Sumit Basu. "ICA: A Critical Review of Three Prominent Approaches." Technical Report. April, 2000.

ICA: A Critical Review of Three Prominent Approaches

Sumit Basu
Perceptual Computing Section, The MIT Media Laboratory
20 Ames St., Cambridge, MA 02139 USA
sbasu@media.mit.edu

April 25, 2000

Abstract

We present the problem of Independent Components Analysis (ICA) and review three major approaches as described by Comon [9],Amari et al. [2], and Bell and Sejnowski [5]. We cast all three in a common notational framework, point out their strengths and weaknesses, and show how they are related to each other. We then go on to suggest several empirical studies to further investigate the numerical behavior of the algorithms. Finally, we present a number of novel extensions/applications of ICA. The most interesting of these is our notion of "conditionally independent components analysis," in which we propose factoring the conditional density $p(x|y)$ to greatly reduce the amount of data needed to accurately model a scaled version of $p(y|x)$.

1 Introduction

The problem of Blind Source Separation (BSS) is an old one in signal processing: in the basic setup, there is a vector of statistically independent signals, s, that are mixed together by a mixing matrix M to result in the vector of signals y that is observed. The task is then to estimate a demixing matrix W that will recover the original components of s. At best, we can recover the components modulo a scalefactor and a permutation (i.e., component 1 in s may or may not show up as component 1 in z even under perfect separation). The basic setup is shown in figure 1.

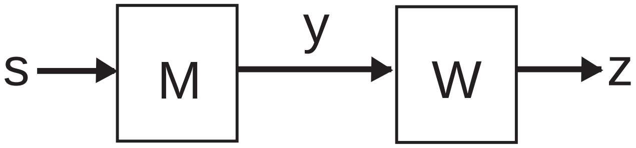

Figure 1: Problem setup for blind source separation. The original sources s are mixed by the matrix M resulting in the observations y; we then attempt to demix them with W resulting in z.

This problem has received renewed attention in recent years in both the signal processing and statistics literature due to a series of new techniques collectively termed "Independent Components Analysis" (ICA). Previous approaches relied primarily on PCA, which models the second order statistics of the data, and as a result can only find the most *decorrelated* outputs. The ICA techniques use higher order statistics in an attempt to recover the most independent components. The statistical tools and approximation techniques thus differ significantly from the usual tools for the analysis of Gaussian random vectors, and as a result have provoked interest in other areas, such as the machine learning community. In addition, the potential for this technique goes beyond mere signal separation. If we have an arbitrary source y that we can split into independent components z, we can estimate the N-dimensional joint density of y by only estimating the 1-dimensional marginals z (and taking their product), which we can do with far less data. This is potentially useful for a large number of estimation and pattern recognition tasks.

In this study, we review three of the major approaches to ICA, discuss their strengths and limitations, and describe a number of potential extensions and applications for these techniques. The approaches we will examine are the pairwise minimization of mutual independence by Pierre Comon [9], the natural gradient work of Shun-ichi Amari and colleagues [2], and the infomax approach of Bell and Sejnowski [5]. This is by no means a comprehensive review of all the work in ICA: [9] and [7] give a much more complete coverage. However, these three approaches do touch upon many of the major ideas in the field, and as such this study may provide a useful guide to the other techniques described in the literature.

Note that there are several related problems often discussed in the ICA context that we will not be discussing here. The first of these is blind deconvolution, in which we wish to separate a one-dimensional signal into a source signal and a filter, where we assume the source signal was white (and thus find the filter that best decorrelates the signal from its past). The other is the combination of blind source separation and blind deconvolution, in which the mixing system M is a matrix of FIR filters, and we wish to find the matrix of filters W that gives us the most independent components. Such a model can account for real-world effects such as the propagation delays between sensors. The formulations of these problems (as mentioned in [5, 1]) are similar to that of BSS, but the nature of their solutions do seem to yield significant new insights into the methods we will describe. Last, while many of the results from BSS can be extended into the complex domain, for the sake of simplicity we will limit our discussion to real variables.

1.1 Notation and Basic Quantities

Before we begin, we will review some basic quantities that will be used repeatedly in the discussion. The first of these is the *differential entropy*, as described in Chapter 9 of [10]:

$$H(z) = -\int p(z) \log p(z) dz \tag{1}$$

This is the analogue of the Shannon entropy for continuous variables. However, it does not share all of the same properties - most importantly, the differential entropy for a deterministic variable diverges to $-\infty$, and it is not invariant to an arbitrary invertible transform (though it is invariant to an orthogonal transform [9]). We will refer to this quantity as the entropy in the context of continuous variables.

The next is the continuous form of the *Kullback-Liebler divergence* (KL divergence) between two densities:

$$KL(p(z)||q(z)) = \int p(z) log \frac{p(z)}{q(z)} dz \tag{2}$$

Though often referred to as a "distance," it does not fulfill the symmetry requirements for a distance measure. However, note that it is minimized at zero and only reaches this minimum when the densities are (almost) equal.

We can express the *mutual information* (MI) between two variables x and y as the KL divergence between the product of their marginals and their joint density, which we can easily interpret as a "divergence from independence:" If x and y are independent, the product of the marginals will equal the joint and divergence will be zero. As a result, minimizing the mutual information of the outputs in figure 1 is a common means of seeking independent components. We notate mutual information as follows:

$$I(x;y) = KL(p(x,y)||p(x)p(y)) \tag{3}$$

Note that we can rewrite this in a number of useful forms:

$$\begin{aligned} I(x;y) &= KL(p(x|y)||p(x)) & (4) \\ &= H(x) + H(y) - H(x,y) & (5) \end{aligned}$$

The next concept is that of *cumulants*, which are the Taylor coefficients of the log of the characteristic function (Fourier transform) of a PDF [12]. Like the moments of a multi-dimensional variable, the nth order cumulant is an nth order tensor. We denote scalar cumulants of order n as κ_n and the $ijkl...$ coefficient of a multi-dimensional cumulant as $K_{ijkl...}$. Note that while K is a multi-dimensional tensor, K_{ijkl} is a scalar component within that tensor.

Cumulants have a number of useful properties in characterizing densities that have higher than second-order statistics. For instance, the cumulants of order greater than 2 are all zero for Gaussian random variables. In addition, all cross-cumulants for a set of variables that can be split by a linear transform into two subsets of independent variables are zero [7]. Lastly, the cumulants of order r can be derived from the moments of order r through the relations described in [12].

The last concept is that of a *contrast function*, which is the expression that we will maximize in seeking independence in the components. Obviously, we want this quantity to be maximized only when the components are truly independent. When we say that a quantity is "proved to be a contrast" (as with Comon), we mean that this property is guaranteed to hold. In other cases, as our analysis of the Bell and Sejnowski work will show, we have no guarantees that even the global maximimum implies independence.

2 Three Approaches to the Problem

In this section, we describe the three approaches listed above and comment on their strengths and weaknesses. We begin with Comon [9] and Amari [2], both of who attempt to model the mutual information through a series of approximations and then minimize it by different means. We will end with Bell and Sejnowski [5], who take the approach of maximizing the output entropy (which, as we will see, is related to minimizing mutual information and exactly matches a maximum-likelihood criterion). Note that in the descriptions below, I have standardized the notation as much as possible to that of figure 1, which thus will often differ from the original notation of the papers.

2.1 Comon: A Pairwise Batch Algorithm

Of the three papers reviewed, Comon [9] gives the most complete analysis of the problem, justifying each of his steps in detail, as well as providing a substantial literature review. He begins his development by pointing out the inherent indeterminacy of ICA (due to the scaling (including negation) and permutation possibilities stated earlier). He thus requires that the columns of W have unit norm, that the entries of Δ (where Δ is diagonal and the covariance of y, $K_y = W\Delta^2W^T$) are sorted in order of decreasing value, and the component in each column of W largest absolute value be positive. This forces the ICA to have a unique solution, though the demixed outputs can still be permuted and scaled with respect to the original inputs.

Comon makes a number of assumptions to make the problem theoretically tractable. First, he assumes that we are working in the noiseless case. Second, he requires that only one component can be Gaussian. Last, he requires that W is orthogonal. We will see the advantages of the latter two assumptions shortly. He then "standardizes" the observations y by whitening them, i.e., normalizing the variables to have an identity covariance matrix. If the covariance can be decomposed as

$$K_y = U\Lambda U^T \tag{6}$$

where Λ is diagonal, then the normalized y can be written as $z = \Lambda^{-1}U'y$, so we are initializing W with $\Lambda^{-1}U^T$. If K_y is singular, Comon suggests using the SVD to determine the projection of the data onto a space (i.e., the basis vectors corresponding to the non-zero singular values) in which we will have a non-singular covariance. Note that this transform is decorrelating the inputs. If we cannot find N decorrelated components, we can hardly hope to find N independent components, so such a projection is quite reasonable. In this case, we can break down the entire N (number of dimensions of y) by T (number of samples) data matrix with the SVD as $Y = V\Sigma U$ (i.e., the estimated covariance is then $\frac{1}{T}V\Sigma^2 V'$) and then initialize W to $\frac{\Sigma^{-1}V^T}{\sqrt{T}}$.

He then goes on to describe the negentropy $J(p_x)$, which is the difference in entropy between x and a Gaussian and can be considered a distance from normality:

$$J(p_z) = S(\Phi_z) - S(p_z) \tag{7}$$

where Φ_z has the same mean and variance as z. Since a Gaussian has the largest entropy for a given mean and variance, this quantity is guaranteed to be positive and is further guaranteed to be zero only when p_z is Gaussian (almost) everywhere. In addition, the negentropy is invariant to any invertible linear transform on z. He then writes the mutual information of the components of z in terms of this quantity:

$$I(z) = J(p_z) - \sum_{i=1}^{N} J(p_{z_i}) + \frac{1}{2}\log\frac{\prod V_{ii}}{|V|} \tag{8}$$

where V denotes the covariance of z. After this point, we will only modify W and thus z by orthogonal transforms (N by N matrices Q). As a result, the first term in the mutual information will not change because of the invariance we mentioned above. As for the third term, remember that we have normalized z with a whitening transform such that its covariance V is diagonal, which obviously makes this term zero. Furthermore, when we transform z to Qz where Q is orthogonal, the resulting covariance is QVQ^T, which is still diagonal. Thus, as a result of restricting ourselves to orthogonal transforms, the first term remains constant and the third term is always zero. We then only need to minimize the second term, the sum of the marginal negentropies, with respect to Q.

Of course, the densities of the marginal z_i's are not available, so Comon approximates the negentropy in terms of higher order moments using the Edgeworth expansion [15]. He chooses the Edgeworth expansion over the apparently more conventional Gram-Charlier expansion because he claims the terms are ordered in terms of decreasing significance as a function of $m^{-1/2}$, where z is modeled as being made up of the sum of m independent random variables. For a standardized (zero-mean, unit variance) scalar variable, this expansion is:

$$J(p_x) = \frac{1}{12}\kappa_3^2 + \frac{1}{48}\kappa_4^2 + \frac{7}{48}\kappa_3^4 - \frac{1}{8}\kappa_3^2\kappa_4 + O(m^{-2}) \tag{9}$$

Dropping $O(m^{-2})$ terms, he then wishes to maximize the following contrast function in order to minimize the mutual information:

$$\Psi(Q) = \sum_{i=1}^{N} 4K_{iii}^2 + K_{iiii}^2 + 7K_{iii}^4 - 6K_{iii}^2 K_{iiii} \tag{10}$$

However, at this point he backs off to a simplified version of this expression:

$$\Psi(Q) = \sum_{i=1}^{N} K_{iiii}^2 \tag{11}$$

and proves that this is a contrast function as well as long as the fourth-order cumulants (or r-th order, if $K_{iii\ldots i}^2$ are used) null for at most one component. Clearly, this is not as powerful as the earlier contrast, but will greatly simplify the maximization of this function with respect to Q.

Comon then shows that the third and fourth-order cumulants of z can be related to the cumulants of y through the following relation, where Γ represents the appropriate cumulant for y:

$$K_{ijk} = \sum_{pqr} Q_{ip} Q_{jq} Q_{kr} \Gamma_{pqr} K_{ijkl} = \sum_{pqrs} Q_{ip} Q_{jq} Q_{kr} Q_{ls} \Gamma_{pqrs} \tag{12}$$

As a result, we can differentiate the contrast function with respect to Q. Comon showed at an earlier point in the paper if at most one source component is Gaussian, then pairwise independence is equivalent to mutual independence, and indeed upon differentiating equation 11 he finds that only terms from two distinct indices of z are involved. As a result, Comon's algorithm boils down to looping through pairs of components in z and rotating them in the plane in which they exist (thus maintaining orthogonality of Q) such that $\Psi(Q)$ is maximized. Amazingly, this inner loop reduces to solving a polynomial of degree at most five, so the solution for each step can be found analytically. Furthermore, he proves in [8] (which is unfortunately only available in French) that the contrast monotonically increases with each such rotation.

Comon does not have theoretical bounds for how many pairs need to be processed, but he finds empirically that $1 + \sqrt{N}$ sweeps (where a sweep is an iteration through all unique pairs) are sufficient. With this in mind, he empirically estimates the order of operations at $O(N^4)$. One unfortunate aspect of the algorithm that adds significantly to its computational complexity is that z is calculated explicitly after each update of W through Q. The author mentions that the relationship between the cumulants of the observation y and the cumulants of z described in (12 above would result in considerable savings per iteration, but this would require computing the entire family of cumulants for y, which would apparently outweigh the computational cost of the entire algorithm as it stands.

Overall, the technique of Comon is well-motivated and seems to perform well. It increases the contrast monotonically and completes in low-order polynomial time. In empirical tests, he shows good convergence results, even when the signal to noise ratio is quite low (though he only shows the distance of WM from a permutation of the scaled identity and not the actual signal to noise ratio of the separated components). However, there are a number of places where his approach leaves something to be desired. First, while it appears fairly efficient, it is a batch procedure and there are no obvious ways to make it online since the the latest W has to be applied to all points before the required cumulants can be estimated. In addition, although requiring the demixing the matrix W to be orthogonal greatly simplifies the math, it greatly reduces the range of solutions as well, effectively removing one degree of freedom (since all pairs of axes must be orthogonal). As a result, for example, two arbitrarily mixed signals could not be separated using only two sensors - three would be necessary unless they happened to be mixed with an orthogonal matrix (as is the case in his experiments). Next, while he has a nice development and an accurate expansion of the mutual information, he drops it for the simpler form in (11). While the latter is still a contrast, it results in the further requirement that the fourth cumulant is non-null for all but one sources. More importantly, it means we are no longer really minimizing the mutual information or even an approximation thereof. Last, while this is not a criticism, it is interesting to notice from (10) that even in the ideal form, all we are doing is minimizing the sum of the marginal negentropies, which is equivalent to maximizing the sum of the marginal entropies with the additional constraint that we have an orthogonal demixing matrix. This will give us something of a link when we consider Bell and Sejnowski's infomax approach later on.

2.2 Amari: Maximizing the Contrast Using Natural Gradient

Amari [2] follows the same path as Comon in the early stages – he also attempts to minimize the mutual information of the components of z (though he always refers to it in the KL divergence form), and notes that it can be written as:

$$I(z) = -H(z) + \sum_{i=1}^{N} H(z_i) \tag{13}$$

He also normalizes the components to have zero mean and unit variance (but does not decorrelate them as Comon). He then expands the marginal densities $p(z_i)$ using the Gram-Charlier expansion to find

$$p(z_i) \approx \frac{1}{\sqrt{2\pi} e^{-\frac{z_i^2}{2}}} \left[1 + \frac{K_{iii}}{3!} H_3(z_i) + \frac{K_{iiii}}{4!} H_4(z_i) \right] \tag{14}$$

where $H_i(x)$ are the Hermite polynomials [15]. Amari uses the Gram-Charlier expansion instead of the Edgeworth (as used by Comon) because it more clearly shows the role of the third and fourth order cumulants in estimating the marginal PDF's. In the above, we can easily see how when these cumulants are zero, the density is Gaussian, and how the density varies as their values change. Furthermore, he argues that the Edgeworth expansion only shows the ordering in magnitude that Comon takes advantage of for special distibutions, such as the sum of iid random variables. This argument strikes us as somewhat strange, in that the ICA problem by definition involves the sum of a number of independent random variables, though perhaps not identically distributed. Note that Comon claimed the ordering held as long as the marginal was a sum of independent random variables with finite cumulants, referring to a theorem by Cramer in [14] – he did not qualify that that the variables had to be identically distributed.

Returning to our development, the author then finds the entropy of the marginals using the expansion in (14):

$$H(z_i) \approx \frac{1}{2}log(2\pi e) - \frac{K_{iii}^2}{2 \cdot 3!} - \frac{K_{iiii}^2}{2 \cdot 4!} + \frac{5}{8}K_{iii}^2 K_{iiii} + \frac{1}{16}K_{iiii}^3 \tag{15}$$

Since $H(z)$ can be written as simply $H(y) + \log|W|$, he can approximate the whole of the mutual information as

$$I(z) \approx -H(y) - \log|W| + \frac{N}{2}log(2\pi e) - \sum_{i=1}^{N}\left[\frac{K_{iii}^2}{2 \cdot 3!} + \frac{K_{iiii}^2}{2 \cdot 4!} - \frac{5}{8}K_{iii}^2 K_{iiii} - \frac{1}{16}K_{iiii}^3\right] \tag{16}$$

The negative of this approximation to the information is thus his contrast function. In order to find the optimal W, he uses "natural gradient" descent directly on the mutual information. In order to formulate this, he first takes the ordinary gradient $\frac{dI(z)}{dW}$, and then converts this to a derivative in time $\frac{dI(z)}{dt}$ by multiplying by a learning rate $\eta(t)$ which controls the step size over time. He expresses this elementwise for W:

$$\frac{dW_{i,j}}{dt} = \eta(t)\left[W_{i,j}^{-T} - f(K_{iii}, K_{iiii})z_i^2 y_j - g(K_{iii}, K_{iiii})z_i^3 y_j\right] \tag{17}$$

where f and g are simple second order polynomials. This initial derivative is of course in terms of the third and fourth order cumulants of the marginals, which are difficult to estimate since W is continuously changing. Amari thus proposes replacing their estimated values with their instantaneous values, resulting in

$$\frac{dW_{i,j}}{dt} = \eta(t)\left[W_{i,j}^{-T} - f(z_i)y_j\right] \tag{18}$$

where $f(z)$ is

$$f(z) = \frac{3}{4}y^{11} + \frac{25}{4}y^9 - \frac{14}{3}y^7 - \frac{47}{4}y^5 + \frac{29}{4}y^3 \tag{19}$$

In vector form, this gradient can be written as

$$\frac{dW}{dt} = \eta(t) \left[I - f(y) y^T \right] W^{-T} \tag{20}$$

At this point, Amari invokes the concept of the natural gradient. The basic idea is that many times, the parameter space in which we are working (in this case, the space of non-singular N by N matrices) forms a Riemannian manifold in some Euclidean space (i.e., the space of all N by N matrices). In such cases, instead of doing the gradient descent directly in the Euclidean space, we can achieve a more direct descent by exploiting the Riemannian structure of the manifold (see [3, 1] for details of the development). The disadvantage of this method is that determining the structure of the parameter space can be difficult and can lead to extremely complex algorithms. However, in the case of non-singular matrices, it is quite simple and amounts to postmultiplying the gradient in (20) by $W^T W$. This gives us the following modified gradient:

$$\frac{dW}{dt} = \eta(t) \left[I - f(y) y^T \right] W \tag{21}$$

Note that even though this may be a more effective gradient, we still have the standard gradient descent problem of choosing a step size. In principle, we could reevaluate $I(z)$ after each proposed step and adjust the step size until we are guaranteed a decrease at every step, but Amari does not seem to think this is necessary. As far as we could see, there are thus no guarantees of monotonic convergence using an arbitrary $\eta(t)$.

The author shows one experiment in which he chooses the mixing matrix M randomly and uses a learning rate $\eta(t) = 250e^{-5t}$. We have reproduced hi plots from this experiment in figure 2. The error measure $E1$ is again a distance of WM from a scaled, permuted identity. We can see that the separated signals z seem to be fairly similar to the stated forms of the sources s by the end (see the caption for details on the sources). However, $E1$ does appear to converge at all - it is not principally smaller or even more stable at the end of the trial than it is at early points. It is quite surprising that this is the case and yet the signals appear to approach the original sources. It would be interesting to see the value of the MI at each timestep to see if at least this quantity is achieving some kind of convergence.

Overall, the method of Amari has a number of advantages over that of Comon. His method is online – it adapts one data point at a time – making real-time implementations a possibility. His allowable class of transformations is the set of non-singular matrices as opposed to Comon's restriction to orthogonal ones. However, there are some issues that remain to be resolved, particularly in terms of convergence. Certainly from the one example he presents in the paper, it is not clear that the algorithm is converging stably to a permutation of the inputs. While his argument for using the natural gradient is appealing, its convergence properties in the case of BSS are not immediately obvious. In [1], section 7, he takes a variant on the algorithm which matches the Bell and Sejnowski

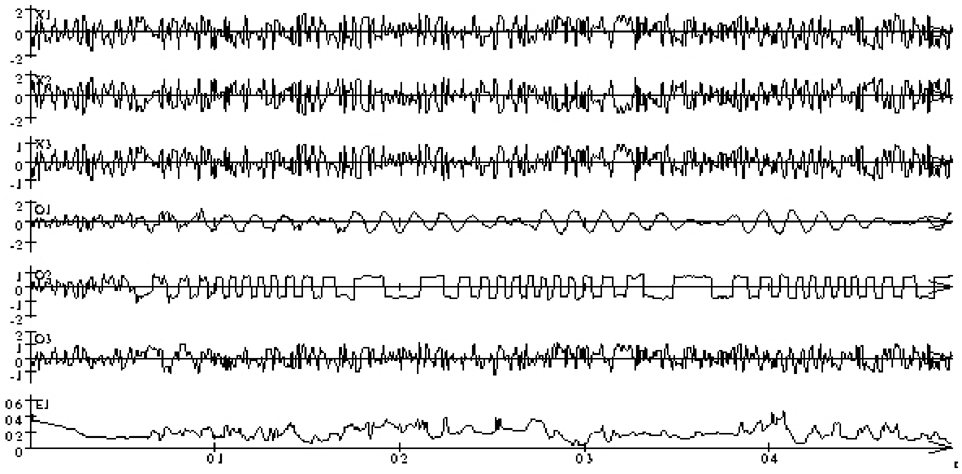

Figure 2: Empirical results from Amari's method (from Amari, figure 1). The first three signals are the mixed observations y (a sine modulated by a low-frequency cosine, a square wave modulated in frequency by a low-frequency sine, and a noise signal), the next three are the unmixed outputs, z, and the last is $E1$, the measure of convergence for the unmixing system MW. Note the lack of convergence of $E1$ despite the good separation results.

formulation. In this case, he argues that the natural gradient algorithm is "efficient and stable," referring to another paper [4] for the details. There is also the issue of choosing the step size in such a way to monotonically decrease the MI. Amari presents an algorithm in section 4 of [1] for adaptively modifying $\eta(t)$. However, he does not present numerical results or theoretical guarantees for either of these techniques. In the end, while Amari's arguments are mathematically appealing, it is not clear how well his methods work in practice.

2.3 Bell and Sejnowski: An Infomax Approach

The last approach we will discuss is the infomax formulation of Bell and Sejnowski [5]. Their method is quite different from the preceding two. Given an input y, they transform it through

$$z = g(Wx + w_0) \tag{22}$$

where g is a fixed, monotonic, component-wise non-linearity that maps the real line to [0,1] (we will see more on the meaning of this later). To find the best demixing matrix, they then simply maximize the output entropy, $H(z)$, as their contrast function. Their justification for this rule is that since

$$H(z) = H(z_1) + H(z_2) + \cdots + H(z_N) - I(z) \tag{23}$$

maximizing the output entropy amounts to minimizing the mutual information and thus finding the statistically most independent components. This has obvious problems in that we have no guarantees that the global maximum of $H(z)$ requires the minimization of $I(z)$ – the increase in the marginal entropies could easily outweigh the reduction in mutual information. We will discuss later how this method can fail and how we can set up the non-linearity so that it is guaranteed to succeed in certain cases.

First, though, let us go through the rest of their formulation and solution. Because z can be written as a function of y, we can relate their densities as follows:

$$p_z(z) = \frac{p_y(f^{-1}(z))}{|J|} \tag{24}$$

Where $f(y)$ is the function relating y and z and J is the Jacobian (the determinant of the matrix of partial derivatives) of f. We can then write the entropy of z as

$$\begin{aligned} H(z) &= -\int p_z(z)\log p_z(z) = -\int \frac{p_y(f^{-1}(z))}{|J|}\log\frac{p_y(f^{-1}(z))}{|J|} && (25) \\ &= \frac{1}{|J|}(H(y) + \log|J|) && (26) \end{aligned}$$

as a result, we can maximize $H(z)$ by maximizing $\log|J|$. When using the logistic function as the nonlinearity, i.e., $g(u) = (1 + e^u)^{-1}$, the authors show how this gives rise to the following gradients:

$$\begin{aligned} \frac{d(H(z))}{dW} &= W^{-T} + (1-2y)x^T && (27) \\ \frac{d(H(z))}{dw_0} &= 1 - 2y && (28) \end{aligned}$$

This system avoids singular matrices since the W^{-T} term begins to blow up and drives the solution away from such points. We still have the problem of choosing the learning rate, but the authors do not address this issue at all.

In their experiments section, the authors show that their algorithm can separate a combination of many (10) signals containing speech, music, laughter, etc. quite well using the logistic function above as the nonlinearity. They also show convergence of WM to a permutation of a scaled identity. However, there are simple cases where it fails, as shown in figure 3 below. In this case, the increase in entropy for the "wrong" components outweighs the increase in mutual information, as we had earlier hinted could be possible. We now investigate why this occurs and how maximizing (23) relates to seeking independence amongst our outputs.

As we described earlier, arbitrarily maximizing $H(z)$ does not guarantee the minimization of the mutual information. However, if we know the distributions of the sources, we can set up a situation in which the

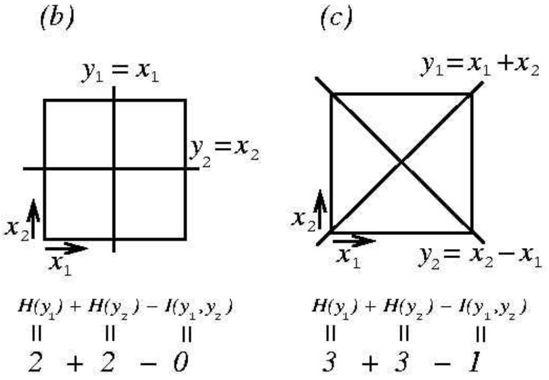

Figure 3: A simple case where the Bell and Sejnowski method fails. The components chosen in (c) maximize the output entropy, while the components in (b) minimize mutual information and are the true independent components (from Bell and Sejnowski, figure 4).

global maximum of $H(z)$ does correspond to finding the independent components. To see this, we have to first realize that the CDF (cumulative distribution function) of a random variable maps its distribution to the uniform distribution on [0,1] (i.e., if $g(x)$ is the CDF of random variable x, the density of $y = g(x)$ is uniform over [0,1]. This is illustrated in figure 4.

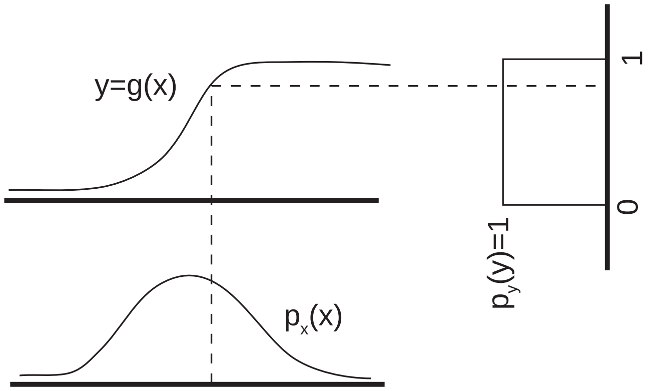

Figure 4: An illustration of how the CDF maps a random variable to one of uniform density. $g(x)$ is the CDF of random variable x, and the density of $y = g(x)$ is uniform over [0,1].

Also recall that over a fixed domain, the uniform is the distribution with the maximum entropy. Now consider the case in which all the input sources have the same distribution. Let us set $g(y)$, the non-linearity from (22) to the CDF for this distribution. When we find a demixing matrix that results in the observations y being split in such a way that the outputs z match the known source distributions, the marginal entropies of $g(y)$ will be maximized, since they will all be distributed as the uniform. In general, this will only occur when we have correctly separated the sources. This is the case since only with Gaussians can two random variables of the same distribution sum to produce another with the same distribution, so as long as more than one of our sources are not Gaussian, this will not present a problem. Note that if the sources are distributed differently, we can set various component of $g(y)$ to match the different source statistics, in which case the global maximization will force the chosen component to appear in the corresponding component of z.

Cardoso shows this result quite elegantly in [6], and it is instructive to follow his development. He slightly simplifies things by removing the bias term w_0, thus making the quantity to be maximized $H(z) = H(g(Wy))$. Since the mapping functions g_i all map the real line to [0,1], they can be interpreted as the CDF's for some set of distributions q_i (i.e., the derivatives of g_i). He then defines $\tilde{s_i}$ as a set of variables that have distributions q_i. Since the (1) negative of the entropy of a random variable is equivalent to its KL divergence from the uniform and (2) the KL divergence is invariant to a transform applied to both arguments, he applies both properties to $H(z)$ to rewrite the function we are maximizing to:

$$-K(g(Wy)||u) = -K(Wy||\tilde{s}) \tag{29}$$

In other words, the infomax principle is *equivalent* to minimizing the KL divergence between the distributions of the demixed outputs z_i and $\tilde{s_i}$. Cardoso also shows how this approach can be equated to a maximum likelihood where we are maximizing over the parameter W such that the distribution of $W^{-1}\tilde{s}$ best matches the samples of Ms (i.e., y). There is thus a clear interpretation of the $\tilde{s_i}$ as our "models" for the source inputs and the q_i as our source input distributions. If our models match our sources, then the global minimization results in finding the correct inputs. On the other hand, if our models do not match our sources, we will find the demixing matrix that makes our output components z look as close as possible in distributions to those models. The results can be dramatically wrong, as in figure 3. The authors claim that the sigmoid non-linearity works well for most source distributions. In particular, they conjecture that for distributions that are "super-gaussian" – that have a positive kurtosis (fourth order cumulant) – their approach will separate the signals correctly. We believe that they may have overly restricted in their thinking to the BSS problem of signal processing. When we consider the range of possible densities in arbitrary learning problems, it is difficult to believe that "most" densities will have this property, and even if they do that this property is sufficient to ensure proper separation via the infomax approach.

The authors are aware of the interpretation in (29), and propose the solution of having a flexible g function that can be trained to the source PDF's. In their view, that this makes the ICA into a two-stage process, where the CDF's are first estimated and the resulting g is then used to do the separation. Ideally, though, both the g function and the W matrix would be pursued jointly and automatically without having to provide a "training" signal of the sources. One possible way to approach this may be to use the maximization of mutual information criterion employed by Comon or Amari to do the original separation, which require no assumptions about the source distributions. As these approaches converge, the output statistics of z_i can be used to estimate the g_i

functions. At that point, though, unless the mixing matrix were changing over time, there would be no reason to switch to the infomax algorithm.

In summary, the Bell and Sejnowski approach is an online method that is very simple to derive and implement, but can be ineffective if our models of the source distributions are sufficiently poor. In addition, it has no guarantees for convergence to the global maximum and faces the same step size problems as the Amari approach. However, when we do have good models of the distributions, it may in some ways work better than the other methods. First of all, it would not be throwing out higher order effects through approximations as the other methods do, and as a result may be better able to separate out the independent components. Next, it would allow us not only to separate the components but have WM precisely equal the identity - in the case when the sources had different (known) statistics, we could use g_i to determine which source each component z_i would match. The scalefactor is accounted for by the g_i as well - the wrong scaling will result in non-uniform distribution for $g_i(y_i)$ and thus will not maximize $H(z)$. A last point is to notice an interesting similarity with the contrast of Comon (10. While the two methods come from quite different directions, Comon ends up maximizing the sum of the marginal entropies (or an approximation thereof) under a restricted class of transformations (orthogonal W), while Bell and Sejnowski are maximizing the joint entropy under arbitrary invertible linear transformations and post-processed by the non-linearity g. In the case where we have matched the input distributions, maximization of the joint entropy is equivalent to maximizing the transformed marginal entropies. While these approaches are clearly not the same, it is interesting to note that both approaches in seeking independence end up maximizing marginal entropies in some sense.

3 Discussion and Extensions

The three methods we have surveyed each have their strengths and weaknesses. Comon's approach is a relatively slow batch method and is restricted in its class of transforms, but it improves its contrast function monotonically with each iteration and requires no knowledge of the source distributions (though they must have non-null fourth-order cumulants). Amari's method is mathematically appealing in that it maximizes its contrast directly in the manifold of possible solutions, yet the step size is an issue and there are no guarantees of smooth convergence, especially considering the estimate of the cumulants by their "instantaneous" values. Furthermore, the meager empirical results are questionable in terms of their convergence. Last, the Bell and Sejnowski infomax method is simple to implement and seems to work well empirically (and perhaps very well when we know the source distributions), but is known to have problems when there is a sufficient degree of mismatch between the CDF's of the source distributions and the components of the non-linearity.

In addition, there are a number of situations which none of the algorithms address. For example, given that the first two methods approximate the density of the marginals around the neighborhood of a Gaussian and the third makes claims of generality for its chosen nonlinearity, it would be interesting to see how they would perform on a multi-modal distribution (e.g., a mixture of Gaussians). If we have the luxury of training on the source distributions, It is quite conceivable that the Sejnowski approach would do the best here because it would be cutting no corners with approximations. However, in the general machine learning scenario where we would be trying to break up a complex joint density into indpendent factors, no such luxury would be available. Second, it would be interesting to see how the approaches compare against the simple PCA approaches. While it is obvious that PCA can only find decorrelated components, it is not obvious how much worse it would do than the methods surveyed here. Furthermore, it would be instructive to compare the results of all three algorithms and PCA (and others in the field) on the same set of data.

In the remainder of this section, we propose a number of novel extensions that stem from the ICA concepts we have discussed. While we have not developed them in any great detail, they may be interesting diretions to pursue in future research.

3.1 Correlated Components Analysis

One thing that none of the above approaches take into account is the time-dependence of the data. In the general data case, of course, we cannot rely on this, but it could certainly be useful for separating audio components or any other form of relatively continuous input. For such signals, we propose seeking out the signals that have the *lowest* conditional entropy for $p(z_i[k+1]|z_i[k])$ as well as the lowest mutual information. We would thus be minimizing mutual information across components and minimizing entropy across time. Our contrast would then be of the form

$$\Phi(W) = -\alpha I(z) - \beta \sum_{i=1}^{N} H(z_i[k+1]|z_i[k]) \tag{30}$$

As we have seen already, the second term amounts to maximizing the KL divergence between the marginal entropies and uniform distribution. If we can model the dynamics of our source statistics $p(z_i[k+1]|z_i[k])$ with some model distribution $q(z_i[k+1]|z_i[k])$, we can incorporate this into our contrast:

$$\Phi(W) = -\alpha I(z) + \beta \sum_{i=1}^{N} KL(p(z_i[k+1]|z_i[k])||q(z_i[k+1]|z_i[k])) \tag{31}$$

By using both the time dependency of the signal and the mutual information metric, it would follow that we could reduce the number of samples we need to accurately estimate W, since we are effectively using more information from each sample.

3.2 Conditionally Independent Components Analysis

In the context of pattern recognition, we often want to estimate one quantity y on the basis of many features x. For example, we may want to estimate a person's height based on their shoe size, their hair color, their race, their eye color, etc. The typical approach to this problem is to model the entire joint density, which of course requires a very large number of samples. Recent approaches [13, 11] have proposed reducing the number of samples (and modeling components) necessary by modeling the conditional density explicitly using (respectively) decision trees and mixtures of Gaussians. We propose instead to find a factorial form for the conditional density. The quantity we are trying to model is $p(y|x_1, x_2, \ldots, x_N)$. We can rewrite this through Bayes' rule as

$$p(y|x_1, x_2, \ldots, x_N) = \frac{p(x_1, x_2, \ldots, x_N|y)p(y)}{p(x_1, x_2, \ldots, x_N)} \tag{32}$$

In general, we are not concerned with the denominator, since we are estimating y with MAP and thus can simply pick the maximum of this density. If we can find a transform $z = Wx$ that splits it into independent components, we would have:

$$p(y|x_1, x_2, \ldots, x_N) \propto \prod p(z_i|y)p(y) \tag{33}$$

This factorial form allows us to estimate only N two-dimensional marginals (assuming y is one-dimensional) rather than an N dimensional joint. Furthermore, we can use the mutual information between z_i and y to prune components that are not useful, since $I(y, z_i) = KL(p(y|z_i)||p(z_i))$ as shown in the introduction. The components that are not good predictors of y will have low mutual information and can be discarded.

Making use of this method will require some additional work beyond the formulations we have discussed, since now we are factoring two-dimensional densities and all our approximations, cumulants, etc. must be generalized to account for this. However, if it is not computationally overpowering, this method could prove very useful for reducing the amount of data necessary for a variety of learning tasks.

3.3 ICA-based clustering

A related problem is using ICA for clustering, i.e., finding the clusters that best factor into independent components. The basic idea is simple: imagine we have data distributed as shown in figure 5. We then want to choose membership for a given number of clusters j such that we minimize the sum of the mutual information over all clusters. Our contrast for this problem is thus:

$$\Phi(\theta, W_1 \ldots W_N) = -\sum_{j=1}^{N} I(W_j y|j) \tag{34}$$

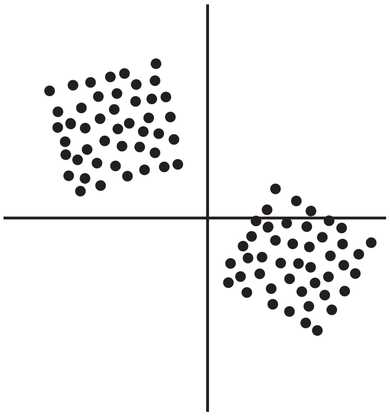

Figure 5: A distribution of data that can be broken into factorizable clusters

where θ represents the parameters that determine the membership in the clusters. Obvious choices for this would include parameters for density models such that data would be assigned to the MAP cluster in a greedy way. With simple enough models, perhaps we can find computationally reasonable gradients for maximizing this contrast.

Another way of looking at this problem is that we have a slowly varying *context signal* y, an appropriate binning of which results in Wx being mutually independent for some W. In the above description, the y would be the value j (which cluster a given point belongs to). The relation to the conditionally independent components analysis proposed above is clear here.

3.4 Phase-Aligned ICA

This is a simple idea for cases in the audio domain in which we have both propagation delays between sensors and more sources than sensors, so that we cannot use the combination of blind source separation and blind deconvolution described in the introduction. We propose to cross-correlate the observations to find a particular source, phase align the signals against that source, and then perform ICA on the resulting signals. Under various W's, the other sources in each signal will appear as sums of delayed versions of themselves, since none of the other sources will be aligned. Since we are only demixing the signals with a matrix, we will not be able to completely isolate the given source. However, we can do better than the typical array approach in which all of the phase-aligned components are simply summed together. Here, we will be weighting (through the row of W corresponding to the desired output) each observation channel with what we hope is related to the amount of information contributed from that channel. We can perhaps deliver a cleaner signal by then finding the best local autoregressive (LPC) model for the signal using an error metric involving all of the observation channels, but weighted again by the appropriate row of W.

4 Conclusions

We have described the Independent Components Analysis problem, surveyed three prominent techniques for computing the independent channels, and commented on their strengths and weaknesses. Overall, it appears that Comon's method is the most dependable, with guaranteed monotonic increases in the contrast function on every iteration. Amari provides a nicely formulated online approach based on the natural gradient, but it is not clear how well his algorithm will converge. Bell and Sejnowski's infomax method perhaps claims more than is warranted, but works well empirically on some distributions and can work very well if we can model the source distributions accurately.

In addition, we have suggested some empirical trials and described a number of interesting potential extensions/applications for ICA. The most interesting of these is probably the idea of conditionally independent components analysis, in which we propose a means for factorizing a conditional density and selecting the most relevant components for the purpose of estimating the quantity they are conditioned on. We have also sketched a method in this vein for clustering data into factorizable components.

References

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

[2] S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. In *NIPS'96*, volume 8, pages 752–763. MIT Press, 1996.

[3] S. Amari and S. C. Douglas. Why natural gradient. In *Proceedings of the IEEE Int'l. Conf. on Acous., Speech, and Signal Processing (ICASSP)*, pages 1213–1216, 1998.

[4] S. Amari and M. Kawanabe. Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli*, (3), 1997.

[5] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[6] Jean-François Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4(4):112–114, April 1997.

[7] Jean-François Cardoso and Pierre Comon. Independent component analysis, a survey of some algebraic methods. In *Proc. ISCAS'96*, volume 2, pages 93–96, 1996.

[8] P. Comon. Analyse en composantes independantes et identification aveugle. *Traitement du Signal*, 7(5):435–350, December 1990.

[9] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley, 1991.

[11] T. Jebara and A. Pentland. Maximum conditional likelihood via bound maximization and the cem algorithm. In *Neural Information Processing Systems 11*, 1998.

[12] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, 1984.

[13] Kris Popat. *Conjoint Probabilistic Subband Modeling.* PhD thesis, MIT, 1997.

[14] D. L. Wallace. Asymptotic approximations to distributions. *Ann. Math. Statist.*, 29:635–654, 1958.

[15] E. W. Weisstein. Eric weisstein's world of mathematics. http://mathworld.wolfram.com/, 2000.