishing kurtosis, regardless of the sign of the nonzero kurtosis (Cardoso & Souloumiac, 1993). The logistic ICA required only about 50% more time than JADE. The SHIBBS algorithm is

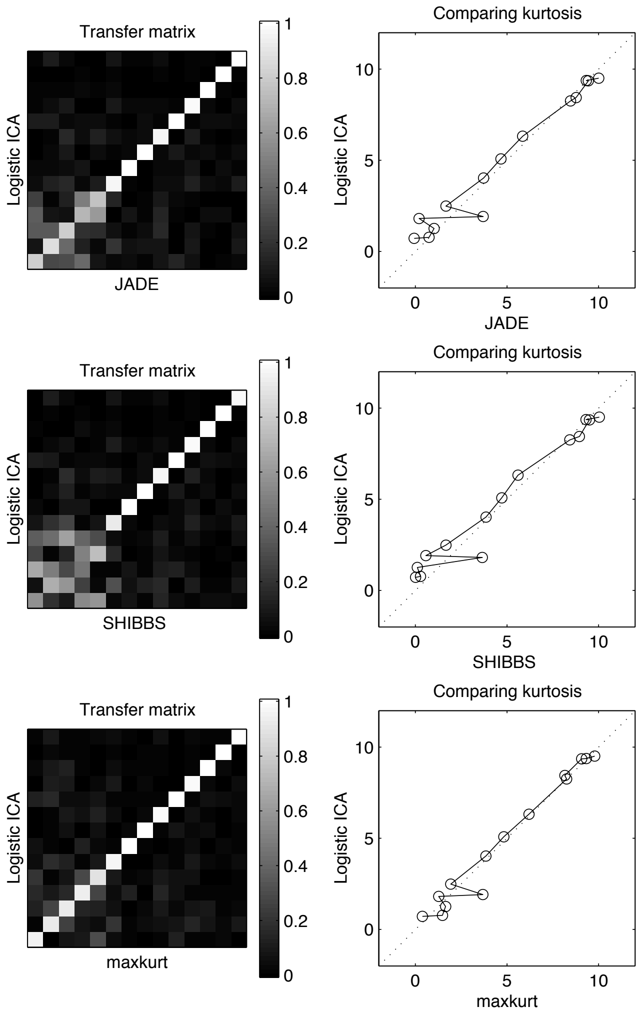

Figure 2: (Left column) Absolute values of the coefficients $|C_{ij}|$ of a matrix relating the signals obtained by two different methods. A perfect agreement would be for $C = I$: deviation from diagonal indicates a disagreement. The signals are sorted by kurtosis, showing a good agreement for high kurtosis. (Right column) Comparing the kurtosis of the sources estimated by two different methods. From top to bottom: JADE versus logistic ICA, SHIBBS versus logistic ICA, and maxkurt versus logistic ICA.

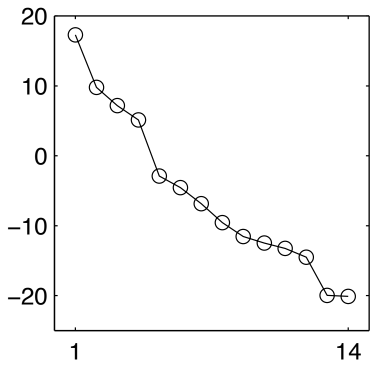

Figure 3: Eigenvalues of covariance matrix R^X of the data in dB (i.e., $10\log_{10}(\lambda_i)$).

Table 2: Number of Floating-Point Operations and CPU Time.

	Flops	CPU Secs.	Flops	CPU Secs.
Method	14 Components		12 Components	
Logistic ICA	5.05e+07	3.98	3.51e+07	3.54
JADE	4.00e+07	2.55	2.19e+07	1.69
SHIBBS	5.61e+07	4.92	2.47e+07	2.35
MaxKurt	1.19e+07	1.09	5.91e+06	0.54

slower than JADE here because the data set is not large enough to give it an edge. These remarks are even more marked when comparing the figures obtained in the extraction of 12 components. It should be clear that these figures do not prove much because they are representative of only a particular data set and of particular implementations of the algorithms, as well as of the various parameters used for tuning the algorithms. However, they do disprove the claim that algebraic-cumulant methods are of no practical value.

6 Summary and Conclusions

The definitions of classic entropic contrasts for ICA can all be understood from an ML perspective. An approximation of the Kullback-Leibler divergence yields cumulant-based approximations of these contrasts. In the or-

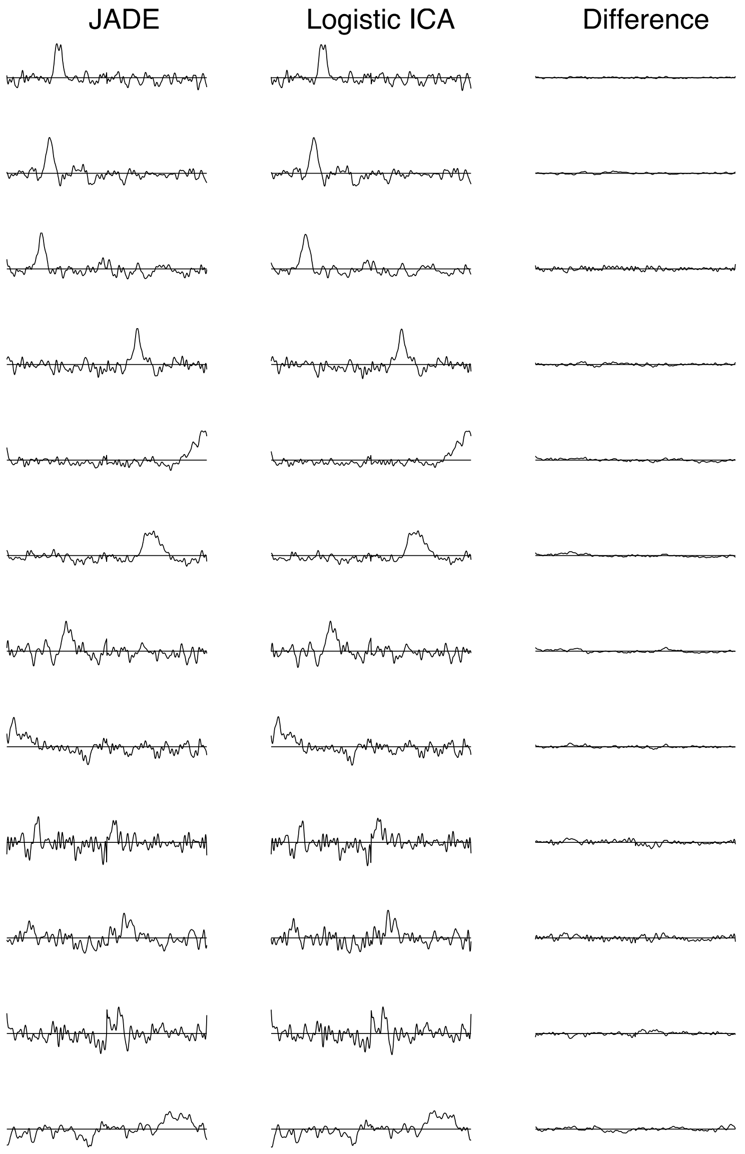

Figure 4: The 12 source signals estimated by JADE and a logistic ICA out of the first 12 principal components of the original data.

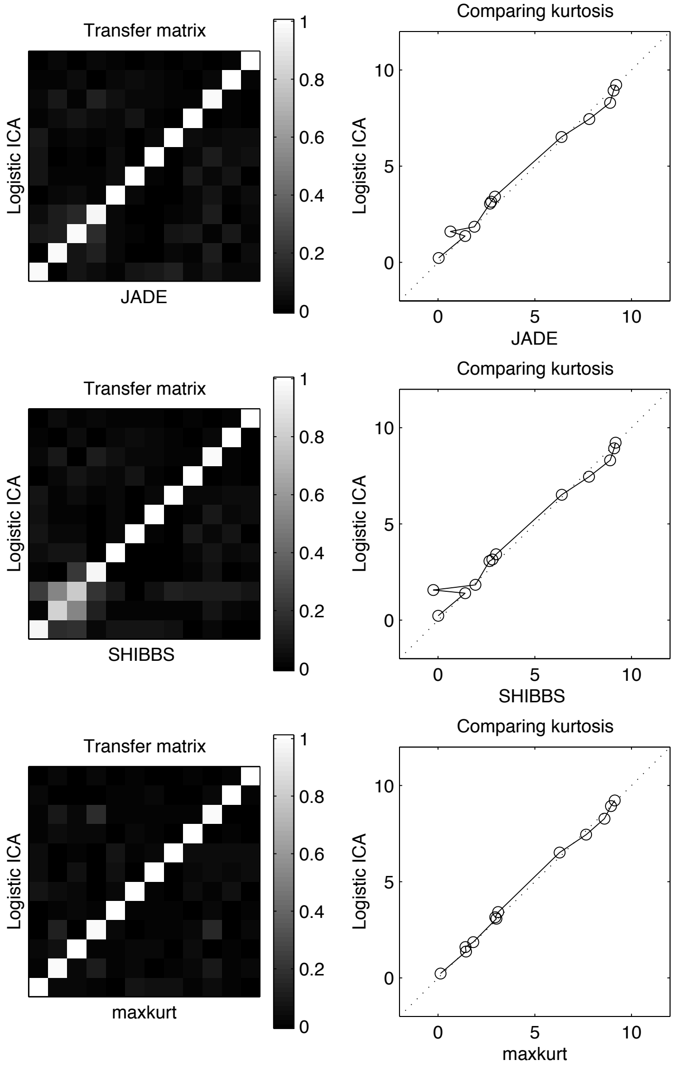

Figure 5: Same setting as for Figure 2 but the processing is restricted to the first 12 principal components, showing a better agreement among all the methods.

thogonal approach to ICA where decorrelation is enforced, the cumulant-based contrasts can be optimized with Jacobi techniques, operating on either the data or statistics of the data, namely, cumulant matrices. The structure of the cumulants in the ICA model can be easily exploited by algebraic identification techniques, but the simple versions of these techniques are not equivariant. One possibility for overcoming this problem is to exploit the joint algebraic structure of several cumulant matrices. In particular, the JADE algorithm bridges the gap between contrast-based approaches and algebraic techniques because the JADE objective is both a contrast function and the expression of the eigenstructure of the cumulants. More generally, the algebraic nature of the cumulants can be exploited to ease the optimization of cumulant-based contrasts functions by Jacobi techniques. This can be done in a data-based or a statistic-based mode. The latter has an increasing relative advantage as the number of available samples increases, but it becomes impractical for large numbers n of components since the number of fourth-order cumulants grows as $O(n^4)$. This can be overcome to a certain extent by resorting to SHIBBS, which iteratively recomputes a number $O(n^3)$ of cumulants.

An important objective of this article was to combat the prejudice that cumulant-based algebraic methods are impractical. We have shown that they compare very well to state-of-the-art implementations of adaptive techniques on a real data set.

More extensive comparisons remain to be done involving variants of the ideas presented here. A technique like JADE is likely to choke on a very large number of components, but the SHIBBS version is not as memory demanding. Similarly, the MaxKurt method can be extended to deal with components with mixed kurtosis signs. In this respect, it is worth underlining the analogy between the MaxKurt update and the relative gradient update, equation 1.1, when function $H(\cdot)$ is in the form of equation 1.5.

A comment on tuning the algorithms: In order to code an all-purpose ICA algorithm based on gradient descent, it is necessary to devise a smart learning schedule. This is usually based on heuristics and requires the tuning of some parameters. In contrast, Jacobi algorithms do not need to be tuned in their basic versions. However, one may think of improving on the regular Jacobi sweep through all the pairs in prespecified order by devising more sophisticated updating schedules. Heuristics would be needed then, as in the case of gradient descent methods.

We conclude with a negative point about the fourth-order techniques described in this article. By nature, they optimize contrasts corresponding somehow to using linear-cubic nonlinear functions in gradient-based algorithms. Therefore, they lack the flexibility of adapting the activation functions to the distributions of the underlying components as one would ideally do and as is possible in algorithms like equation 1.1. Even worse, this very type of nonlinear function (linear cubic) has one major drawback: potential sensitivity to outliers. This effect did not manifest itself in the examples presented in this article, but it could indeed show up in other data sets.

Appendix: Derivation and Implementation of MaxKurt

A.1 Givens Angles for MaxKurt. An explicit form of the MaxKurt contrast as a function of the Givens angles is derived. For conciseness, we denote $[ijkl] = Q^Y_{ijkl}$ and we define

$$a_{ij} = \frac{[iiii] + [jjjj] - 6[iijj]}{4} \quad b_{ij} = [iiij] - [jjji] \quad \lambda_{ij} = \sqrt{a_{ij}^2 + b_{ij}^2}. \tag{A.1}$$

The sum of the kurtosis for the pair of variables Y_i and Y_j after they have been rotated by an angle θ depends on θ as follows (where we set $c = \cos(\theta)$ and $s = \sin(\theta)$):

$$\begin{aligned}
& k(\cos(\theta)Y_i + \sin(\theta)Y_j) + k(-\sin(\theta)Y_i + \cos(\theta)Y_j) && \text{(A.2)}\\
&= c^4[iiii] + 4c^3s[iiij] + 6c^2s^2[iijj] + 4cs^3[ijjj] + s^4[jjjj] && \text{(A.3)}\\
&\quad + s^4[iiii] - 4s^3c[iiij] + 6s^2c^2[iijj] - 4sc^3[ijjj] + c^4[jjjj] && \text{(A.4)}\\
&= (c^4 + s^4)([iiii] + [jjjj]) + 12c^2s^2[iijj] + 4cs(c^2 - s^2)([iiij] - [jjji]) && \text{(A.5)}\\
&\overset{c}{=} -8c^2s^2\frac{[iiii] + [jjjj] - 6[iijj]}{4} + 4cs(c^2 - s^2)([iiij] - [jjji]) && \text{(A.6)}\\
&= -2\sin^2(2\theta)a_{ij} + 2\sin(2\theta)\cos(2\theta)b_{ij} \overset{c}{=} \cos(4\theta)a_{ij} + \sin(4\theta)b_{ij} && \text{(A.7)}\\
&= \lambda_{ij}\left(\cos(4\theta)\cos(4\Omega_{ij}) + \sin(4\theta)\sin(4\Omega_{ij})\right) = \lambda_{ij}\cos(4(\theta - \Omega_{ij})). && \text{(A.8)}
\end{aligned}$$

where the angle $4\Omega_{ij}$ is defined by

$$\cos(4\Omega_{ij}) = \frac{a_{ij}}{\sqrt{a_{ij}^2 + b_{ij}^2}} \quad \sin(4\Omega_{ij}) = \frac{b_{ij}}{\sqrt{a_{ij}^2 + b_{ij}^2}}. \tag{A.9}$$

This is obtained by using the multilinearity and the symmetries of the cumulants at lines A.3 and A.4, followed by elementary trigonometrics.

If Y_i and Y_j are zero-mean and sphered, $\mathrm{E}Y_iY_j = \delta_{ij}$, we have $[iiii] = Q^Y_{iiii} = \mathrm{E}Y_i^4 - 3\mathrm{E}^2Y_i^2 = \mathrm{E}Y_i^4 - 3$ and for $i \neq j$: $[iiij] = Q^Y_{iiij} = \mathrm{E}Y_i^3Y_j$ as well as $[iijj] = Q^Y_{iijj} = \mathrm{E}Y_i^2Y_j^2 - 1$. Hence an alternate expression for a_{ij} and b_{ij} is:

$$a_{ij} = \frac{1}{4}\mathrm{E}\left(Y_i^4 + Y_j^4 - 6Y_i^2Y_j^2\right) \quad b_{ij} = \mathrm{E}\left(Y_i^3Y_j - Y_iY_j^3\right). \tag{A.10}$$

It may be interesting to note that all the moments required to determine the Givens angle for a given pair (i, j) can be expressed in terms of the two variables $\xi_{ij} = Y_iY_j$ and $\eta_{ij} = Y_i^2 - Y_j^2$. Indeed, it is easily checked that for a zero-mean sphered pair (Y_i, Y_j), one has

$$a_{ij} = \frac{1}{4}\mathrm{E}\left(\eta_{ij}^2 - 4\xi_{ij}^2\right) \quad b_{ij} = \mathrm{E}\left(\eta_{ij}\xi_{ij}\right). \tag{A.11}$$

A.2 A Simple Matlab Implementation of MaxKurt. A Matlab implementation could be as follows, where we have tried to maximize readability but not the numerical efficiency:

```
function Y = maxkurt(X) %
[n T] = size(X)  ;
Y     = X - mean(X,2)*ones(1,T);  % Remove the mean
Y     = inv(sqrtm(X*X’/T))*Y   ;  % Sphere the data
encore = 1                     ;  % Go for first sweep
while encore, encore=0;
  for p=1:n-1,                    % These two loops go
    for q=p+1:n,                  % through all pairs
      xi    = Y(p,:).*Y(q,:);
      eta   = Y(p,:).*Y(p,:) - Y(q,:).*Y(q,:);
      Omega = atan2( 4*(eta*xi’), eta*eta’ - 4*(xi*xi’) );

      if abs(Omega) > 0.1/sqrt(T)  % A ‘statistically small’
                                   % angle
         encore = 1              ; % This will not be the
                                   %last sweep
         c          = cos(Omega/4);
         s          = sin(Omega/4);
         Y([p q],:) = [ c s ; -s c ] * Y([p q],:) ; % Plane
                                                    % rotation
        end

      end
    end
end
return
```

Acknowledgments

We express our thanks to Scott Makeig and to his coworkers for releasing to the public domain some of their codes and the ERP data set. We also thank the anonymous reviewers whose constructive comments greatly helped in revising the first version of this article.

References

Amari, S.-I. (1996). Neural learning in structured parameter spaces—Natural Riemannian gradient. In *Proc. NIPS*.

Amari, S.-I., Cichocki, A., & Yang, H. (1996). A new learning algorithm for blind signal separation. In *Advances in neural information processing systems, 8* (pp. 757–763). Cambridge, MA: MIT Press.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation, 7*, 1004–1034.

Cardoso, J.-F. (1992). Fourth-order cumulant structure forcing. Application to blind array processing. In *Proc. 6th SSAP workshop on statistical signal and array processing* (pp. 136–139).

Cardoso, J.-F. (1994). On the performance of orthogonal source separation algorithms. In *Proc. EUSIPCO* (pp. 776–779). Edinburgh.

Cardoso, J.-F. (1995). The equivariant approach to source separation. In *Proc. NOLTA* (pp. 55–60).

Cardoso, J.-F. (1997). Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing, 4*, 112–114.

Cardoso, J.-F. (1998). Blind signal separation: statistical principles. *Proc. of the IEEE. Special issue on blind identification and estimation*.

Cardoso, J.-F., Bose, S., & Friedlander, B. (1996). On optimal source separation based on second and fourth order cumulants. In *Proc. IEEE Workshop on SSAP*. Corfu, Greece.

Cardoso, J.-F., & Laheld, B. (1996). Equivariant adaptive source separation. *IEEE Trans. on Sig. Proc., 44*, 3017–3030.

Cardoso, J.-F., & Souloumiac, A. (1993). Blind beamforming for non Gaussian signals. *IEEE Proceedings-F, 140*, 362–370.

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

Comon, P. (1997). Cumulant tensors. In *Proc. IEEE SP Workshop on HOS*. Banff, Canada.

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

De Lathauwer, L., De Moor, B., & Vandewalle, J. (1996). Independent component analysis based on higher-order statistics only. In *Proc. IEEE SSAP Workshop* (pp. 356–359).

Gaeta, M., & Lacoume, J. L. (1990). Source separation without a priori knowledge: The maximum likelihood solution. In *Proc. EUSIPCO* (pp. 621–624).

Golub, G. H., & Van Loan, C. F. (1989). *Matrix computations*. Baltimore: Johns Hopkins University Press.

MacKay, D. J. C. (1996). *Maximum likelihood and covariant algorithms for independent component analysis*. In preparation. Unpublished manuscript.

Makeig, S., Bell, A., Jung, T.-P., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proc. Nat. Acad. Sci. USA, 94*, 10979–10984.

McCullagh, P. (1987). *Tensor methods in statistics*. London: Chapman and Hall.

Moreau, E., & Macchi, O. (1996). High order contrasts for self-adaptive source separation. *International Journal of Adaptive Control and Signal Processing, 10*, 19–46.

Moulines, E., Cardoso, J.-F., & Gassiat, E. (1997). Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. ICASSP'97* (pp. 3617–3620).

Nadal, J.-P., & Parga, N. (1994). Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer. *NETWORK, 5*, 565–581.

Obradovic, D., & Deco, G. (1997). Unsupervised learning for blind source separation: An information-theoretic approach. In *Proc. ICASSP* (pp. 127–130).

Pearlmutter, B. A., & Parra, L. C. (1996). A context-sensitive generalization of ICA. In *International Conference on Neural Information Processing* (Hong Kong).

Pham, D.-T. (1996). Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Trans. on Sig. Proc., 44*, 2768–2779.

Pham, D.-T., & Garat, P. (1997). Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Tr. SP, 45*,1712–1725.

Souloumiac, A., & Cardoso, J.-F. (1991). Comparaison de méthodes de séparation de sources. In *Proc. GRETSI*. Juan les Pins, France (pp. 661–664).

Received February 7, 1998; accepted June 25, 1998.