Abstract

The independent component analysis (ICA) data model, in its simplest form, assumes that $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a random vector where each component is a linear combination of $n$ statistically independent, non-normal random variables or source signals. So $\mathbf{X} = \mathbf{A}\mathbf{S}$ and $X_i = a_{i1}S_1 + \cdots + a_{in}S_n$ where $\mathbf{A}$ is unknown and is referred to as the "mixing" matrix, and $S_1, \ldots, S_n$ are indirectly observed and called the independent components. Independent component analysis attempts to determine an "unmixing" matrix $\mathbf{W}$ so that $\mathbf{W}\mathbf{X} = \hat{\mathbf{S}} \approx \mathbf{S}$. Along with infomax and JADE, the fastICA algorithm, developed by Aapo Hyvärinen of the University of Helsinki, is one of the most popular algorithms for performing ICA. This paper can be considered a detailed explication of this algorithm. First, the information theory concepts of entropy, negentropy, the Kullback-Leibler divergence, and mutual information are introduced. Second, an approximation of negentropy is developed using an approximative probability density function that gives a rough upper bound for the entropy of a random variable. Third, the fastICA algorithm is described. Lastly, the fastICA algorithm is applied to some digital image processing problems.
# Contents

1 Introduction 3

2 Preliminary Information Theory 4
   2.1 Shannon Entropy and Differential Entropy 4
      2.1.1 Shannon Entropy 4
      2.1.2 Differential Entropy 5
   2.2 Negentropy 7
   2.3 Mutual Information 9

3 Approximating Negentropy 11
   3.1 The Maximum Entropy Density and its Approximation 11

4 Simulation 16
   4.1 First Part of Simulation 16
   4.2 Second Part of Simulation 18

5 Hyvärinen’s fastICA Algorithm for Maximizing Negentropy 20

6 Applications 25
   6.1 Signal Separation 26
   6.2 Separating Two Digitally Mixed Natural Images 26
   6.3 Separating Five Digitally Mixed Natural Images 27
   6.4 Separating Reflections 29
   6.5 Possible Application to a Road Safety Device Currently in Development 30

7 References 33
1 Introduction

A common statistical analysis problem is to transform observed data in order to infer some information about latent variables hidden in the observation that are difficult or impossible to observe directly. In its simplest form, the independent component analysis (ICA) data model [2] assumes that \( X = (X_1, \ldots, X_n) \in \mathbb{R}^n \) is a random vector where each component is a linear combination of \( n \) statistically independent, non-normal random variables or source signals. So \( X = AS \) and \( X_i = a_{i1}S_1 + \cdots + a_{in}S_n \) where \( A \) is unknown and is referred to as the "mixing" matrix, and \( S_1, \ldots, S_n \) are indirectly observed and called the independent components. Independent component analysis attempts to determine an "unmixing" matrix \( W \) so that \( WX = \hat{S} \approx S \).

A rigorous justification of the non-normality assumption is not given in this paper, but some intuitive reasoning is offered. The normal distribution is the most random or unstructured real-valued probability distribution (this property is shown later), and one might expect the source signals to have significant structure. In fact, the ICA method described in this paper breaks down to maximizing a measure of non-normality or structuredness of the transformed vector \( \hat{S}_i = w_i^T X \) called negentropy.

The problem the technique derived in this paper attempts to solve is easily stated: Transform an observed random vector, \( X \in \mathbb{R}^n \) by a matrix \( W = [w_1, \cdots, w_n]^T \) (i.e. \( \hat{S} = WX \)) so that the mutual independence of the components of the transformed vector \( \hat{S} = (\hat{S}_1, \ldots, \hat{S}_n)^T = (w_1^T X, \ldots, w_n^T X)^T \) is maximized. There is more than one way to do ICA, depending on how the mutual independence of the components of the transformed vector \( \hat{S} = WS \) is measured. In this paper, the measure of independence is based on a few key concepts in information theory including entropy, negentropy, and the Kullback-Leibler divergence. Once a measure of mutual independence is defined, ICA becomes an optimization problem, where this measure must be maximized.
2 Preliminary Information Theory

2.1 Shannon Entropy and Differential Entropy

2.1.1 Shannon Entropy

Shannon entropy, named after the father of information theory, Claude Shannon, is a measure of information, surprisal, or randomness associated with a discrete random variable. The Shannon entropy for a discrete random variable $Y$ with possible values $\{y_1, y_2, ..., y_n\}$ is

$$H(Y) = E[-\log(P(Y))] = -\sum_{i=1}^{n} P(Y = y_i) \log(P(Y = y_i))$$

The average amount of information that is gained when a discrete random variable takes on a value is the Shannon entropy. The greater the uncertainty is that the random variable takes on one of its possible values, the greater the Shannon entropy. If there is no randomness or uncertainty associated with a random variable, that is, if a random variable will always take on a certain value with a probability of 1, then the Shannon entropy is zero.

To illustrate these ideas, consider a coin toss. Let $Y$ represent the outcome of the toss and $p$ represent the probability that the coin lands heads up. The figure below shows how the Shannon entropy changes as a function of $p$.

![Figure 1: The Shannon entropy of a coin toss as a function of the probability of observing a heads](image_url)
Here, Shannon entropy is measured in units called *bits* since the logarithm function is base two. Base 2 is appropriate here since the response is binary. Sometimes it is more convenient to use other bases, like $e$, where the units are *nats*, or 10, where the units are *dits*. The Shannon entropy of $Y$ is maximized when $P(Y = \text{heads}) = 0.5$. This is when the outcome of the coin toss is most uncertain or most random. As it becomes more sure that the coin will land heads up (or conversely, more surely tails up), the Shannon entropy nears zero.

2.1.2 Differential Entropy

Shannon entropy is only defined for discrete random variables. To define a measure of average surprisal/information gain of a continuous random variable, some concessions must be made. This can be seen in the derivation of differential entropy [3], the extension of Shannon entropy defined for continuous random variables. Differential entropy is defined as

$$h(Y) = -\int f(y) \log(f(y)) dy.$$ 

**Derivation of Differential Entropy**

Let $Y$ be a continuous random variable with probability density function $f$. Without loss of generality, let the support set of $Y$ be $(0, \infty)$. Divide the support set into sub-intervals of length $\delta > 0$ and define $Y_i^*$ for $i = 1, 2, 3, \ldots$ such that $\int_{\delta(i-1)}^{i\delta} f(y) = f(y_i^*)\delta$. The mean value theorem guarantees that such a $Y_i^*$ exists. Let $Y^* = y_i^*$ when $\delta(i - 1) < Y \leq i\delta$. Note that for any given $\delta > 0$

$$\sum_{i=1}^{\infty} f(y_i^*)\delta = \int_{0}^{\infty} f(y) dy = 1,$$

and $Y^*$ is a discrete random variable.
Figure 2: Breaking the area under the p.d.f. into bins of size $\delta$

Calculate limit of the Shannon entropy of $Y^*$ as $\delta \to 0$:

$$
\lim_{\delta \to 0} H(Y^*) = \lim_{\delta \to 0} \sum_{i=1}^{\infty} f(y_i^*) \delta \log(f(y_i^*)) - \lim_{\delta \to 0} \sum_{i=1}^{\infty} f(y_i^*) \delta \log(\delta)
$$

$$
= - \lim_{\delta \to 0} \sum_{i=1}^{\infty} f(y_i^*) \delta \log(f(y_i^*)) - \lim_{\delta \to 0} \log(\delta)
$$

Since $\lim_{\delta \to 0} \log(\delta) = -\infty$ differential entropy must be specially defined as

$$
\lim_{\delta \to 0}(H(Y^*) + \log(\delta)) = \lim_{\delta \to 0}(- \sum_{i=1}^{\infty} f(y_i^*) \delta \log f(y_i^*))
$$

$$
= \int_{0}^{\infty} f(y) \log f(y) dy
$$

$$
= h(Y)
$$

This is the differential entropy of the continuous random variable $y$. Differential entropy does not share all the properties of Shannon entropy, like non-negativity. Also, differential entropy is not invariant under invertible linear transformations. Shown below is the differential
entropy of $\hat{s} = Wx$ where $\hat{s}, x \in \mathbb{R}^n$ and $W$ is an $n \times n$ invertible matrix.

$$h(\hat{s}) = -\int f_{\hat{s}}(\hat{s}) \log f_{\hat{s}}(\hat{s}) d\hat{s} = -\int f_x(W^{-1}\hat{s}) \frac{1}{|detW|} \log(f_x(W^{-1}\hat{s}) \frac{1}{|detW|}) d\hat{s}$$

$$= -\int f_x(x) \frac{1}{|detW|} \log(f_x(x) \frac{1}{|detW|}) |detW| dx$$

$$= -\int f_x(x) \log(f_x(x)) dx + \log |detW| = h(x) + \log |detW|$$

2.2 Negentropy

The negentropy of a random vector, $X$ with support $\mathbb{R}^n$ is defined as

$$J(X) = h(X_{gauss}) - h(X)$$

where $X_{gauss}$ is a multivariate normal random vector with the same mean and covariance matrix as $X$. (Note: Let $\varphi$ and $\phi$ represent the p.d.f.’s of $N(0,1)$ and $N_p(\mu, \Sigma)$, respectively, $h_\varphi(\cdot)$ and $h_\phi(\cdot)$ will also be used to denote the entropy of a normally distributed random variable.) Negentropy can be thought of as a measure of non-normality or degree of unstructuredness of the distribution of $x$. Negentropy is both non-negative and invariant under invertible linear transformations (i.e. $J(WX) = J(X)$), properties that prove to be useful later on.

The nonnegativity of negentropy is a result of the maximum entropy property of the multivariate normal distribution. $N_p(\mu, \Sigma)$ has the greatest entropy among all probability distributions with support $\mathbb{R}^p$, mean vector $\mu$, and covariance matrix $\Sigma$. This is shown below using the Kullback-Leibler divergence, which is defined as

$$D(f||g) = E_f[\log \frac{f(x)}{g(X)}]$$
The K-L divergence can be thought of as a measure of how close the distribution with p.d.f. $f$ is to the distribution with p.d.f. $g$, or, how much information is lost when a random variable with true p.d.f $g$ is modeled using the p.d.f $f$. In fact, negentropy is precisely the K-L divergence when $g$ is the normal p.d.f.. It is not a true measure of distance, however, since, in general, $D(f||g) \neq D(g||f)$, but it is non-negative (which can be easily proved using Jensen’s inequality) and is zero iff $f(x) = g(x)$.

**Maximum Entropy Property of $N_p(\mu, \Sigma)$ and Non-negativity of Negentropy**

Let $\phi(x)$ be the p.d.f. of $N_p(\mu, \Sigma)$ and $f(x)$ be the p.d.f. of some arbitrary distribution with support $\mathbb{R}^p$ and the same mean vector and covariance matrix. First, note that the entropy, in nats, of $N_p(\mu, \Sigma)$ is

$$h_\phi(X) = \frac{1}{2} \ln[(2\pi e)^p|\Sigma|].$$

Now use the K-L divergence to show the maximum entropy property.

$$0 \leq D(f||\phi) = \int_{\mathbb{R}^p} f(x) \ln\left(\frac{f(x)}{\phi(x)}\right)dx = -h_f(X) - \int_{\mathbb{R}^p} f(x) \ln(\phi(x))dx$$

$$= -h_f(X) + \int_{\mathbb{R}^p} f(x) \ln((2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}})dx + \int_{\mathbb{R}^p} f(x)(x - \mu)'\Sigma^{-1}(x - \mu)dx$$

$$= -h_f(X) + \ln[(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}] + \frac{p}{2}$$

$$= -h_f(X) + h_\phi(X)$$

$$\Rightarrow h_f(X) \leq h_\phi(X).$$

**Invariance Property of Negentropy Under Invertible Linear Transformations**
\[ J(\hat{S}) = J(WX) = \frac{1}{2} \ln[(2\pi e)^p|W\Sigma W'|] - h(X) - \ln(detW) \]
\[ = \frac{1}{2} \ln[(2\pi e)^p|\Sigma|] + \ln(detW) - h(X) - \ln(detW) \]
\[ = \frac{1}{2} \ln[(2\pi e)^p|\Sigma|] - h(X) = J(X) \]

### 2.3 Mutual Information

Independent component analysis requires some measure of mutual independence of the components of the vector \( \hat{S} = (\hat{S}_1, \ldots, \hat{S}_n)^T = (w_1^T X, \ldots, w_n^T X)^T \). Comon [2] offers mutual information as such a measure. We know \( \hat{S} \) has mutually independent components iff \( f_{\hat{S}}(\hat{s}) = \prod_{i=1}^{n} f_{\hat{S}_i}(\hat{s}_i) \). To determine the degree of mutual independence of the components of \( \hat{S} \) one might construct some measure of distance or similarity \( \delta(f_{\hat{S}}, \prod_{i=1}^{n} f_{\hat{S}_i}) \). Mutual information is based on this idea. It is defined as the K-L divergence \( D(f_{\hat{S}}|| \prod_{i=1}^{n} f_{\hat{S}_i}) \) and is expressed in terms of entropy.

\[
D(f_{\hat{S}}|| \prod_{i=1}^{n} f_{\hat{S}_i}) = E[\log \frac{f_{\hat{S}}(\hat{S})}{\prod_{i=1}^{n} f_{\hat{S}_i}(\hat{S}_i)}] = E[\log f_{\hat{S}}(\hat{S})] - \sum_{i=1}^{n} E[\log f_{\hat{S}_i}(\hat{S}_i)]
\]
\[ = \sum_{i=1}^{n} h(\hat{S}_i) - h(\hat{S}) = I(\hat{S}) \text{ (mutual information)} \]

Mutual information will serve as the measure of independence of the components of \( \hat{S} \). \( W \) is chosen so that \( I(\hat{S}) \) is minimized. It is convenient to express mutual information in terms
of negentropy:

\[ J(\hat{S}) - \sum_{i=1}^{n} J(\hat{S}_i) = h(\hat{S}_{gauss}) - h(\hat{S}) - \sum_{i=1}^{n} [h(\hat{S}_{gauss i}) - h(\hat{S}_i)] \]

\[ = I(\hat{S}) + h(\hat{S}_{gauss}) - \sum_{i=1}^{n} h(\hat{S}_{gauss i}) \]

\[ = I(\hat{S}) + \frac{1}{2} \ln[(2\pi e)^n|\Sigma|] - \sum_{i=1}^{n} \frac{1}{2} \ln[2\pi e \sigma_{ii}] \]

\[ = I(\hat{S}) + \frac{1}{2} \ln\left[\frac{\prod_{i=1}^{n} \sigma_{ii}}{|\Sigma|}\right] \]

\[ \Rightarrow I(\hat{S}) = J(\hat{S}) - \sum_{i=1}^{n} J(\hat{S}_i) + \frac{1}{2} \ln\left[\prod_{i=1}^{n} \sigma_{ii}\right] \]

If \( \hat{S}_1 \ldots \hat{S}_n \) are constrained to be uncorrelated, then \( \prod_{i=1}^{n} \sigma_{ii} = |\Sigma| \) and

\[ I(\hat{S}) = J(\hat{S}) - \sum_{i=1}^{n} J(\hat{S}_i) \]

and if \( W \) is constrained to be \( n \times n \) and invertible, then by the invariance property of negentropy under invertible linear transformations

\[ I(\hat{S}) = J(\hat{S}) - \sum_{i=1}^{n} J(\hat{S}_i) = J(WX) - \sum_{i=1}^{n} J(\hat{S}_i) \]

\[ = J(X) - \sum_{i=1}^{n} J(w_i^T X) \]

These constraints will be maintained through the rest of the paper. The constraint of uncorrelatedness is sensible since the true independent components \( s_1 \ldots, s_n \) are mutually independent and therefore uncorrelated.

Since \( J(X) \) does not depend on \( W \), minimizing mutual information is a problem of finding vectors \( w_1, \ldots, w_n \) so that \( \sum_{i=1}^{n} J(w_i^T X) \), a sum of one-dimensional negentropy measures, is
maximized. But because of the invariance property of negentropy, $J(w_i^T X) = J(cw_i^T X)$ for $c \in \mathbb{R}$, so each independent component can be determined only up to a multiplicative factor.

It makes sense that in order to approximate one of the independent components, $s_i$, one might try to transform the data in a way that the non-normality or structuredness is maximized. If the independent components were all very unstructured, they would behave very much like noise, and it would be difficult to distinguish from one another. Even in the case that there are only two independent components whose distributions are "close" to normal, they will most likely be indistinguishable in ICA. There are, however, methods for performing ICA where there is at most one normal or close to normal independent component [6].

3 Approximating Negentropy

Since the distribution of $w^T X$ is usually unknown in practice, negentropy cannot be calculated directly so it must be approximated. This requires developing an approximation of the entropy of $w^T X$.

Hyvärinen [5] offers a method of approximating negentropy based on approximating the "maximum entropy" probability density given by Cover and Thomas [3]. This approximation of negentropy might be considered a conservative one since it is derived using an estimation of an upper bound for entropy. This method is described in detail below.

3.1 The Maximum Entropy Density and its Approximation

This section is concerned with finding a probability density function of a random variable $X$ that has a distribution with the greatest entropy among all distributions with the same mean, variance and support set. Once the general form of the p.d.f. is determined, it is
approximated using a first-order Maclaurin series approximation. Using this p.d.f. approximation, an approximation of entropy and negentropy is developed.

Consider the problem of finding a function $f : \mathbb{R} \to \mathbb{R}$ that maximizes $h(x) = -\int_S f(x) \log(f(x))dx$ for $x \in S \subseteq \mathbb{R}$ where $f$ is subject the following constraints:

1. $f(x) \geq 0$ with equality if $x \notin S$
2. $\int_S f(x)dx = 1$
3. $\int_S f(x)G_j(x)dx = c_j$ for $j = 1, \ldots, m$

So $f$ is the p.d.f of the random variable $X$ and the expectations of some functions $G_1, \ldots, G_m$ exist. Using lagrange multipliers (in a functional analysis context), it is shown in [3] that the maximum entropy density is of the form

$$f^*(x) = \exp[\lambda_0 - 1 + \sum_{j=1}^{m} \lambda_j G_j(x)],$$

where $\lambda_0, \lambda_1, \ldots, \lambda_m$ are chosen so that the above constraints are satisfied.

The K-L divergence can be used to verify that this is the unique form of the p.d.f. of the distribution with maximum entropy among all distributions on the same support set, with the same mean and variance. Say the function $g$ satisfies the above constraints. Then

$$h_g(X) = -\int_S g(x) \ln(g(x))dx = -\int_S g(x) \ln(f^*(x)) \frac{g(x)}{f^*(x)}dx$$

$$= -D(g||f^*) - \int_S g(x) \ln(f^*(x))dx$$

$$\leq -\int_S g(x) \ln(f^*(x))dx = -\int_S g(x)(\lambda_0 + \sum_{j=1}^{m} \lambda_j G_j(x))dx$$

$$= -\int_S f^*(x)(\lambda_0 + \sum_{j=1}^{m} \lambda_j G_j(x))dx = -\int_S f^*(x) \ln(f^*(x))dx$$

$$= h_{f^*}(X) \Rightarrow h_g(X) \leq h_{f^*}(X).$$
What is left to do is solve for the constraints $\lambda_0, \ldots, \lambda_m$ given the following system of equations.

$$
\begin{align*}
\int_S f^*(x)G_j(x)dx &= c_j \text{ for } j = 1, \ldots, m \\
\int_S f^*(x)dx &= 1.
\end{align*}
$$

In order to make this system easier to solve, Hyvärinen adds the functions $G_{m+1} = x$ and $G_{m+2} = x^2$ with corresponding $c_{m+1} = 0$ and $c_{m+2} = 1$. This amounts to constraining $x$ to have zero mean and unit variance.

$$
\begin{align*}
\int_S f^*(x)xdx &= 0 \\
\int_S f^*(x)x^2dx &= 1.
\end{align*}
$$

Also, the functions $G_1, \ldots, G_m$ are constrained to form an orthonormal system of functions according to the inner product (1) and be orthogonal to polynomials of degree $\leq 2$.

$$
\langle G_i, G_j \rangle = \int_S \varphi(x)G_i(x)G_j(x)dx = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{1}
$$

$$
\int_S \varphi(x)G_j(x)x^kdx = 0 \text{ for } k = 0, 1, 2, \tag{2}
$$

where $\varphi(x)$ is the p.d.f. of the standard normal distribution. Also, for $S = \mathbb{R}$, the maximum entropy density $f^*$ is assumed to be close to $\varphi(x)$, since the normal distribution is a maximum entropy distribution. So if $f^*$ is written as

$$
f^*(x) = \exp[\lambda_0 - 1 + \lambda_{m+1}x + \lambda_{m+2}x^2 + \sum_{j=1}^{m} \lambda_jG_j(x)]
$$

$$
= A \exp[-\frac{x^2}{2} + \lambda_{m+1}x + (\lambda_{m+2} + \frac{1}{2})x^2 + \sum_{j=1}^{m} \lambda_jG_j(x)], \text{ where } A = \exp[\lambda_0 - 1]
$$

$\lambda_1, \ldots, \lambda_{m+1}$ should be very small in relation to $\lambda_{m+2} \approx -\frac{1}{2}$. Using the first-order Taylor series approximation $e^\epsilon \approx 1 + \epsilon$ where $0 < \epsilon << 1$ to approximate
exp[\lambda_{m+1}x + (\lambda_{m+2} + \frac{1}{2})x^2 + \sum_{j=1}^{m} \lambda_j G_j(x)], f^* can be written as

\[ f^*(x) = \tilde{A} \varphi(x)(1 + \lambda_{m+1}x + (\lambda_{m+2} + \frac{1}{2})x^2 + \sum_{j=1}^{m} \lambda_j G_j(x)) \]

where \( \tilde{A} = A\sqrt{2\pi} \). Finally the non-linear system above can be solved with relative ease.

\[
\begin{align*}
\int_{\mathbb{R}} f_*(x) dx &= \tilde{A}(1 + (\lambda_{m+2} + \frac{1}{2})) = 1 \\
\int_{\mathbb{R}} f^*(x) dx &= \tilde{A} \lambda_{m+1} = 0 \\
\int_{\mathbb{R}} f^*(x) x^2 dx &= \tilde{A}(1 + 3(\lambda_{m+2} + \frac{1}{2})) = 1 \\
\int_{\mathbb{R}} f^*(x) G_i(x) dx &= \tilde{A} \lambda_{j} = c_j \text{ for } j = 1, \ldots, m
\end{align*}
\]

This system yields \( \tilde{A} = 1, \lambda_{m+1} = 0, \lambda_{m+2} = \frac{1}{2}, \) and \( \lambda_{j} = c_j = E[G_j(X)] \) for \( j = 1, \ldots, m \).

**Approximating Entropy**

Now the approximative maximum entropy density

\[ \hat{f}^*(x) = \varphi(x)(1 + \sum_{j=1}^{m} c_j G_j(x)) \]

can be used to develop an approximation for entropy and, in turn, negentropy.

\[
\begin{align*}
h_{\hat{f}^*}(X) &= -\int_{\mathbb{R}} \hat{f}^*(x) \ln(\hat{f}^*(x)) dx = -\int_{\mathbb{R}} \varphi(x)(1 + \sum_{j=1}^{m} c_j G_j(x)) [\ln(\varphi(x)) + \ln(1 + \sum_{j=1}^{m} c_j G_j(x))] dx \\
&= h_{\varphi}(X) - \int_{\mathbb{R}} \varphi(x) \sum_{j=1}^{m} c_j G_j(x) \ln(\varphi(x)) dx - \int_{\mathbb{R}} \varphi(x)[(1 + \sum_{j=1}^{m} c_j G_j(x)) \ln(1 + \sum_{j=1}^{m} c_j G_j(x))] dx
\end{align*}
\]

Since \( \ln(\varphi(x)) = kx^2 \), where \( k \in \mathbb{R} \) the second term in the sum is zero by orthogonality and the third term can be approximated with the first-order Taylor series expansion.
$e \log(1 + e) \approx e + \frac{e^2}{2}$ to give

$$h_{\hat{J}}(X) = h_{\phi}(X) - 0 - \int_{\mathbb{R}} \varphi(x) \left[ \sum_{j=1}^{m} c_j G_j(x) + \frac{1}{2} \left( \sum_{j=1}^{m} c_j G_j(x) \right)^2 \right] dx$$

$$= h_{\phi}(X) - 0 - \int_{\mathbb{R}} \varphi(x) \sum_{j=1}^{m} c_j G_j(x) dx - \int_{\mathbb{R}} \varphi(x) \frac{1}{2} \left( \sum_{j=1}^{m} c_j G_j(x) \right)^2 dx$$

$$= h_{\phi}(X) - 0 - \frac{1}{2} \sum_{j=1}^{m} c_j^2 \text{ (due to the orthogonality constraints)}$$

$$= h_{\phi}(X) - \frac{1}{2} \sum_{j=1}^{m} E[G_j(x)]^2.$$ 

This approximation of entropy gives the following approximation of negentropy.

$$J(X) = h_{\phi}(X) - h(X) \approx h_{\phi}(X) - \frac{1}{2} \sum_{j=1}^{m} E[G_j(x)]^2 = \frac{1}{2} \sum_{j=1}^{m} E[G_j(x)]^2 = \hat{J}(X) \quad (3)$$

Like true negentropy, this approximation is nonnegative and zero when $X$ has a standard normal distribution since $\int_{\mathbb{R}} \varphi(x) G_j(x) dx = 0$. $\mathbf{w}_1, \ldots, \mathbf{w}_n$ will be determined so that $\hat{J}(\mathbf{w}_i^T \mathbf{X})$ is maximized and sample averages will be used to estimate the $E[G_j(\mathbf{w}_i^T \mathbf{X})]$.

**Choosing the functions $G_1, \ldots, G_m$**

Hyvärinen, Karhunen, and Oja [6] suggest the following three rules of thumb for choosing $G_1, \ldots, G_m$.

1. The expectations $E[G_j(\mathbf{w}_i^T \mathbf{X})]$ should not be very difficult to calculate. Choose simple functions over complicated ones.

2. Do not choose functions that grow faster than quadratically as a function of $|x|$.

3. If the exact distribution of $\mathbf{w}_i^T \mathbf{X}$ were known, the exact entropy could be calculated without using (3). Since the density is, in practice, almost always unknown, the log-densities of some well-known or important distributions might be used in (3).
4 Simulation

In this simulation we will compare the approximation of negentropy derived above with true negentropy measurements for the family of mixture distributions

\[ X \sim w_1 N(0, 1) + w_2 \text{LOG}(1, \frac{1}{5}) + w_3 \text{LAP}(\frac{3}{2}, \frac{2}{5}) \]  
where \( w_1 + w_2 + w_3 = 1 \) \( (4) \)

where \( \text{LOG}(1, \frac{1}{5}) \) is the logistic distribution with location and scale parameters 1 and \( \frac{1}{5} \), respectively, and \( \text{LAP}(\frac{3}{2}, \frac{2}{5}) \) is the Laplace (or double exponential) distribution with location and scale parameters \( \frac{3}{2} \) and \( \frac{2}{5} \).

4.1 First Part of Simulation

Let \( f(x) \) be the p.d.f. of the transformed mixture distribution above so that \( x \) has zero mean and unit variance. First we will use only one function, \( G(x) = \log(f(x)) \), the log-density of the mixture distribution. This choice was inspired by Hyvärinen’s third rule of thumb for choosing the functions used in the negentropy approximation. We must ensure that \( G \) is normalized and orthogonal to polynomials up to degree 2 according to (1). To achieve this, we can use Graham-Schmidt orthonormalization. Let \( G_0(x) = 1 \), \( G_1(x) = x \), and \( G_2(x) = x^2 \).

\[
G_0(x) \leftarrow \frac{1}{\langle 1, 1 \rangle} = \frac{1}{\|1\|} = 1
\]

\[
G_1(x) \leftarrow \frac{x - \text{proj}_{G_0(x)}[G_1(x)]}{\|x - \text{proj}_{G_0(x)}[G_1(x)]\|} = x
\]

\[
G_2(x) \leftarrow \frac{x^2 - \text{proj}_{G_0(x)}[G_2(x)] - \text{proj}_{G_1(x)}[G_2(x)]}{\|x - \text{proj}_{G_0(x)}[G_2(x)] - \text{proj}_{G_1(x)}[G_2(x)]\|} = \frac{x^2}{2} - \frac{1}{2}
\]

\[
G(x) \leftarrow \frac{x^2 - \text{proj}_{G_0(x)}[G(x)] - \text{proj}_{G_1(x)}[G(x)]}{\|x - \text{proj}_{G_0(x)}[G(x)] - \text{proj}_{G_1(x)}[G(x)]\|}
\]

We will compare the true negentropy of the family of mixture distributions with the negentropy approximation using the log-density function. The figure below shows the results
of the simulation.

Figure 3: Results of simulation using the log-density of the mixture distribution in the negentropy approximation: Negentropy measures are represented on the z-axis while the other two axes are \((w_1, w_2)\) space. The un-meshed surface is the true negentropy and the meshed surface shows the approximation.

The two surfaces are very close in shape and intersect the \((w_1, w_2)\) plane at \((1, 0)\) when \(X\) is normally distributed, as should be expected. The approximation overestimates the non-normality of \(X\) most noticeably for \(w_1 \approx 0.1\), with the largest overestimates occurring at the extreme values of \(w_2\) (i.e. \(\approx 0\) and \(\approx 1\)). This happens to be the set of weights for which the kurtosis and the skewness are highest. The figure below shows in \((w_1, w_2)\) space the differences in skewness and kurtosis of the mixture distributions in (4). There is a clear correspondence between the lighter region in the figure below and the region where negentropy is overestimated in figure 3.

Figure 4: Differences in skewness and kurtosis of the family of mixture distributions in (4), lighter regions correspond to larger measures (in magnitude) of skewness and kurtosis
4.2 Second Part of Simulation

Jones and Sibson [7] offer the following approximation of negentropy for a random variable of zero mean and unit variance, a function of skewness and kurtosis:

\[
J(X) \approx \frac{1}{12} E[X^3]^2 + \frac{1}{48} (E[X^4] - 3)^2 = \frac{1}{12} E[X^3]^2 + \frac{1}{48} \text{kurt}(X)^2
\]  

Hyvärinen, Karhunen, and Oja [6] take inspiration from this approximation and show that when two functions \(G_1\) and \(G_2\) are used in the the approximation in (3) such that \(G_1\) is odd and \(G_2\) is even, the approximation can be written as

\[
J(X) \approx k_1 E[F_1(X)]^2 + k_2 (E[F_2(X)] - E_\varphi[F_2(X)])^2
\]

where \(k_1\) and \(k_2\) are chosen so that the orthogonality constraints are satisfied. The odd \(G_1\) corresponds to the function \(x^3\) in (5), and the even \(G_2\) corresponds to the function \(x^4 - 3\), making (6), in a sense, a generalization of (5). In (6), \(G_1\) and \(G_2\) are chosen following the three aforementioned rules of thumb. So \(x^3\) and \(x^4 - 3\) are not admissible functions for (6) since they both grow faster than quadratically.

Figure 5 shows the results of a simulation where the approximations (5) and (6) are compared. The odd function used for (6) was \(x \exp(-x^2/2)\), and the even function was \(\exp(-x^2/2)\) -two simple, slowly growing functions. (6) is clearly a closer approximation almost everywhere.
Figure 5: Results of simulation using approximations given in (5) and (6): Negentropy measures are represented on the z-axis while the other two axes are \((w_1, w_2)\) space. The transparent surface is the kurtosis-skewness approximation, the dark surface is the even-odd approximation, and the lighter surface is true negentropy.

Notice that both approximations are farthest from the true negentropy for distributions with large kurtosis. It may be that the faster the functions used in (3) grow, the greater the overestimate for leptokurtic distributions. The figure below compares some of the functions used in the simulation with \(x^2\).

Figure 6: Comparing some of the \(G_i\) functions used in the simulation with \(x^2\). Dotted curve: log-density of the mixture distribution given in (4) with weights \((w_1, w_2, w_3) = (0.1, 0.85, 0.05)\), dashed curve: \(x \exp\left(-\frac{x^2}{2}\right)\), dot-dashed curve: \(\exp\left(-\frac{x^2}{2}\right)\), solid curve: \(x^2\)
5 Hyvärinen’s fastICA Algorithm for Maximizing Negentropy

Aapo Hyvärinen’s popular ICA algorithm, fastICA, determines the vectors $w_1, \ldots, w_n$ that maximize the negentropy approximation in (3) evaluated at $w_i^T x$. In this section, a general derivation of the fastICA algorithm is given under the assumptions that only one function is used in (3), $x \in \mathbb{R}^n$ is a mixture of $n$ independent source signals, the observation $x$ has mean vector $0$ and covariance matrix $I$ (i.e. the data have been centered and whitened), and the independent component estimates $\hat{s}_1, \ldots, \hat{s}_n$ are uncorrelated.
Estimating One Independent Component

To estimate one independent component, a vector \( w \) must be found so that 
\[
\hat{J}(w^T X) = \frac{1}{2} E[G(w^T X)]^2
\]
is maximized. Since \( X \) has been centered and whitened, the independent component estimates will have zero mean and unit variance. This constrains \( w \) to have unit length since 
\[
E[(w^T X)(w^T X)] = ||w||^2 = 1.
\]
Formally stated, the the optimization problem for estimating one independent component is

\[
\max_w \frac{1}{2} E[G(w^T X)]^2 \\
\text{s.t } E[(w^T X)^2] = ||w||^2 = 1.
\]

Optimal vectors, \( w^T * \) occur where

\[
\frac{d}{d w^T} \frac{1}{2} E[G(w^T X)]^2 - \lambda \frac{d}{d w^T} (||w||^2 - 1) = 0,
\]

where \( \lambda \) is the Lagrange multiplier. Calculate the first addend on the left hand side of (7), denoting \( f \) as the p.d.f of \( X \).

\[
\frac{d}{d w^T} \frac{1}{2} E[G(w^T X)]^2 = \frac{1}{2} \frac{d}{d w^T} \left( \int_{\mathbb{R}^n} G(w^T x) f(x) dx \right)^2 = \left( \frac{\partial}{\partial w_1} \int_{\mathbb{R}^n} G(w^T x) f(x) dx, \ldots, \frac{\partial}{\partial w_n} \int_{\mathbb{R}^n} G(w^T x) f(x) dx \right)^T = \left( \int_{\mathbb{R}^n} x_1 G'(w^T x) f(x) dx, \ldots, \int_{\mathbb{R}^n} x_n G'(w^T x) f(x) dx \right)^T = E[xG'(w^T X)].
\]

And now the second addend:

\[
\frac{d}{d w^T} (||w||^2 - 1) = \frac{d}{d w^T} (w_1^2 + w_2^2 + \cdots + w_n^2 - 1) = 2w\]
So (7) can now be written as (with $\lambda$ absorbing the the 2 in $2w$)

$$E[xG'(w^T X)] - \lambda w = 0. \quad (8)$$

Let $w^* T$ be the optimal vector. Solve for $\lambda$.

$$E[XG'(w^* T X)] = \lambda w^* T$$

$$w^* T E[XG'(w^* T X)] = w^* T \lambda w^* T = \lambda$$

So we have the equation

$$E[XG'(w^T X)] - E[w^* T XG'(w^* T X)] w = 0. \quad (9)$$

Hyvärinen employs the Newton-Raphson method to solve for the zeros of this vector equation. The Newton-Raphson iterative scheme for this system of $n$ equations of $n$ variables is

$$w_{n+1} \leftarrow w_n - \mathbb{J}[E[XG'(w^T X)] - \lambda w_n]^{-1}(E[XG'(w^T X)] - \lambda w_n),$$

where $\mathbb{J}(\cdot)$ is the Jacobian matrix, which is calculated below.

$$\mathbb{J}[E[XG'(w^T X)] - \lambda w] = \begin{pmatrix}
\frac{\partial}{\partial w_1} (\int x_1 G'(w^T x) f_1(x_1) dx_1 - \lambda w_1) & \ldots & \frac{\partial}{\partial w_n} (\int x_1 G'(w^T x) f_1(x_1) dx_1 - \lambda w_1) \\
\vdots & \ddots & \vdots \\
\frac{\partial}{\partial w_1} (\int x_n G'(w^T x) f_n(x_n) dx_n - \lambda w_n) & \ldots & \frac{\partial}{\partial w_n} (\int x_n G'(w^T x) f_n(x_n) dx_n - \lambda w_n) \\
\int x_1 x_1 G''(w^T x) f_1(x_1) dx_1 - \lambda & \ldots & \int x_1 x_n G''(w^T x) f_1(x_1) dx_1 - \lambda \ast 0 \\
\vdots & \ddots & \vdots \\
\int x_n x_1 G''(w^T x) f_n(x_n) dx_n - \lambda \ast 0 & \ldots & \int x_n x_n G''(w^T x) f_n(x_n) dx_n - \lambda
\end{pmatrix}

= E[XX^T G''(w^T X)] - \lambda I
To ease calculation, Hyvärinen makes the approximation

\[
E[XX^T G''(w^T x)] \approx E[XX^T] E[G''(w^T X)] = IE[G''(w^T X)].
\]

With this approximation the Jacobian matrix \((E[G''(w^T X)] - \lambda) I\) is diagonal and therefore easily invertible. Lastly, Hyvärinen replaces \(w^*\) with the current iteration \(w_n\) so \(\lambda \approx E[w^T X G''(w^T X)]\). The approximative Newton-Raphson iteration is

\[
\begin{align*}
    w_{(n+1)} &\leftarrow w_n - \frac{E[XG'(w_n^T X)] - \lambda w_n}{E[G''(w_n^T X)] - \lambda}, \\
    w_{n+1} &\leftarrow \frac{w_{(n+1)}}{||w_{(n+1)}||}.
\end{align*}
\]

By multiplying both sides of the first part of the iteration above by \(E[G''(w_n^T X)] - \lambda\), one gets

\[
\begin{align*}
    w_{(n+1)}(E[G''(w_n^T X)] - \lambda) &\leftarrow w_n E[G''(w_n^T X)] - E[XG'(w_n^T X)],
\end{align*}
\]

and because \(E[G''(w_n^T X)] + \lambda \in \mathbb{R}\), it will cancel out when \(w_{(n+1)}(E[G''(w_n^T X)] + \lambda)\) is normalized. So, finally, we get what Hyvärinen refers to as the fast fixed-point iteration for estimating one independent component.

\[
\begin{align*}
    w_{(n+1)} &\leftarrow E[XG'(w_n^T X)] - w_n E[G''(w_n^T X)] \\
    w_{n+1} &\leftarrow \frac{w_{(n+1)}}{||w_{(n+1)}||}
\end{align*}
\]

This iteration is repeated, starting with a random \(w_1\), until the left hand side of the equation given in (9) (with \(w_n\) instead of \(w^*\)) is within a predetermined tolerance of 0.
Estimating Several Independent Components

Repeating the above iteration until convergence n times to estimate n independent components may not work since the same independent component could be converged upon more than once. If the optimum vectors $w_1^*, \ldots, w_n^*$ are orthogonal, then the independent component estimates should be distinct since $w_i^*^T X = w_j^*^T X \Rightarrow \hat{I} X = w_i^* w_j^T X = 0$ and it’s reasonable to assume $X \neq 0$. In fact, $w_1^*, \ldots, w_n^*$ are constrained to be orthogonal since $\hat{S}_1, \ldots, \hat{S}_n$ are uncorrelated.

$$E[\hat{S}_i \hat{S}_j] = E[(w_i^*^T X)(w_j^*^T X)] = w_i^T w_j^* = 0$$

$\Rightarrow w_i^* w_j^*$ orthogonal

So the optimal matrix $W^*$ should be orthogonal which means that at each iteration the vectors $w_{1,(m)}, \ldots, w_{n,(m)}$ (here at the $m^{th}$ iteration) must be orthogonalized. Here, the vectors are orthogonalized by finding the nearest orthogonal matrix to $W_{(m)} = [w_{1,(m)}, \ldots, w_{n,(m)}]$. Horn [4] gives the minimization problem

$$\min_{W_{ortho}} ||W - W_{ortho}||_F^2$$

$$s.t. W_{ortho}^T W_{ortho} = I,$$

where $|| \cdot ||_F^2$ is the Frobenius norm (i.e. $||A||_F^2 = \text{trace}(A^T A)$) and $W_{ortho}$ is an orthogonal matrix. Horn shows that the solution is $W_{ortho} = W(W^T W)^{-\frac{1}{2}}$. So

$$W_m \leftarrow W_{(m)} (W_{(m)}^T W_{(m)})^{-\frac{1}{2}}.$$  

Using this orthogonalization approach, the algorithm is as follows:  

1. Starting with $n$ random vectors $w_{1,(1)}, \ldots, w_{n,(1)}$, form the orthogonal matrix

\[\text{See [6] for proof of convergence.}\]
\[ W_1 \leftarrow W(W^TW)^{-\frac{1}{2}}. \]

2. Perform the iteration for estimating one independent component on each column of
\[ W_1 = \begin{bmatrix} w_{1,1}, \cdots, w_{n,1} \end{bmatrix}. \]

\[ w_{i,(2)} \leftarrow E[XG'(w_{i,(1)}^T X)] - w_{i,(1)}E[G''(w_{i,(1)}^T X)] \]

3. Orthogonalize the matrix \( W_2 = \begin{bmatrix} w_{1,(2)}, \cdots, w_{n,(2)} \end{bmatrix}. \)

\[ W_2 \leftarrow W_2(W_2^T W_2)^{-\frac{1}{2}} \]

4. If for each column of \( W_2 \) the left hand side of the equation given in (9) (again, with \( w^* \) approximated by the column of \( W \)) is within a predetermined tolerance of 0, stop. Otherwise, repeat 2.

6 Applications

In this section, ICA is applied to a signal separation problem and then several simple digital image processing problems. The algorithm above was used via the fastICA package for R to estimate the independent components with the function \( G(s) = \log(\cosh(s)) \) used in the negentropy estimate in (3) and a stopping criterion tolerance of 0.0001.
6.1 Signal Separation

Figure 7: Original Source Signals

Figure 8: Mixture Observations: The original signals were mixed with a matrix with elements sampled from a normal distribution with mean 1 and standard deviation .45.

Figure 9: Estimated Independent Components

6.2 Separating Two Digitally Mixed Natural Images

All the digital images in these applications were processed using the EBImage package for R. The black and white images were read as matrices where the matrix element in column i and row j represents a grayscale value for the pixel in location (i, j). The mixture
images were created by two convex combinations of image data matrices, one with weights $(w_1, w_2) = (0.4, 0.6)$ and one with weights $(w_1, w_2) = (0.6, 0.4)$.

Figure 10: Original Natural Images

Figure 11: Digitally Mixed Images

Figure 12: Estimated Original Images

### 6.3 Separating Five Digitally Mixed Natural Images

5 natural images were mixed using a $5 \times 5$ matrix with random elements from a normal distribution with mean 0.8 and variance 0.25.
Figure 13: Original Natural Images

Figure 14: Mixture Images
6.4 Separating Reflections

Inspired by [1], two photographs were taken of a print of M.C. Escher’s *Band of Union* behind a pane of glass. Reflected in the glass is a photograph taken in the 1950’s of a West German border guard. A circular polarizer was used to bring out the reflection in one photograph slightly more that in the other. Using ICA, we can attempt to separate the print and the reflection.
6.5 Possible Application to a Road Safety Device Currently in Development

Dr. Jiann-Shiou Yang, professor of electrical engineering, and graduate student Rohit Sharma at the University of Minnesota Duluth are developing a device to track a vehicle’s position within a lane and alarm the inattentive driver if the vehicle begins to veer off the road. The device attaches to the windshield and has a built-in camera. It determines the vehicle’s position in the lane by tracking the white line on the side of the road.

Road tests on rainy days showed that the device malfunctions each time a windshield wiper passes through the camera’s field of view, triggering an alarm to alert the driver that the car is veering off the road. The camera’s frame rate and the speed at which the wiper passes through the field of view are slow and fast enough, respectively, to produce one frame in which the white line on the side of the road is clearly visible and a consecutive frame in which the line is partially obscured by a wiper. Because the frames are consecutive, the orientation of the white line within each frame is, or should be, nearly identical (assuming the vehicle isn’t actually veering off the road), and only the position of the wiper changes significantly.
Performing an ICA very similar to the problem of separating two digitally mixed images might be applied to this problem. Figure 18 shows two consecutive frames of a dataset provided by Dr. Yang containing 300 frames recorded during a road test on a rainy day while the windshield wipers were in use.

Figure 18: Frames 106 and 107. In frame 107, a windshield wiper is partially obscuring the view of the white line.

The results of the ICA are shown below.

Figure 19: ICA results: (a) and (b) show portions of frames 106 and 107. (c) shows the portion of frame 106 in (a) with (b) lightly superimposed (This is done since it is assumed that each observation is a mixture of the independent components. Here, one independent component is the white line and the other is the windshield wiper.) (b) and (c) serve as the two observations. (d) shows one of the estimated independent components, the white line is clearly visible. (e) shows the other estimated independent component, practically all of the interference from the wiper is in this image.
If the device could perform this sort of ICA each time it detected a problem, or perhaps continuously, then it may be able to tell if the problem is due to a wiper crossing the field of view or if the vehicle actually is veering off the road. One issue is the speed at which the ICA can be done, the triggering of a necessary alarm must not be significantly delayed by the process of determining whether or not to suppress it. Another ICA was done using another set of frames with similar results. See below.

Figure 20: Frames 184 and 185. In frame 185, a windshield wiper is partially obscuring the view of the white line.

Figure 21: ICA results: (a) and (b) show portions of frames 184 and 185. (c) shows the portion of frame 106 in (a) with (b) lightly superimposed, (b) and (c) serve as the two observations. (d) shows one of the estimated independent components, the white line is clearly visible. (e) shows the other estimated independent component, practically all of the interference from the wiper is in this image.
7 References


R Code for Signal Processing Problem

```r
>S1=rep(c(c(1:45),c(45:25),c(25:80),c(80:50),c(50:65),c(65:35))/79.5)
>S2=sin((1:1000)/20)
>S3=rep(((1:100)^2-600)/5000), 10)
>plot(S1,type="l",xlab="",ylab="")
>plot(S2,type="l",xlab="",ylab="")
>plot(S3,type="l",xlab="",ylab="")
>Amix=matrix(c(rnorm(16,1.,45)),byrow=T,ncol=4)
>X=Amix%*%matrix(c(S1,S2,S3,S4),byrow=T,ncol=1000)
>sig <- fastICA(t(X), 4, alg.typ = "parallel", fun = "logcosh", alpha = 1,method = "C", row.norm = FALSE, maxit = 200,tol = 0.00001, verbose = TRUE)
>plot(sig$S[1],type="l")
>plot(sig$S[2],type="l")
>plot(sig$S[3],type="l")
```