ICA: Cocktail Parties and Nature Scenes

Tayler Blake[1]

[1]Department of Statistics
The Ohio State University

DMSL Reading Discussion, October 27, 2010

Outline of Topics

- ICA: A Solution to the Cocktail Party Problem

Outline of Topics

- ICA: A Solution to the Cocktail Party Problem
- Why Do We Forbid Gaussian Projections?

Outline of Topics

- ICA: A Solution to the Cocktail Party Problem
- Why Do We Forbid Gaussian Projections?
- Measures of Gaussianity

Outline of Topics

- ICA: A Solution to the Cocktail Party Problem
- Why Do We Forbid Gaussian Projections?
- Measures of Gaussianity
- Solving the Optimization Problem

Outline of Topics

- ICA: A Solution to the Cocktail Party Problem
- Why Do We Forbid Gaussian Projections?
- Measures of Gaussianity
- Solving the Optimization Problem
- FastICA: An Algorithm

Outline of Topics

- ICA: A Solution to the Cocktail Party Problem
- Why Do We Forbid Gaussian Projections?
- Measures of Gaussianity
- Solving the Optimization Problem
- FastICA: An Algorithm
- Application: Identifying 'Independent Components" in Natural Scenes

The Cocktail Party Problem

The Cocktail Party Problem and a Candidate Solution

Denote the time signals recorded by each microphone by $x_1(t)$ and $x_2(t)$, which are weighted sums of the two sources signals emitted by the two speakers, $s_1(t)$ and $s_2(t)$.

$$x_1(t) = a_{11}s_1 + a_{12}s_2$$

$$x_2(t) = a_{21}s_1 + a_{22}s_2$$

where $a_{11}, a_{12}, a_{21}, a_{22}$ are parameters depending on each speakers' distance from the microphones.

Our goal:

Untangle the two speakers and identify s_1 and s_2 from only x_1 and x_2 (and without any knowledge of the a_{ij}s!)

Blind Source Separation

We observe

$$\mathrm{x} = \mathrm{As}$$

and we wish to uncover the signals $\mathbf{s}$. We seek a projection of the data,

$$\mathrm{u} = \mathrm{Wx}$$

which recovers the original signals, possibly reordered and rescaled. Clearly, if the knowledge of A were available, we just take

$$W = A^{-1}$$

and we require that the signals be assumed (not only uncorrelated, but also) independent. We further assume the s_i each have non-gaussian distributions.

Preprocessing and Assumptions

Denote

$$\mathbf{x}' = (x_1, x_2, \dots, x_n)'$$
$$\mathbf{s}' = (s_1, s_2, \dots, s_n)'$$

and assume

- $\mathbf{x}$ has been centered so that $E[\mathbf{x}] = \mathbf{0}$
- $\mathbf{x}$ has been **whitened**, or $E[\mathbf{x}\mathbf{x}'] = I$

One can achieve this property using the eigenvalue-eigenvector decomposition of $\Sigma = E[\mathbf{x}\mathbf{x}'] = UDU'$ and transforming $\mathbf{x}$, taking $\tilde{\mathbf{x}} = D^{-\frac{1}{2}}U'\mathbf{x}$

- The $\{s_i\}$ are mutually independent
- The $\{s_i\}$ have non-gaussian distributions

Principle Components Analysis and Identifiability

Why Forbid Normality?

For ICA to be possible, we must require that the independent components be non-gaussian.

Principal Components Analysis also seeks an "optimal" representation of the data, restricting solutions, W_p to orthogonal projections of the data (or $WW' =$ diagonal). Using the eigenvalue- eigenvector decomposition of Σ as before, the PCA solution is $W_p = D^{-\frac{1}{2}}U'$.

If we assume that the $\{s_i\}$ are normally distributed, then the joint distribution of the $\{x_i\}$ is determined entirely by the covariance matrix AA', and this covariance matrix is preserved if we simply replace A by AR' for any orthogonal "rotation" matrix, R. Hence, for PCA, the solution W is only attainable up to a rotation, leaving ambiguity in interpretation of the principal components.

Recovery of Signals via Non-gaussianity

We wish to recover s via some transformation of the form $\hat{s} = \mathrm{W}x$ for

$$\mathrm{W} = \begin{bmatrix} {w_1}' \\ {w_2}' \\ \dots \\ {w_n}' \end{bmatrix}$$

Take one of the rows of W, w', denote $y = w'x$ and define $z' = A'w$ so that

$$y = w'x = w'As = z's = \sum_{i=1}^{n} z_i s_i$$

Note that if w were one of the rows of A^{-1}, then $z' = w'A$ would have exactly one nonzero element. However, without knowledge of A inhibits such a wise choice of w, but the Central Limit Theorem allows us to choose a satisfactory w without being prophets.

The CLT saves the day! (again.)

By the Central Limit Theorem, $z's = \sum_{i=1}^{n} z_i s_i$ is more gaussian than just a single one of the s_i. Hence the z with only one nonzero element corresponds to a w that is one of the rows of A^{-1} and minimizes the gaussianity of $y = w'x$.

Hence, our solution W makes the "non-gaussianity" of Wx the largest!

Measuring Non-gaussianity

Several proposed measures of non-gaussianity:

- Kurtosis: the Classical measure
- Entropy and Negentropy
- Mutual Information

Kurtosis

Kurtosis is a measure of the "peakedness" of the probability distribution of a random variable.

$$
\begin{aligned}
Kurt(y) &= E\left[y^4\right] - 3E\left[y^3\right]^2 \\
&= E\left[y^4\right] - 3
\end{aligned}
$$

and for Normal random variables, this quantity is zero (and nonzero for almost all non-gaussian random variables.)

- Large positive values correspond to spiky distributions (leptokurtic)
- Large negative values correspond to flat, diffuse distributions (platykurtic)
- not robust

Negentropy

Entropy is a measure of randomness (or how unpredictable/unstructured a random variable is.)

$$\begin{aligned} H(y) &= -\int logf(y)f(y)\mathrm{d}y \\ &= E\left[\log\left(\frac{1}{\mathrm{f(y)}}\right)\right] \end{aligned}$$

and considering all random variables of equal variance, Normal random variables have the largest entropy. Define negentropy, J

$$J(y) = H\left(y_{gauss}\right) - H\left(y\right)$$

where y_{gauss} is a Normally distributed random variable with the same covariance matrix as y.

Approximations to Negentropy

Calculation of negentropy requires knowledge (estimation) of a probability density. Alternatively,

$$J(y) \approx \frac{1}{12} E\left[y^3\right]^2 + \frac{1}{48} kurt(y)^2 \quad \text{(Jones, Sibson 1987)}$$

where y is mean zero, unit variance.

- problems with robustness

$$J(y) \approx \sum_{i=1}^{p} k_i (E[G_i(y)] - E[G_i(\eta)])^2 \quad \text{(Hyvärinen, 1998b)}$$

for positive constants $\{k_i\}$ and certain choice of non-quadratic functions $\{G_i\}$ and where η is a standard Normal random variable. More simply, for p = 1,

$$J(y) \propto (E[G(y)] - E[G(\eta)])^2 \tag{1}$$

Approximations to Negentropy

The relationship in (1) holds for practically any choice of "measuring function" G, but the approximation improves with improved choice of G.

$$G_1(t) = \frac{1}{a_1}\log\cosh(a_1 t) \quad (2)$$

$$G_2(t) = -e^{-\frac{t^2}{2}} \quad (3)$$

for some constant $1 \leq a_1 \leq 2$ are typical choices.

- Kernel ICA

The Maximum Density Entropy

Assume that any knowledge, or information, we have about the density of x takes the form

$$c_i = \int f(x)\, G_i(x)\, dx \;;\; i = 1, \ldots, n$$

We call the $\{G_i\}$ **measuring functions**.

Under mild regularity conditions, the density satisfying the above conditions having maximum entropy has form

$$f_0(x) = A e^{\sum_i a_i G_i(x)}$$

Solving for A and $\{a_i\}$ requires solving

$$\begin{aligned} c_i &= \int G_i(x)\, A\, e^{\sum_i a_i G_i(x)} dx \\ 1 &= \int A\, e^{\sum_i a_i G_i(x)} dx \end{aligned}$$

The Maximum Density Entropy: Approximation

Assuming f is not far from $\phi(\cdot)$, lets approximate f_0 by adding three additional constraints:

1. $G_{n+1}(u) = u$, $c_{n+1} = 0$
2. $G_{n+2}(u) = u^2$, $c_{n+2} = 1$
3. We assume the G_i are orthonormal wrt $\phi(\cdot)$ and are orthogonal to all polynomials of degree 2.

If f is indeed near $\phi(\cdot)$, then $a_i \ll a_{n+2} = -\frac{1}{2}$ and we can approximate the **maximum entropy density** by

$$\hat{f}(x) = \phi(x)\left(1 + \sum_{i=1}^{n} c_i G_i(x)\right)$$

where $c_i = E[G_i(x)]$

Connection to Negentropy

Using a Taylor approximation to the natural log function (and some algebra), we can show that

$$\begin{aligned} H(x) &= -\int \hat{f}(x) \log \hat{f}(x)\, dx \\ &\approx H(\nu) - \frac{1}{2}\sum_{i=1}^{n} {c_i}^2 \end{aligned}$$

Hence, minimizing H(x) is equivalent to maximizing $\sum_{i=1}^{n} {c_i}^2$, and equation (1) is finally clear.

Choosing Measuring Functions

If f(x) were known, the clear choice of measuring function would be $G_{opt} = -\log f(x)$ since $-E[\log f(X)]$ gives directly the entropy, H(x).
Our considerations when choosing the $\{G_i\}$:

1. The $\{G_i\}$ satisfy the orthogonality assumptions discussed previously.
2. Estimation of $E[G_i(X)]$ must be "easy" and not too sensitive to outliers.
3. $f_0(x) = \mathrm{A}e^{\sum_i \mathrm{a}_i G_i(x)}$ must be integrable.

For (1), apply Gram-Schmidt orthonormalization to any set of n linearly independent G_i and $\{x^k\}$, k = 0,1,2
For (3) to hold, the $\{G_i\}$ should not grow faster than quadratically as a function of $|x|$ Reasonably, one might take G_i as the log density of some well-known important densities.

Mutual Information

The mutual information, I, between the components of y is given by

$$\begin{aligned} \mathrm{I}(y_1, y_2, \ldots, y_n) &= \left(\sum_{i=1}^{n} H\left(y_i\right)\right) - H\left(y\right) \\ &= \mathrm{D_{KL}}\left(f\left(y\right) || \prod_{i=1}^{n} m_i\left(y_i\right)\right) \end{aligned}$$

For invertible linear transformation W,

$$\mathrm{I}(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} H\left(y_i\right) - H\left(x\right) - \log \det \mathrm{W}$$

$$\begin{aligned} \mathrm{I} &= E\left[yy'\right] = E\left[\mathrm{W}xx'\mathrm{W}'\right] = \mathrm{W}E\left[xx'\right]\mathrm{W}' = \mathrm{W}\mathrm{W}' \\ &\Rightarrow 1 = \det\left(\mathrm{W}\ E\left[xx'\right]\ \mathrm{W}'\right) = \det \mathrm{W}\ \det \mathrm{W}' \\ &\Rightarrow \mathrm{I}\left(y\right) = C - \sum_{i=1}^{n} J\left(y_i\right) \end{aligned}$$

Maximum Likelihood

We can write the log-likelihood of y

$$\mathscr{L} = \sum_{i=1}^{n} \log f_i \left({w_i}' x \right) + \log \det |W|$$

where the $\{f_i\}$ are the pdf's of the $\{s_i\}$ (assumed here to be known), and taking expectations on both sides we obtain

$$E\left[\mathscr{L}\right] = \sum_{i=1}^{n} E\left[\log f_i \left({w_i}' x \right)\right] + \log \det |W|$$

and if the $\{f_i\}$ are identically the densities of the $\{s_i\}$, this quantity is the negative mutual information up to additive constant.

FastICA for one unit

Our solution, W*, will maximize

$$\mathrm{J}(y) = \mathrm{J}(\mathrm{Wx}) \propto \left(E\left[G(\mathrm{Wx})\right] - E\left[G(\mathrm{Wx})\right]\right)^2$$

$\Rightarrow$ W* will occur at certain optima of $E\left[G(\mathrm{Wx})\right]$ under the constraint that ${w_i}'x$ has unit variance $\forall$ i = 1,. . . , n. So, we maximize the objective function

$$E\left[G\left(w'x\right)\right] - \frac{\beta}{2}\left(w'w - 1\right)$$

and differentiating, we obtain

$$E\left[xg\left(w'x\right)\right] - \beta w = 0 \qquad (4)$$

$$\text{giving } \beta = E\left[w^{*\prime}xg\left(w^{*\prime}x\right)\right]$$

where w^* is the value of w at the optimum.

FastICA for one unit

To simplify the inversion of the Jacobian matrix for the LHS of (4), take

$$\begin{aligned} JF(w) &= E\left[xx'g'\left(w'x\right)\right] - \beta \mathrm{I} \\ &\approx E\left[xx'\right] E\left[g'\left(w'x\right)\right] - \beta \mathrm{I} \\ &= \left(E\left[g'\left(w'x\right)\right] \ - \beta\right) \mathrm{I} \end{aligned}$$

So an approximate Newton iteration is given by

$$w^{+} = w - \frac{E\left[xg\left(w'x\right)\right] - \beta w}{E\left[g'\left(w'x\right)\right] \ - \beta}$$

which can be further simplified by multiplying both sides by $\beta - E\left[g'\left(w'x\right)\right]$ to give

$$\begin{aligned} w^{+} &= E\left[xg\left(w'x\right)\right] - E\left[g'\left(w'x\right)\right] w \\ w^{+} &= \frac{w^{+}}{\parallel w^{+} \parallel} \end{aligned}$$

after initializing some value of w.

Extending the algorithm to several units

Assuming W is square:

$$\mathrm{y} = \mathrm{W}'x$$
$$\beta_i = E\left[y_i g\left(y_i\right)\right]$$
$$\mathrm{D} = \mathrm{diag}\left(\beta_i - E\left[g'\left(y_i\right)\right]\right)$$

so that we obtain

$$W^{+} = W - W\left(E\left[yg\left(y\right)\right] - \mathrm{diag}\left(\beta_i\right)\right)\mathrm{D}$$

and after each iteration, the outputs are decorrelated and normalized to unit variance. The stability of the algorithm depends heavily on this condition. ((Hyvärinen, 1999)

$$E\left[xx'g'\left(w'x\right)\right] \approx E\left[xx'\right] E\left[g'\left(w'x\right)\right]$$

is reasonable for pre-whitened data. Other gradient methods may be preferred without pre-whitening to avoid complicated matrix inversion. (Cardoso, Laheld 1996)

Extracting the Independent Components of Natural Scenes

Extracting the Independent Components of Natural Scenes

Extracting the Independent Components of Natural Scenes

Extracting the Independent Components of Natural Scenes

The Data

Each image was converted to grey scale byte values, and then n $= 17,595$ observations were randomly sampled from the these images.Each observation was a 12x12 pixel patch, hence $x_i = (x_{i1}, x_{i2}, \ldots, x_{i,144})$, i $= 1, \ldots,$ n is the vector containing the grey scale values assigned to each of the 144 pixels.

The data were centered and whitened using the filter given by

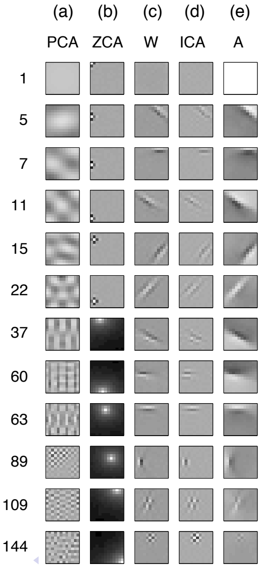
$$W_Z = \widehat{Cov\left(x\right)}^{-\frac{1}{2}}$$

and the data were transformed using the logistic measuring function:

$$G\left(y\right) = \frac{1}{1+e^{-y}}$$

...The Punchline!

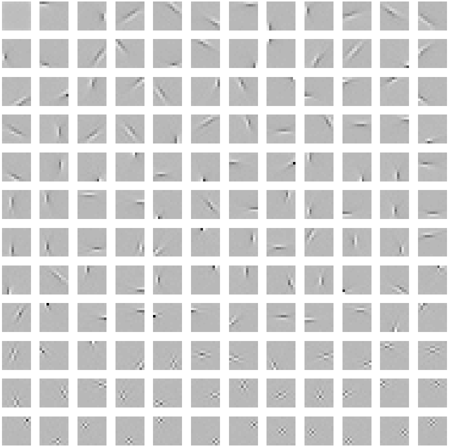
(a) The basis functions (columns of A) given by PCA (which are identical to the rows of ${W_P}^{-1}$
(b) The first 6 rows give the ZCA filters (rows of W_Z), the last 6 shows the corresponding basis functions
(c) The filters learned by ICA on the ZCA pre-whitened data
(d) The ICA filters $W_I = WW_Z$ (whitened versions of the W-filters.)
(e) The ICA basis functions (columns of ${W_I}^{-1}$)



Results

The matrix, W, of ICA filters. Each filter is a single row of W, ordered from top left to bottom right by length of the filter vectors.

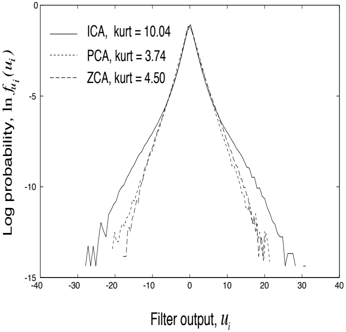
- 1 DC (low-pass) filter
- 106 oriented filters (35 diagonal, 34 horizontal, 37 vertical)
- 37 localised filters



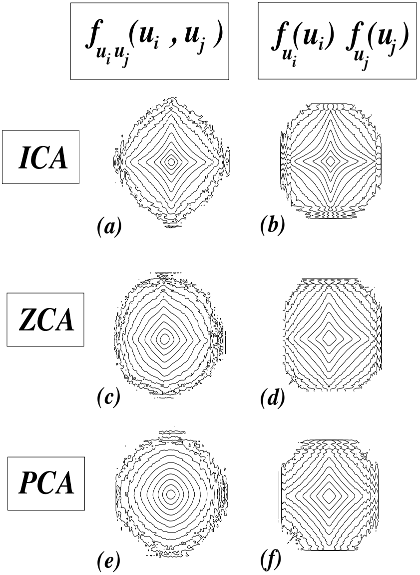
Results

The estimated log density of a fixed output component, u_i, produced by ICA, ZCA, and PCA, averaged over all filters of each type.

The sparsest signals are produced by ICA, as evidenced by the kurtosis estimates for each log histogram.



Results



The average of all bivariate distributions of pairs of output components produced by each filter and the corresponding "independent" density, the product of the marginal densities of each component. We see that the ICA filters capture best the sparseness of each univariate distribution in the joint densities.