Independent Component Analysis

Ying-Chao Hung
Department of Statistics
National Chengchi University
E-mail: hungy@nccu.edu.tw

Preface-1

- FA and ICA all assume the following basic function:

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \xi$$

- The goal of ICA and FA is to estimate the mixing matrix **A** (or its inverse) and the factors **z** based only on the data.
- In FA, it is assumed that the unobserved variables **z** are uncorrelated, while ICA makes a stronger assumption that all **z** are statistically independent (i.e. higher moments constraints).
- Remember in FA the solution of **A** is not unique (any orthogonal transformation works), assuming independence can avoid this problem.

Preface-2

- Note that in ICA, the model assumes the same number of unobserved variables as that of observed variables.

  ➔ Not used for dimension reduction !!

- ICA is a method for transforming the principal components (or factor analysis coefficients) into components which are statistically independent.

Introduction to ICA

- ICA has been used to solve Blind Source Separation (BSS) problems in signal processing.

- Example: the "cocktail party problem".

  Multiple people are all speaking simultaneously in a room. There are as many microphones as individuals in the room, each recording an audio time series signal . Each microphone will pick up a different mixture of the speakers.

  ➔The problem is to identify each speaker's audio signal individually from the mixture data.

Formulation of ICA

- The cocktail party problem is governed by the following set of linear equations:

$$
\begin{aligned}
x_1(t) &= b_{11}s_1(t) + \ldots + b_{1d}s_d(t) \\
x_2(t) &= b_{21}s_1(t) + \ldots + b_{2d}s_d(t) \\
&\vdots \qquad \vdots \\
x_d(t) &= b_{d1}s_1(t) + \ldots + b_{dd}s_d(t)
\end{aligned}
$$

where each speaker (or source) is represented by $s_j(t)$, the parameters b_{jk} represent the mixing coefficients, and the $x_j(t)$ are the mixtures.

Formulation of ICA

- The linear equations describing the true system can be represented in matrix form as:

$$\mathbf{x} = \mathbf{Bs}$$

- Drop the time index t and treat each signal $s_1,\ldots s_d$ as a random variable, the ICA model is then presented as

$$\mathbf{x} = \mathbf{Az}$$

把時間的index拿掉

where the column vector **z** are the independent components and is an estimate of **s**, and the matrix **A** is an estimate of the mixing matrix **B**.

➔ Estimate **A** and **z** based only on the data.

Model Assumptions

- The ICA model assumes the following conditions:

(i) $E(\mathbf{x}) = \mathbf{0}$

(ii) $E(\mathbf{z}) = \mathbf{0}$

(iii) $E(z_j^2) = 1, j = 1,..,m$

(iv) $p(z_1, z_2, ..., z_m) = p(z_1) \times p(z_2) \times ... \times p(z_m)$

- Condition (iii) resolves an identifiability issue of matrix **A** by fixing the variance on z_j to one.

(note that any scalar multiplier of **z** could be cancelled by dividing the same scale in **A**).

Pre-processing

- **Centering**:

  The basic processing is to center **x**, i.e., subtract its mean vector so as to make **x** a zero-mean variable.

  This implies that **s** is zero-mean as well.

- **Whitening**:

  Transform **x** linearly so that its components are uncorrelated and their variances equal unity.

  Thus, we obtain a new $\tilde{\mathbf{x}}$ such that

$$E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}'] = \mathbf{I}$$

Whitening

- The whitening procedure is always possible.
- One popular choice is to consider the eigendecomposition:

$$E[\mathbf{x}\mathbf{x}'] = \mathbf{B}\Lambda\mathbf{B}'$$

➔ Choosing $\tilde{\mathbf{x}} = \mathbf{B}\Lambda^{-1/2}\mathbf{B}'\mathbf{x}$, it is clear that $E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}'] = \mathbf{I}$.

- Whitening transforms the solution of $\mathbf{A}$ to a new one $\tilde{\mathbf{A}}$ so that

$$\tilde{\mathbf{x}} = \mathbf{B}\Lambda^{-1/2}\mathbf{B}'\mathbf{x} = \mathbf{B}\Lambda^{-1/2}\mathbf{B}'\mathbf{A}\mathbf{z} = \tilde{\mathbf{A}}\mathbf{z}.$$

- Since $E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}'] = \tilde{\mathbf{A}}E[\mathbf{z}'\mathbf{z}]\tilde{\mathbf{A}}' = \tilde{\mathbf{A}}\tilde{\mathbf{A}}' = \mathbf{I},$

➔ the solution $\tilde{\mathbf{A}}$ is orthogonal.

Why need Whitening ?

- In larger dimensions, whitening saves a lot of computations
  – since instead of estimating n^2 parameters in $\mathbf{A}$, we only have to estimate the orthogonal matrix $\tilde{\mathbf{A}}$.

  So it is a good idea to reduce the complexity of the original problem.

Estimate of Independent Components

- Let's go back to the original notation that $\mathbf{x} = \mathbf{Az}$, which implies $\mathbf{z} = \mathbf{A}^{-1}\mathbf{x}$.

- If we write one of the independent components as

$$\mathbf{z} = \mathbf{w}'\mathbf{x}$$

某一個ind解

Goal: To determine $\mathbf{w}$ so that it will equal to one of the rows in $\mathbf{A}^{-1}$?

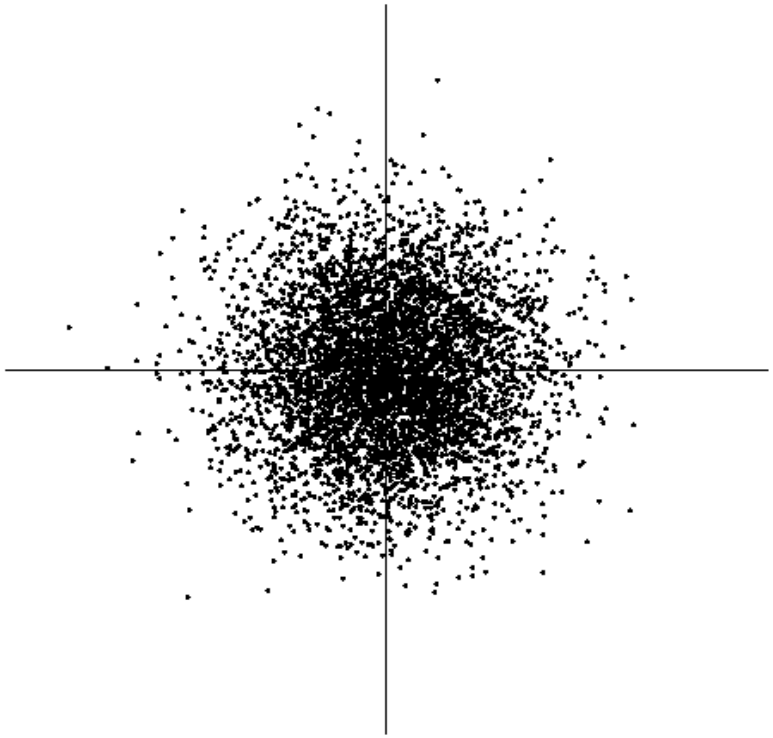
Non-Gaussian Property of **z**

- Note that by the "independence" assumption, the ICA model excludes the possibility that the solution for the **z**'s are Gaussian. (why?)
- To see why Guassian variables (**z**) make ICA impossible, assume that the mixing matrix **A** is orthogonal and z_i are Guassian. For a 2-D case, x_1 and x_2 are then Gaussian, uncorrelated, and have variance one (since **x** = **Az**).

  The joint pdf is then:

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2} \right).$$

Non-Gaussian Property of **z**



- It is easy to see that any orthogonal transformation of (x_1, x_2) has the same distribution ➔ matrix **A** is not identifiable !!

Non-Gaussian Property of **z**

- Therefore, finding the independent components is equivalent to finding the components that are uncorrelated and the furthest away from being Gaussian.

- Our goal now becomes: find **w** which maximizes the nongaussianity of $\mathbf{w}'\mathbf{x}$, thus give us independent components.

Measure of Nongaussianity

- **Kurtosis:**

  For a random variable X,

$$kurt(X) = \frac{\mu_4}{\sigma^4} - 3$$

  where $\mu_4 = E[(X - E[X])^4]$, $\sigma^4 = (Var(X))^2$.

- If X has been scaled so that it has mean 0 and variance 1, then

$$kurt(X) = E[X^4] - 3$$

- For a standard normal z ➔ $kurt(z) = 0$.

Measure of Nongaussianity

- Thus, deviation of z from normality can be measured by

$$|kurt(z)| \quad \text{or} \quad (kurt(z))^2.$$

- Kurtosis is simple to compute based on data.

- However, the kurtosis measure for sample data is sensitive to outliers.

- An alternative measure for normality is negentropy, from information theory.

Another Measure of Nongaussianity

- **Negentropy :** $J(z) = H(z_{gauss}) - H(z)$

  where $H(z)$ is the differential entropy of a random variable z , a basic quantity of information theory (Cover and Thomas, 1991).

- The **entropy** of a random variable z with density $p(z)$ is defined as:

$$H(z) = -\int p(z) \log p(z) dz$$

- $H(z)$ represents the averaged "uncertainty" ($-\log p(z)$) of the r.v. z

  ➔ The larger $H(z)$ is, the larger data size needed to get information about z (e.g. a uniform r.v. with $p(x) = 1/n$)

Entropy/Negentropy

- The more "uncertain" the variable is, the larger it's entropy.
- A Gaussian variable has the largest entropy of all random variables with equal variance.
- The measure of negentropy is quite intuitive since

$$J(z) = H(z_{gauss}) - H(z) \geq 0,$$

which measures the departure from the r.v. z and a Gaussian random variable with the same covariance (z_{gauss}).

Estimation of (Neg)Entropy

- It is clear that estimating the entropy requires an estimate of the probability density function (based on observed data).
- Due to the inherent difficulty in estimating probability densities, various approximations of negentropy are used for ICA.
- One general approximation:

$$J(z) \approx \sum_{i=1}^{p} k_i \left(E\left[g_i(z) \right] - E\left[g_i\left(z_{gauss} \right) \right] \right)^2$$

where k_i are positive constants, z is scaled to have mean 0 and variance 1, $z_{gauss} \sim N(0,1)$.

Simplified Approx. of (Neg)Entropy

- The general approximation of $J(z)$ can be further simplified by using only one term:

$$J(z) \propto \left(E[g(z)] - E[g(z_{gauss})]\right)^2$$

- The choice of g is non-quadratic so that the approx. is more robust to outliers. One popular choice that works well in practice (Hyvärinen & Oja, 2000) is:

$$g(z) = \frac{1}{c}\log\cosh(cz)$$

where c is a constant between 1 and 2.

FastICA Algorithm in R

- The FastICA (Hyvärinen & Oja, 2000) algorithm makes use of the metric

$$J(z) \propto \left(E[g(z)] - E[g(z_{gauss})]\right)^2$$

  and a fixed-point iteration scheme for estimating the independent components.

- This algorithm assumes the data has zero mean and has been whitened.

- It uses $g_1(z) = \frac{1}{c}\log\cosh(cz)$, and $g_2(z) = -\exp(-z^2/2)$.

  $(1 \le c \le 2)$

FastICA Algorithm

fastICA Algorithm:

1. Choose an initial (e.g. random) weight vector $\mathbf{w}$.
2. Let $\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$
3. Let $\mathbf{w} = \mathbf{w}^+ / \|\mathbf{w}^+\|$
4. If not converged, go back to 2.

Note: the fastICA algorithm can be downloaded from the package {fastICA} in R website.

Other Algorithms in R

- The “mlica{mlica}” in R uses the Maximum Likelihood implementation to perform ICA.
- The “ica{e1071}” in R performs ICA based on estimation of kurtosis.

Example 1. (Sources)

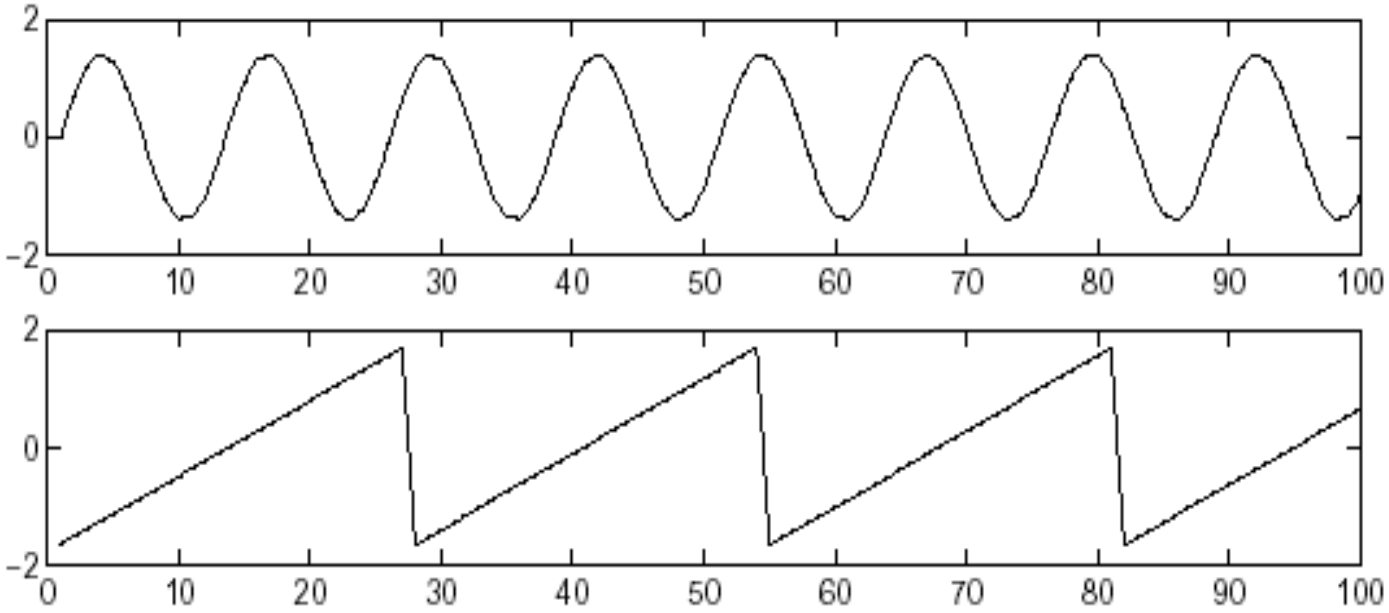


Figure 1: The original signals.

Example 1. (Mixtures)

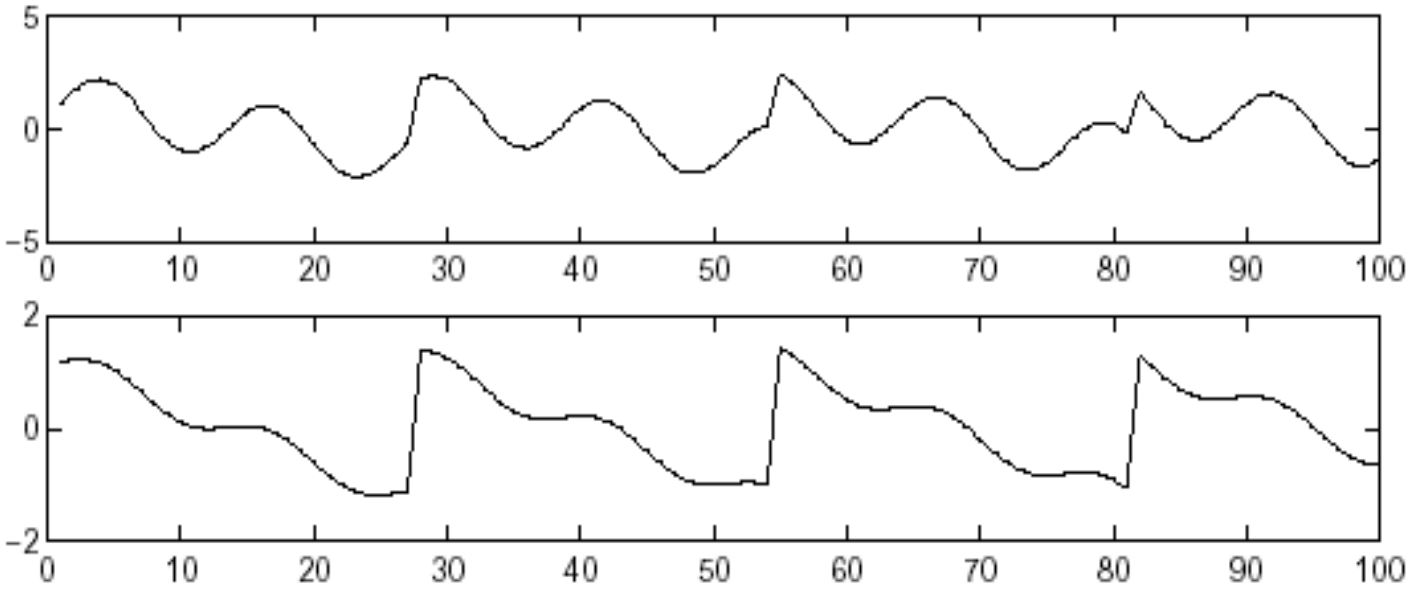


Figure 2: The observed mixtures of the source signals in Fig. 1.

Example 1. (Estimates of Sources)

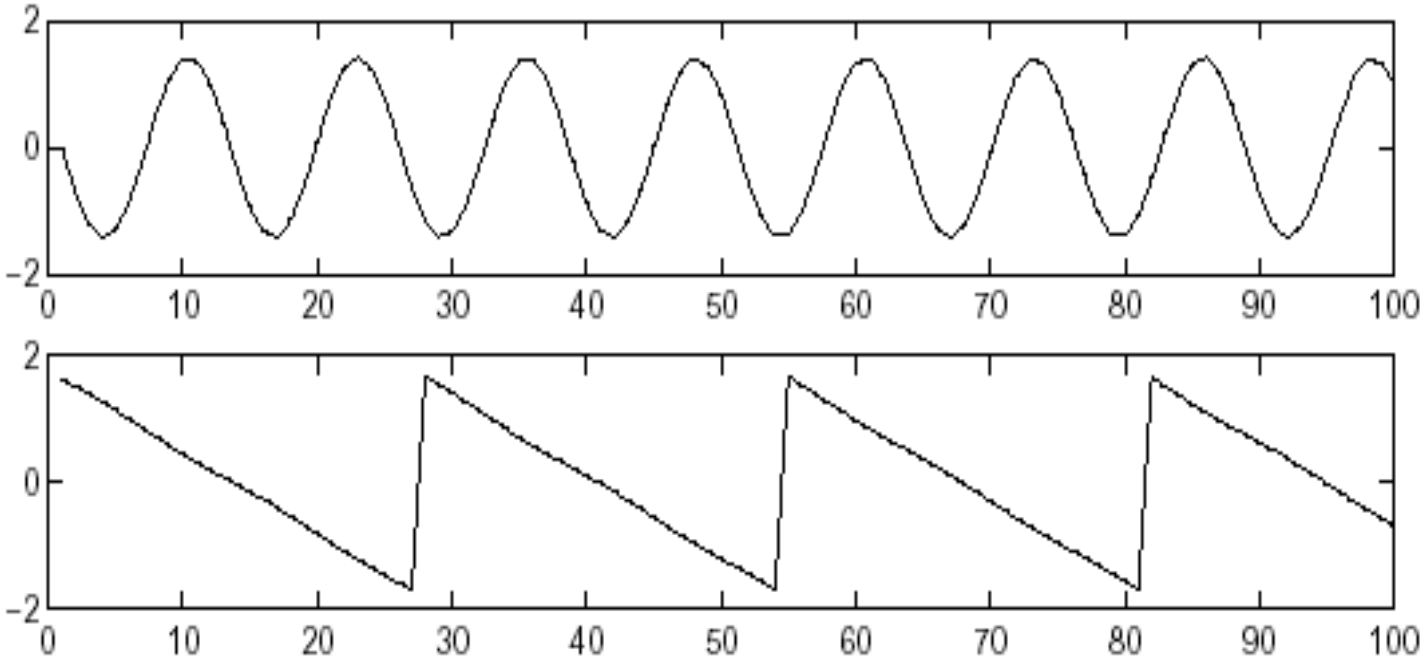


Figure 3: The estimates of the original source signals, estimated using only the observed signals in Fig. 2. The original signals were very accurately estimated, up to multiplicative signs.

Example 2. (Sources)

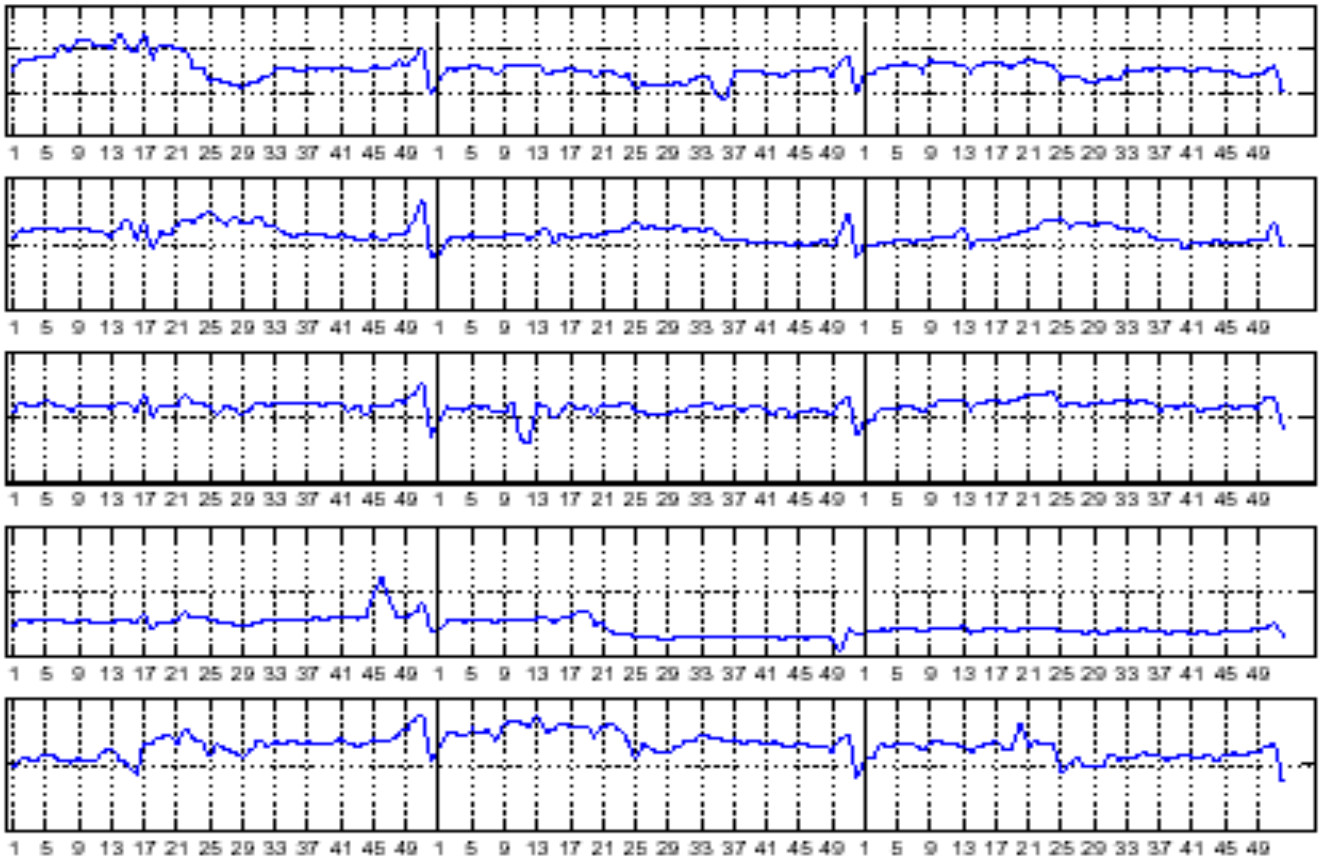


Figure 13: (from Kiviluoto and Oja, 1998). *Five samples of the original cashflow time series (mean removed, normalized to unit standard deviation). Horizontal axis: time in weeks.*

Example 2. (Estimated Independent Components)

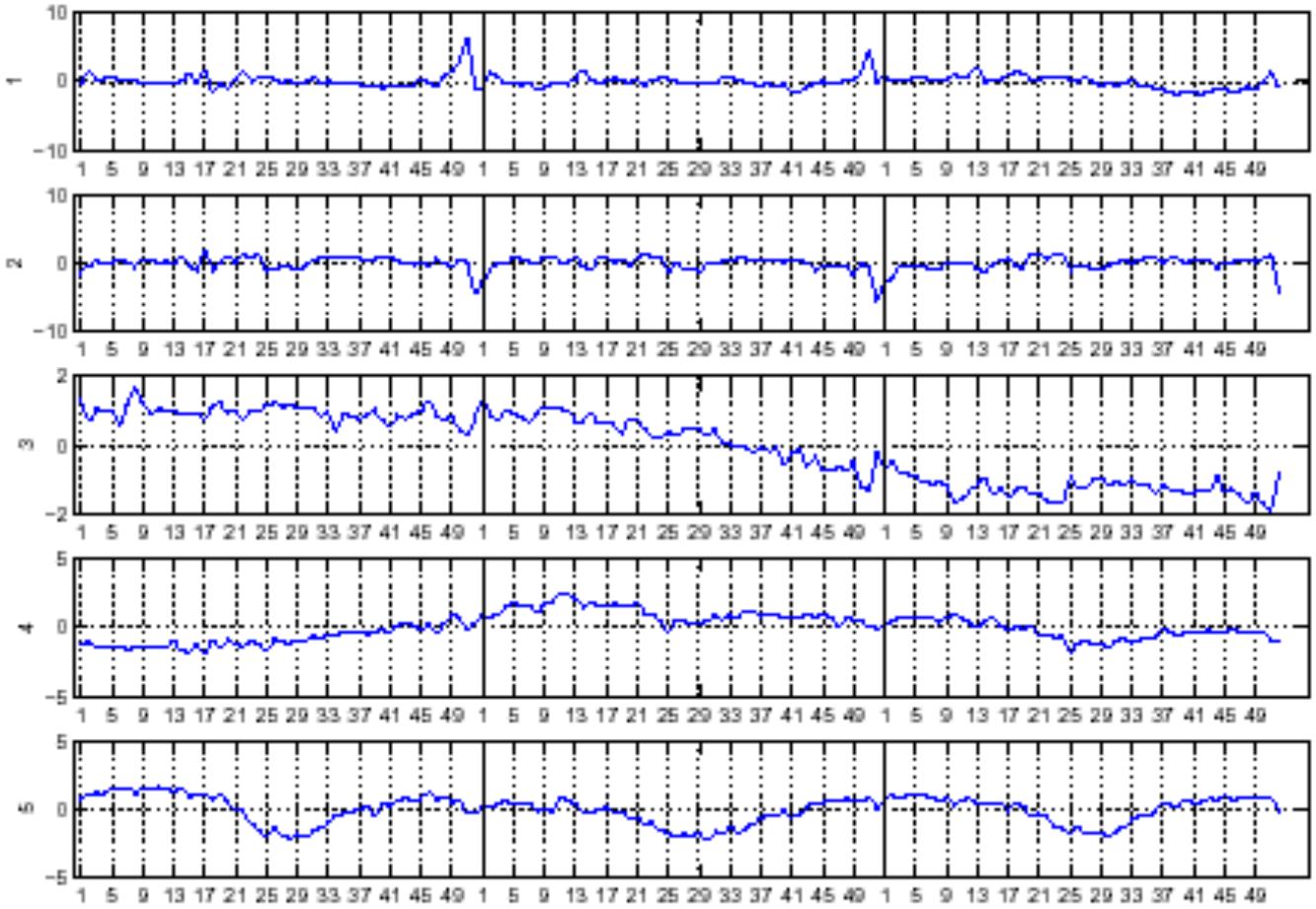


Figure 14: (from Kiviluoto and Oja, 1998). *Five independent components or fundamental factors found from the cashflow data.*

Another Application of ICA: Reducing Noise in Imagines

- Consider a noisy image model:

$$\mathbf{x} = \mathbf{z} + \mathbf{n},$$

where **z** is the vector of pixel grey levels in an image window, **n** is uncorrelated noise, and **x** is the measured image window with noise.

- Assume that **n** is Gaussian and **z** is non-Gaussian.

- **Question**: How to clean the noise?
- **Answer**: ICA is shown useful.

The Maximal Likelihood Solution

- Since we have

$$\mathbf{wx} = \mathbf{wz} + \mathbf{wn} = \mathbf{s} + \mathbf{wn},$$

and then

$$\mathbf{x} = \mathbf{w}'\mathbf{s} + \mathbf{n},$$

where **w** is the best orthogonal (orthonormal) approx. of the inverse of the ICA mixing matrix (i.e. approx. of $\mathbf{A}^{-1}$).

- The solution of **w** can be obtained by the Maximal Likelihood of ICA, thus we can recover **z** by using

$$\mathbf{z} = \mathbf{w}'\mathbf{s}.$$

Example 3. (Clean the Noise)

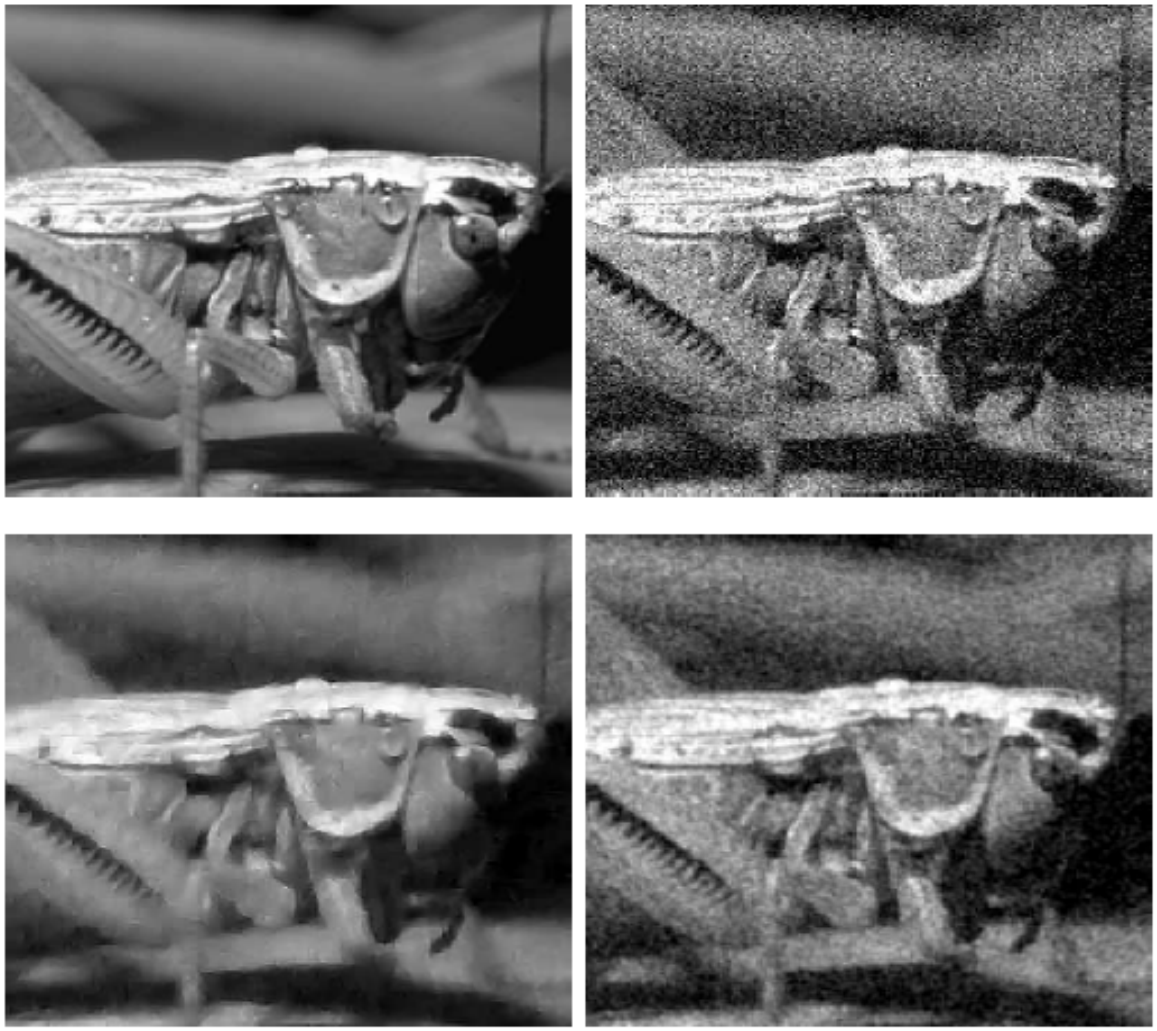


Figure 15: (from Hyvärinen, 1999d). *An experiment in denoising. Upper left: original image. Upper right: original image corrupted with noise; the noise level is 50 %. Lower left: the recovered image after applying sparse code shrinkage. Lower right: for comparison, a wiener filtered image.*

Comparison of PCA, FA, ICA.

	Model equation	**Goal**	**Distribution Assumption**	**Handling Noise**	**Equivalents**
Factor Analysis(FA)	$\mathbf{x} = \mathbf{Az} + \mathbf{u} + \xi$	Minimum Correlation between factors **z**.	Gaussian	Explicitly models noise **u** as variation unique to each input variable.	Equivalent to PCA if unique variation **u** (noise) is small.
Principal Component Analysis (PCA)	$\mathbf{x} = \mathbf{Az} + \xi$	Minimum Covariance between factors **z** (while maximizing variance).	None	Noise shows up as model error.	For Gaussian distribution, PCA provides maximum independence between factors, like ICA.
Independent Component Analysis (ICA)	$\mathbf{x} = \mathbf{Az} + \xi$	Maximum Statistical Independence between factors **z**.	Non-Gaussian	Noise shows up as model error.	A particular transformation of the PCA solution.

Over-determined BSS

- In some applications the number of sources is less than the number of sensors (observed variables in the model)

  ➔ so-called over-determined BSS problem (or non-square mixing problem)

- In this case, "extra" sensor observations are explained as observation noise and the problem can be reduced to the previously introduced determined BSS problem.
- **Solution**: Consider the probability model for non-square ICA, called Probability Independent Component Analysis (PICA).

  Book: ***Independent Component Analysis: Principles and Practice***

  by Stephen Roberts and Richard Everson, 2001, Cambridge University Press.

Under-determined BSS

- In some applications the number of sources is more than the number of sensors (observed variables in the model)

  ➔ so-called under-determined problem, usually technically more challenging !!

- **Possible Solutions:**
  - Bayesian approaches
  - Assumption of signal sparseness
  - Nonparametric maximum likelihood methods