Lectures 10 and 11:

Principal Component Analysis (PCA) and Independent Component Analysis(ICA)

Objectives: Both PCA and ICA are used to reduce measured data into a smaller set of components.

PCA - utilizes the first and second moments of the measured data, hence relying heavily on Gaussian features

ICA - exploits inherently non-Gaussian features of the data and employs higher moments

FastICA website: http://www.cis.hut.fi/projects/ica/fastica/

Principal Component Analysis, or

(PCA, Pearson,1901; Hotelling, 1933)

Proper Orthogonal Decomposition, or

(POD, Kosambi, 1943)

Karhunen(1946)-Loève(1948) Decomposition (KLD)

- It is a model reduction scheme, either for linear or mildly nonlinear problems;

- It extracts the principal eigenvector(s) that can represent trends in stock pricing, population growth, structural mode shapes, etc.

- In engineering, it was perhaps first applied to turbulence modeling by **Homes et al in 1996.**

PCA - Cont’d

Why so many names for what appears to be the same tool?

PCA - statistics, geometry, engineering

EOF(Empirical orthogonal functions) - oceanography, metrology

Factor Analysis - psychology, economics

POD - optimization, statistics, engineering

SVD (singular value decomposition) - numerical analysis

A little dig into its mathematics

Recall that PCA was developed to reduce a large
set of data to extract a reduced set of principal components.

Let us consider a nonlinear system modeled by

$$\ddot{\mathbf{y}}(t) = f(\dot{\mathbf{y}}(t), \mathbf{y}(t)), \quad \mathbf{y}(t) = [y_1(t)\ y_2(t)\ \ldots y_m(t)]^T$$

whose response matrix is given by

$$\hat{\mathbf{X}} = [\mathbf{y}(t_1)\ \mathbf{y}(t_2)\ \ldots \mathbf{y}(t_n)] = \begin{bmatrix} y_1(t_1) & y_1(t_2) & \ldots & y_1(t_n) \\ y_2(t_1) & y_2(t_2) & \ldots & y_2(t_n) \\ \ldots & \ldots & \ldots & \ldots \\ y_m(t_1) & y_m(t_2) & \ldots & y_m(t_n) \end{bmatrix}$$

PCA - Cont'd A little dig into its mathematics

If the mean values of each of the row vector are not zero, we modify the previous X to read

$$\mathbf{X} = [(\mathbf{y}(t_1) - \bar{\mathbf{y}})\ (\mathbf{y}(t_2) - \bar{\mathbf{y}})\ ...(\mathbf{y}(t_n) - \bar{\mathbf{y}})]$$

$$= \begin{bmatrix} (y_1(t_1) - \bar{y}_1) & (y_1(t_2) - \bar{y}_1) & ... & (y_1(t_n) - \bar{y}_1) \\ (y_2(t_1) - \bar{y}_2) & (y_2(t_2) - \bar{y}_2) & ... & (y_2(t_n) - \bar{y}_2) \\ ... & ... & ... & ... \\ (y_m(t_1) - \bar{y}_m) & (y_m(t_2) - \bar{y}_m) & ... & (y_1(t_n) - \bar{y}_m) \end{bmatrix}$$

where $\bar{\mathrm{y}}$ is the mean value of the response data.

A little dig into its mathematics

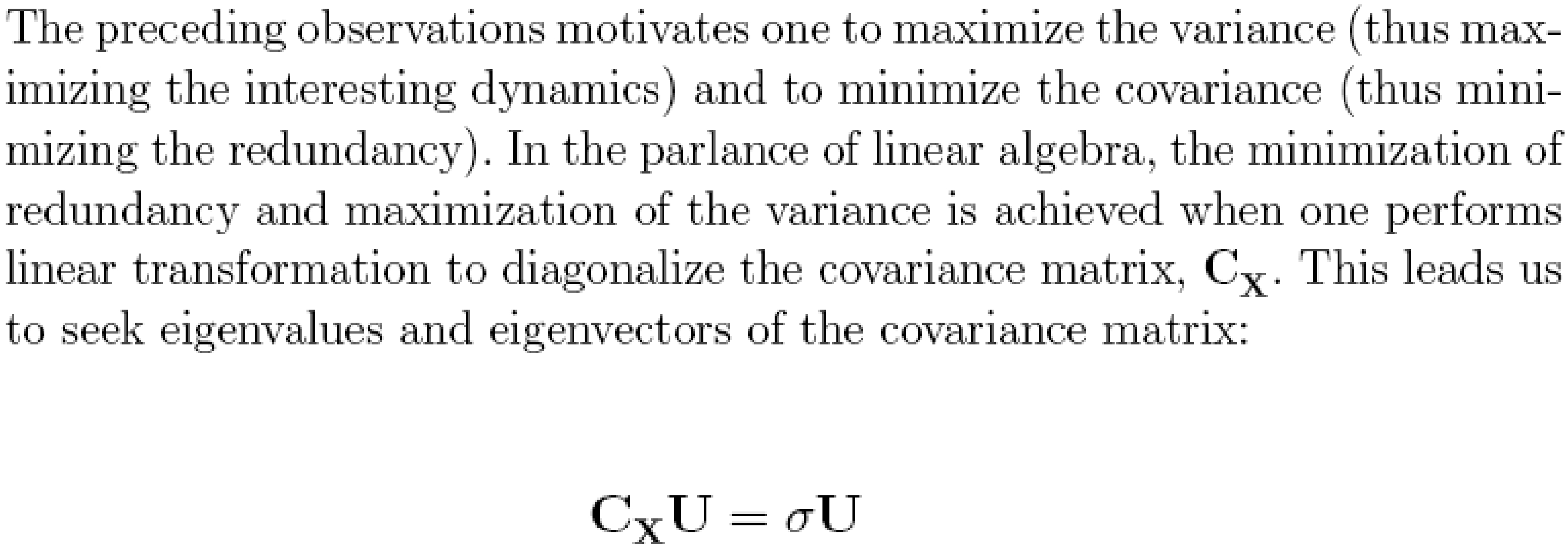
Hence, the covariance matrix of the response matrix X is given by

$$\mathrm{C}_{\mathrm{X}} = \frac{1}{n}\mathrm{X}\mathrm{X}^{T}$$

In statistics, C(m x m) captures the correlations between all possible pairs of measurements for which the diagonal and off -diagonal terms are called the variance and covariance, respectively.

More importantly, the correlation values reflect the noise and redundancy in the measured data. In the field of dynamics, large (small) values in the diagonal terms correspond to interesting dynamics (noises). On the other hand, large(small) values correspond to high(low) redundancy.

A little dig into its mathematics



The preceding observations motivates one to maximize the variance (thus maximizing the interesting dynamics) and to minimize the covariance (thus minimizing the redundancy). In the parlance of linear algebra, the minimization of redundancy and maximization of the variance is achieved when one performs linear transformation to diagonalize the covariance matrix, $\mathbf{C_X}$. This leads us to seek eigenvalues and eigenvectors of the covariance matrix:

$$\mathbf{C_X U} = \sigma \mathbf{U}$$

However, in practice, often the number of interesting dynamics is limited, which means that the rank of the covariance matrix $\mathbf{C_X}$ is small. Thus, we seek the singular values in the form:

$$\begin{aligned} \mathbf{C_X} &= \mathbf{U\Sigma U}^T \\ \mathbf{U} &= \begin{bmatrix} \mathbf{u}_1 \ \mathbf{u}_2 \ ... \ \mathbf{u}_m \end{bmatrix} \\ \mathbf{\Sigma} &= \begin{bmatrix} \sigma_1 & & & & & & & & \\ & \sigma_2 & & & & & & & \\ & & . & & & & & & \\ & & & . & & & & & \\ & & & & \sigma_{noise} & & & & \\ & & & & & 0 & & & \\ & & & & & & . & & \\ & & & & & & & . & \\ & & & & & & & & 0 \end{bmatrix}, \quad \sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{noise} \end{aligned} \tag{6}$$

For example, if there are two non-zero well separated singular values, that is, (σ_1, σ_2) and if σ_2 is considered noise, then $(\sigma_1, \mathbf{u}_1)$ are called the principal value and principal mode, respectively, and the signal-to-noise ratio (SNR) is given by

$$SNR = \frac{\sigma_1}{\sigma_2} \tag{7}$$

We now summarize some salient properties of the principal values (PVs) and principal modes (PMs).

Some Remarks on PCA

1. The principal modes(PMs) can be shown to be a linear combination of structural normal modes when the mass matrix is an identity matrix, provided one utilizes a sufficient number of samples and adequate spatial data density.

2. It has been demonstrated that the PMs can serve as the best representation of the so-called Nonlinear Normal Modes(NNMs).

3. The PVs provide a natural criterion to select the modes as they represent the energy contained in the data. In other words, the smaller the singular values, the less significant is the corresponding mode.

4. However, the principal values (PVs) are not directly related to the structural frequencies. The PVs are simply the energy of the signal measured in Frobenius norm, not necessarily in terms of physical energy, viz., kinetic and/or strain energy. This task will be discussed in the section of the Hilbert-Huang transform.

PCA-Derived NNMs to Reduce Structural EOMs End of PCA

An important application of the PMs is to transform the large-order models to the reduced-order models which can be used to correlate with experiments and to develop control strategies, among others. For example, a structural system with m-degrees of freedom equation

$$\mathbf{M}\ddot{\mathbf{y}} + \mathbf{D}\dot{\mathbf{y}} + \mathbf{K}\mathbf{y} + \mathbf{N}(\dot{\mathbf{y}}, \mathbf{y}) = \mathbf{f}, \quad dim(\mathbf{y}) = (m \times 1) \tag{8}$$

can be transformed via

$$\mathbf{y} = \sum_{j=1}^{\ell} \mathbf{u}_j \mathbf{z}_j(t), \quad \ell << m \tag{9}$$

to the following reduced-order form:

$$\begin{aligned} &\mathbf{m}\ddot{\mathbf{z}} + \mathbf{d}\dot{\mathbf{z}} + \mathbf{k}\mathbf{z} + \mathbf{U}^T\mathbf{N}(\mathbf{U}\dot{\mathbf{z}}, \mathbf{U}\mathbf{z}) = \mathbf{U}^T\mathbf{f} \\ &(\mathbf{m}, \mathbf{d}, \mathbf{k}) = \mathbf{U}^T(\mathbf{M}, \mathbf{D}, \mathbf{K})\mathbf{U} \ \ (\ell \times \ell) \end{aligned} \tag{10}$$

Independent component analysis (ICA)

PCA (Principal Component Analysis) seems to cover a wide range of data analysis; in addition, for transient data we have Hilbert-Huang Transform.

So why do we need ICA?

- PCA is based on the second-order statistics. That is, data fit Gaussian distribution (i.e., exploit correlation/ covariance properties)

- What if data cannot be characterized by the second moment? That is, $R_x = E\{xx^T\} = I$ (white noise)?

ICA- cont’d

- When do we generally need ICA?
  - when data cannot be ensembled (hence, most likely non-Gaussian via Central Limit Theorem);
  - when raw data appear to be very noisy;
  - when a sensor records several source signals simultaneously.
  - mathematically, PCA is adequate if the data are Gaussian, linear, and stationary. If not, then higher-order statistics begin to be essential.

Comparison of PCA with ICA

PCA minimizes the covariance of the data; on the other hand ICA minimizes higher-order statistics such as fourth-order cummulant (or kurtosis), thus minimizing the mutual information of the output.

Specifically, PCA yields orthogonal vectors of high energy contents in terms of the variance of the signals, whereas ICA identifies independent components for non-Gaussian signals.

ICA thus possesses two ambiguities:

First, the ICA model equation is underdetermined system; one cannot determine the variances of the independent components.

Second, one cannot rank the order of dominant components.

Why ICA has been a late bloomer

Historically there is a good reason why ICA came to be studied only recently, whereas random variables are assumed to be Gaussian in most of statistical theory;

ICA thrives on the fact that the data are non-Gaussian.

This implies that ICA exploits the loose end of the Central Limit Theorem which states that the distribution of a sum of independent random variables tends toward a Gaussian distribution. Fortunately for ICA, there are many cases where some real-world data do not have sufficient data pools that can be characterized as Gaussian.

ICA- cont'd

ICA in a nutshell

- Given measurement, $\mathbf{x}$, find the independent components, $\mathbf{s}$, and the associated mixing matrix, $\mathbf{A}$, such that

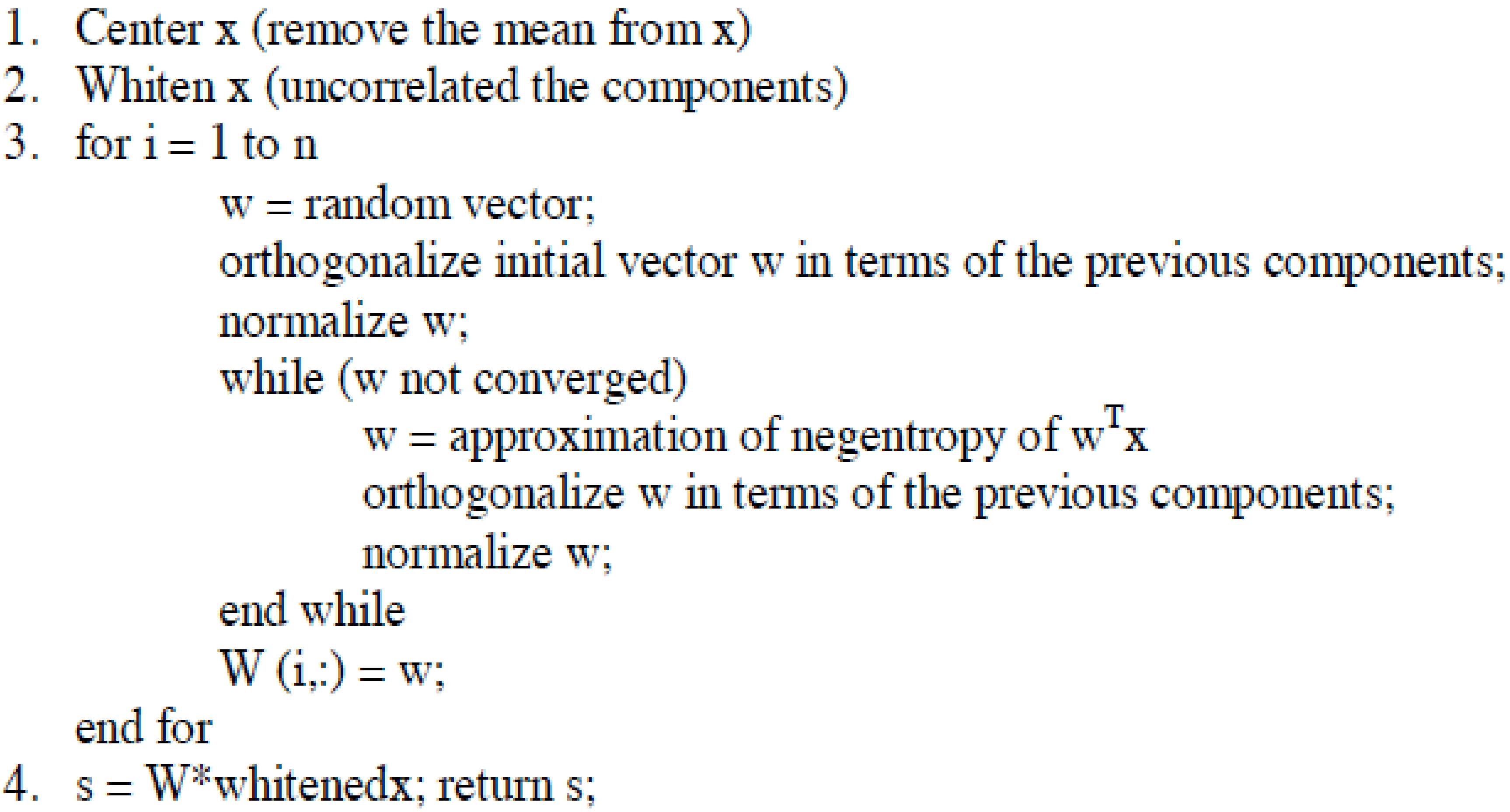
$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

- Find $\mathbf{w}_j$ that maximizes non-Gaussianity of $\mathbf{w}_j^T\mathbf{x}$

- Independent components $\mathbf{s}$ is then found from

$$\mathbf{s} = \mathbf{W}\,\mathbf{x} \qquad \text{where } \mathbf{A}^{-1} = \mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \ldots\ \mathbf{w}_m]$$

Fast ICA in words



```
1. Center x (remove the mean from x)
2. Whiten x (uncorrelated the components)
3. for i = 1 to n
           w = random vector;
           orthogonalize initial vector w in terms of the previous components;
           normalize w;
           while (w not converged)
                   w = approximation of negentropy of $w^Tx$
                   orthogonalize w in terms of the previous components;
                   normalize w;
           end while
           W (i,:) = w;
   end for
4. s = W*whitenedx; return s;
```

A Fast ICA Algorithm (using Negentropy concept)

1. Center the data, x, to make its mean zero (same as PCA):

   $$x \Leftarrow x - x_m, \quad x_m = E\{x\}$$

2. Whiten **x** to maximize non-Gaussian characteristics(PCA with filtering):

   $$z = V \Lambda^{-1/2} V^T x, \quad V \Lambda V^T = E\{xx^T\}$$

3. Choose an initial random vector, w, $\|w\| = 1$

4. Update w (maximally non-Gaussian direction!)

   $$w = E\{z * g(w^T z)\} - E\{g'(w^T z)\}w,$$

   $$g(y) = \tanh(a_1 y) \text{ or } y*\exp(-y^2/2), \quad 1 < a_1 < 2$$

   $$w = w/\|w\|$$

5. If not converged, then go back to step 4.

6. Obtain the independent component, s:

7. $s = [\, w_1 \;\; w_2 \ldots w_n \,] \;\; x$

Example of ICA (Source:Independent Component Analysis, Algorithms and Applications, in: *Neural Networks*, 13(4-5):411-430, 2000 by A. Hyvarinen and E. Oja)

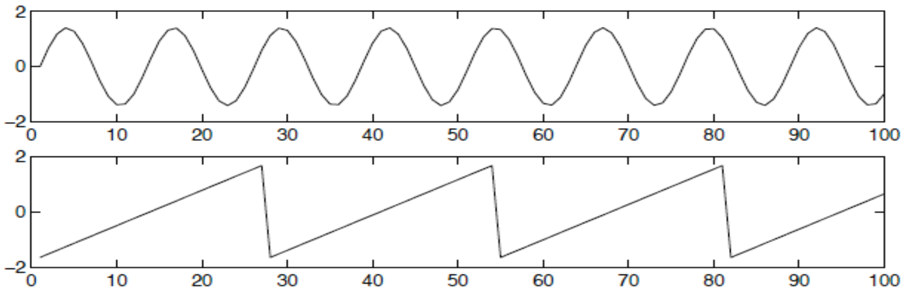


Figure 1: The original signals.

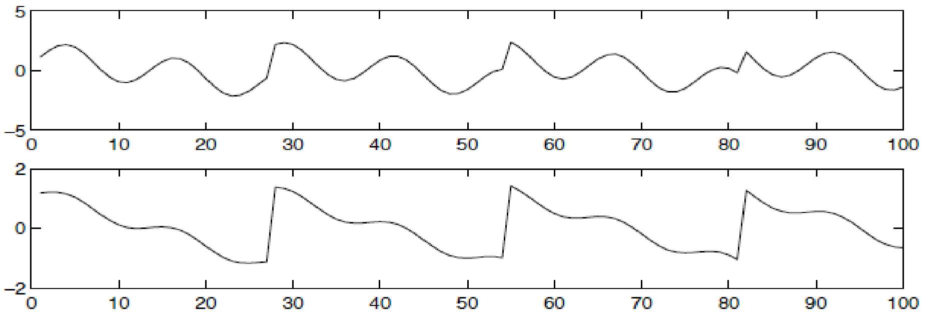


Figure 2: The observed mixtures of the source signals in Fig. 1.

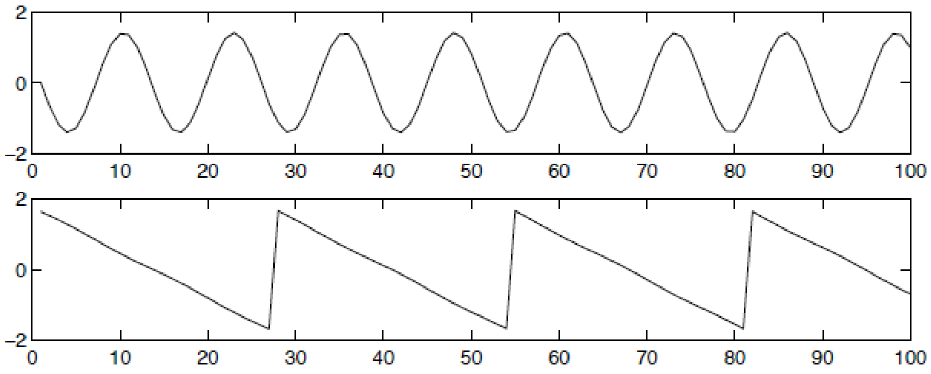
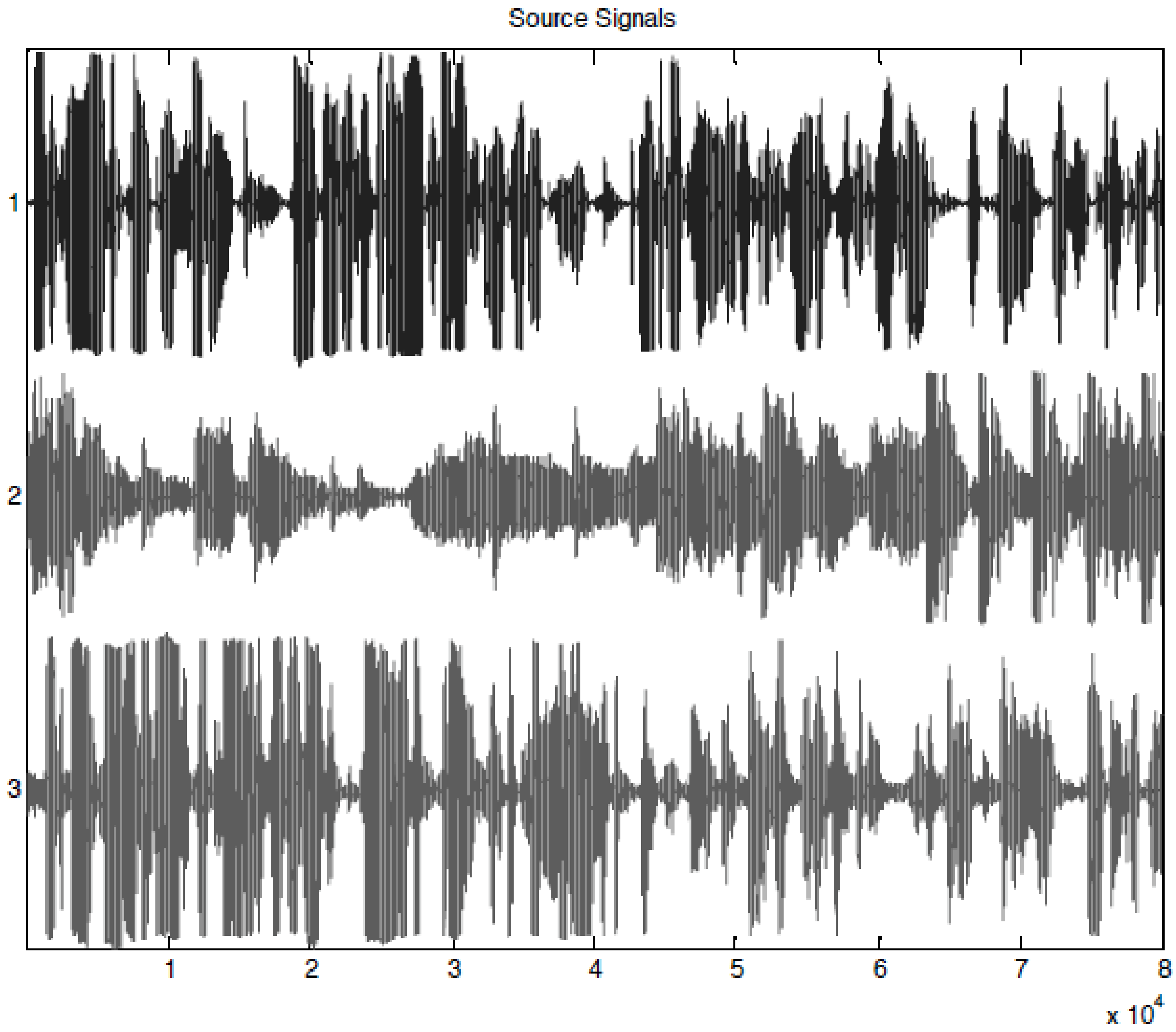
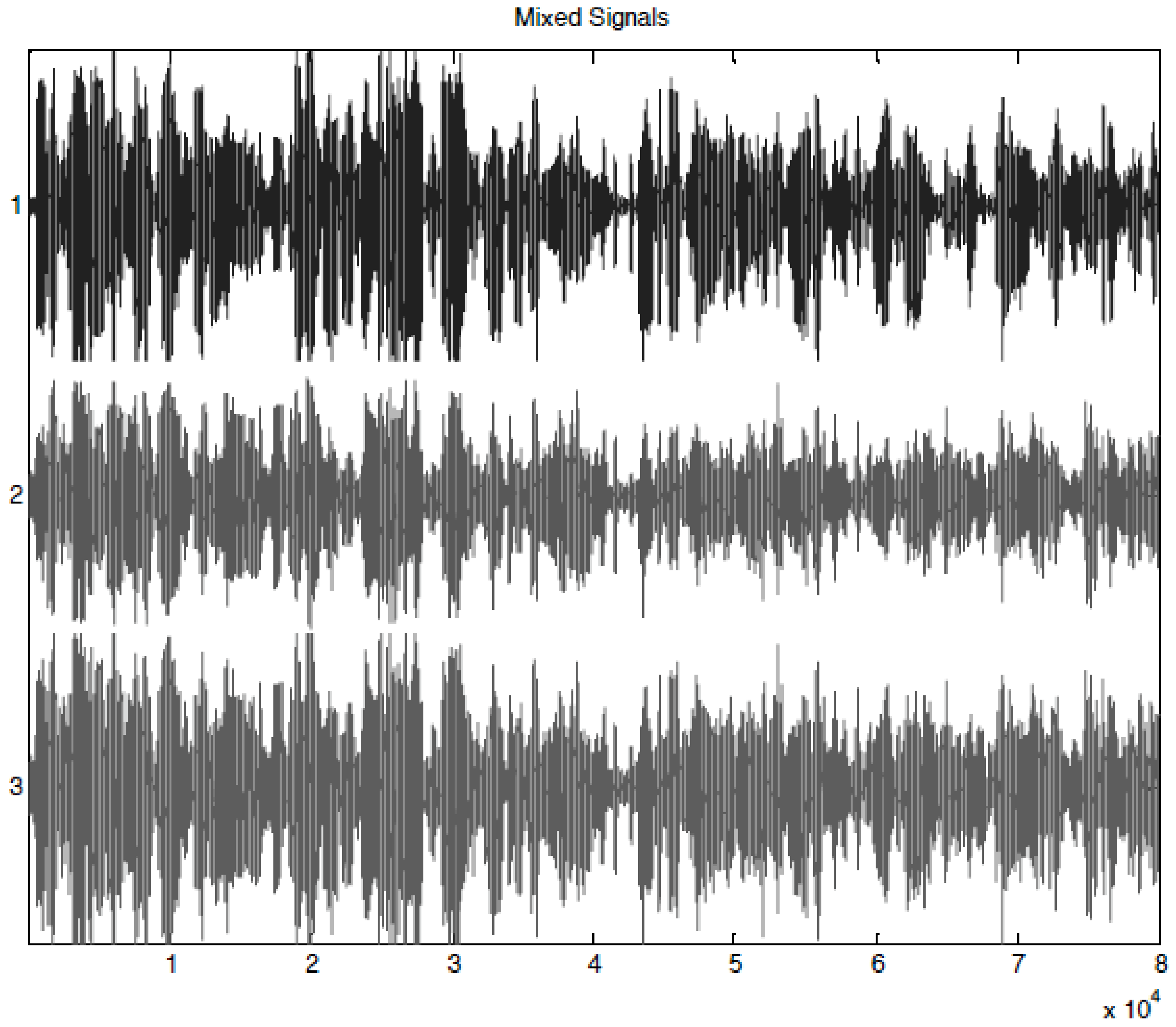
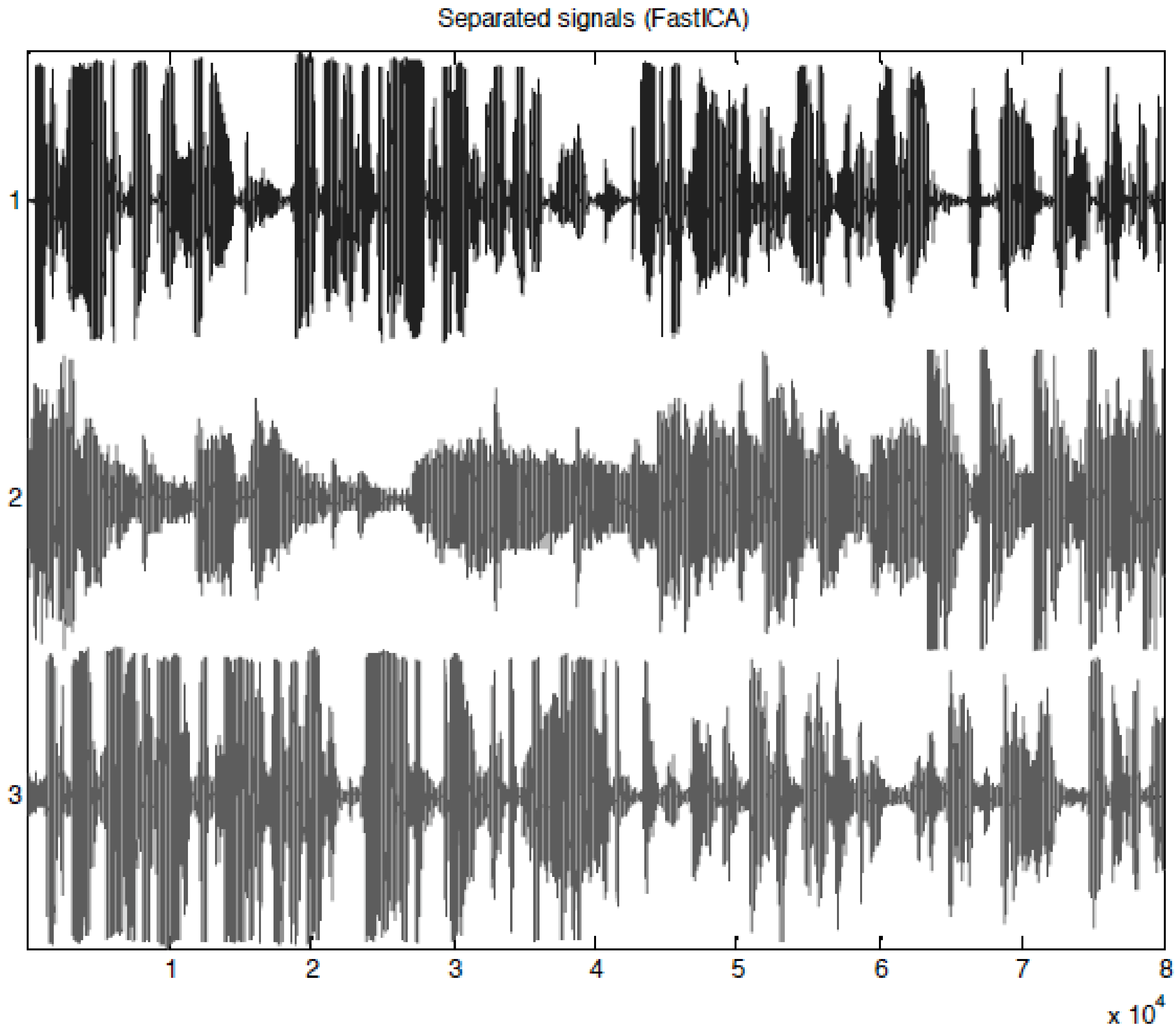


Figure 3: The estimates of the original source signals, estimated using only the observed signals in Fig. 2. The original signals were very accurately estimated, up to multiplicative signs.


Source Signals
1
2
3
1
2
3
4
5
6
7
8
x 10^{4}


Mixed Signals
1
2
3
1
2
3
4
5
6
7
8
x 10^4


Separated signals (FastICA)
1
2
3
1
2
3
4
5
6
7
8
x 10^4

ICA for Nonlinear Problems?

Source: http://videolectures.net/mlss05au_hyvarinen_ica/

Nonlinear is nonparametric?

Usually, this means very general functions:
x=f(s) where f is "almost anything"

Should perhaps be called nonparametric as in statistics

Can this be solved for a general f(s)? No.

Indeterminacy of nonlinear ICA

We can always find an infinity of different nonlinear functions g so that $y=g(x)$ has independent components and these are very different solutions from each other;

We must restrict the problem in some way;

Recently, many people propose
$x_i = f(w_i^t x)$
where f is scalar. "Post-nonlinear" mixtures

Nonlinear ICA

A simple solution?

Do nonlinear PCA
(e.g. Kohonen map) and then linear ICA
(very constrained)

Another solution: constrain the nonlinearity to be smooth, use this as Bayesian prior (Harri Valpola et al)

A lot of work needs to be done...