Advanced Data Analysis: Blind Source Separation

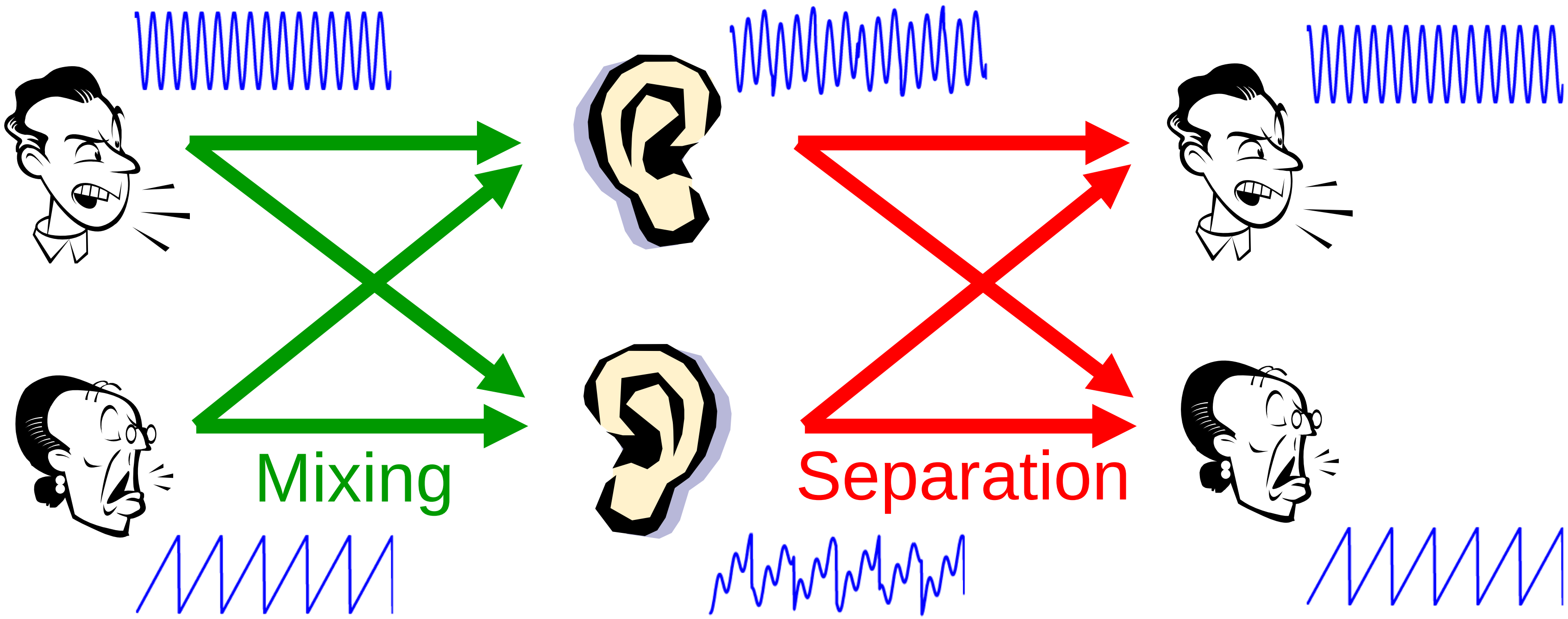
Masashi Sugiyama (Computer Science)

W8E-406, sugi@cs.titech.ac.jp

http://sugiyama-www.cs.titech.ac.jp/~sugi

Blind Source Separation

- Cocktail-party problem:

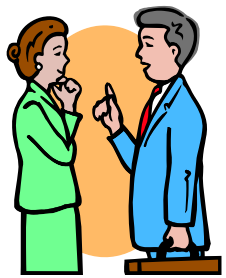


- We want to separate mixed signals into original ones.

Demonstration

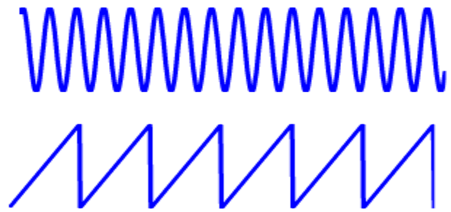
	Mixed signal	Separated signal 1	Separated signal 2
Conversation + Conversation			
Conversation + Instrument			

From http://www.brain.kyutech.ac.jp/~shiro/research/blindsep.html

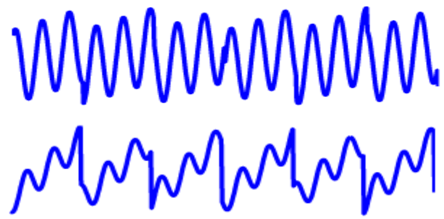


Formulation

- Source signals:
  - Speaker 1: $s_1^{(1)}, s_2^{(1)}, \ldots, s_n^{(1)}$
  - Speaker 2: $s_1^{(2)}, s_2^{(2)}, \ldots, s_n^{(2)}$



- Mixed signals:
  - Left ear: $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$
  - Right ear: $x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}$



$$x_i^{(1)} = m_{11} s_i^{(1)} + m_{12} s_i^{(2)}$$
$$x_i^{(2)} = m_{21} s_i^{(1)} + m_{22} s_i^{(2)}$$

Formulation (cont.)

- In matrix form:

$$\boldsymbol{x}_i = \boldsymbol{M}\boldsymbol{s}_i$$

$$\boldsymbol{x}_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \end{pmatrix} \quad \boldsymbol{M} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \quad \boldsymbol{s}_i = \begin{pmatrix} s_i^{(1)} \\ s_i^{(2)} \end{pmatrix}$$

- More generally
  - $\boldsymbol{x}_i, \boldsymbol{s}_i$: d-dimensional vectors
  - $\boldsymbol{M}$: d-dimensional matrix.

Problem

$$\boldsymbol{x}_i = \boldsymbol{M}\boldsymbol{s}_i$$

- We want to estimate $\{\boldsymbol{s}_i\}_{i=1}^n$ from $\{\boldsymbol{x}_i\}_{i=1}^n$.
- Approach: Estimate $\boldsymbol{M}$, and use its inverse for obtaining $\{\widehat{\boldsymbol{s}}_i\}_{i=1}^n$.

$$\widehat{\boldsymbol{s}}_i = \widehat{\boldsymbol{M}}^{-1}\boldsymbol{x}_i$$

- In BSS, the followings may not be important:
  - Permutation of separated signals
  - Scaling of separated signals
- Therefore, we estimate $\widehat{\boldsymbol{M}}^{-1}$ up to permutation and scaling of rows.

Assumptions

- $\{\boldsymbol{s}_i\}_{i=1}^n$ are i.i.d. random variables with mean zero and covariance identity:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{s}_i = \boldsymbol{0} \qquad \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{s}_i\boldsymbol{s}_i^\top = \boldsymbol{I}_d$$

- $\{s^{(j)}\}_{j=1}^d$ are mutually independent:

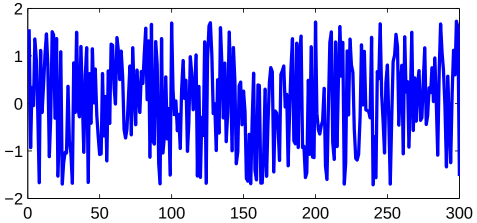
$$P(s^{(1)}, s^{(2)}, \ldots, s^{(d)}) = P(s^{(1)})P(s^{(2)})\cdots P(s^{(d)})$$

- $\{s^{(j)}\}_{j=1}^d$ are non-Gaussian.
- $\boldsymbol{M}$ is invertible.
- BSS under source independence is called independent component analysis.

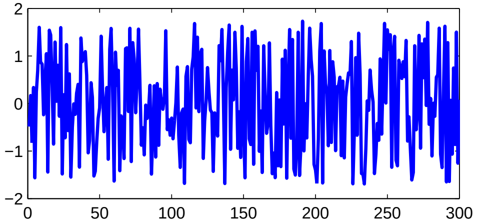
Example

Source signals (uniform)

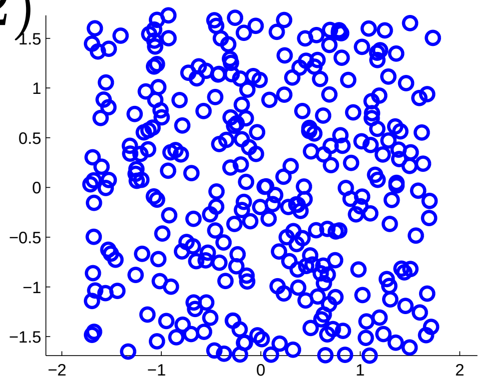
$s^{(1)}$



$s^{(2)}$



$s^{(2)}$

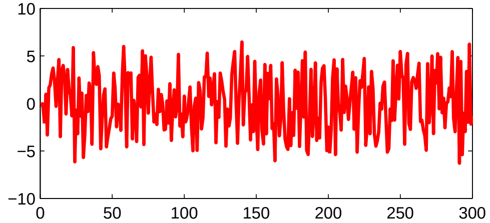


$s^{(1)}$

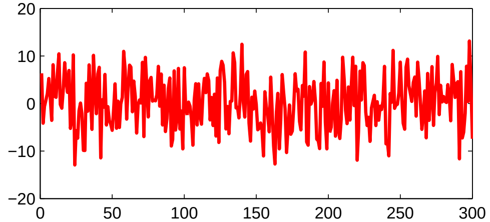
Mixed signals

$$M = \begin{pmatrix} 1 & 3 \\ 5 & 1 \end{pmatrix}$$

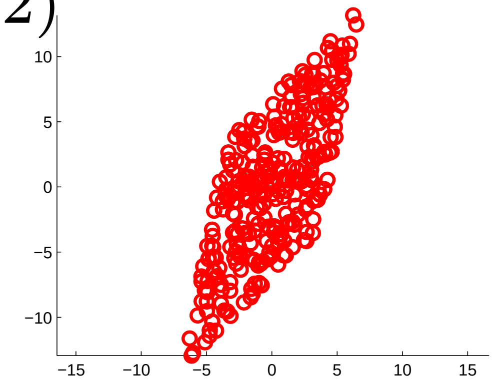
$x^{(1)}$



$x^{(2)}$



$x^{(2)}$



$x^{(1)}$

Data Sphering

- Sphering (or pre-whitening):

$$\widetilde{\boldsymbol{x}}_i = \boldsymbol{C}^{-\frac{1}{2}} \boldsymbol{x}_i \qquad \boldsymbol{C} = \frac{1}{n} \sum_{i'=1}^{n} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}^{\top}$$

- Then

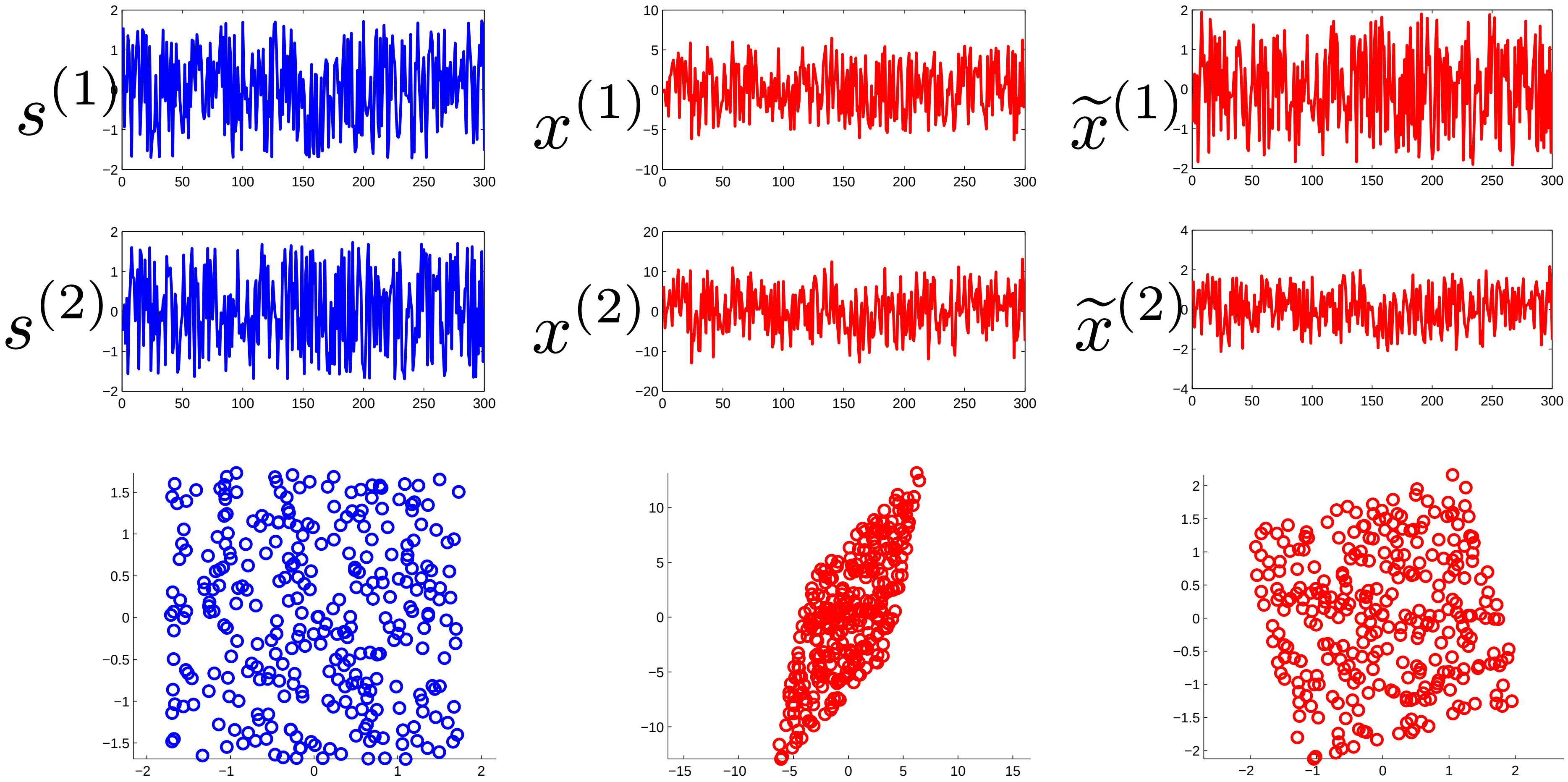
$$\widetilde{\boldsymbol{x}}_i = \widetilde{\boldsymbol{M}} \boldsymbol{s}_i \qquad \widetilde{\boldsymbol{M}} = \boldsymbol{C}^{-\frac{1}{2}} \boldsymbol{M}$$

- Now we want to estimate $\widetilde{\boldsymbol{M}}$ from $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^{n}$, and obtain $\{\widehat{\boldsymbol{s}}_i\}_{i=1}^{n}$ by

$$\widehat{\boldsymbol{s}}_i = \boldsymbol{W} \widetilde{\boldsymbol{x}}_i \qquad \boldsymbol{W} = \widetilde{\boldsymbol{M}}^{-1}$$

Example

Source signals (uniform)	Mixed signals	Sphered signals
	$\boldsymbol{M} = \begin{pmatrix} 1 & 3 \\ 5 & 1 \end{pmatrix}$	$\widetilde{\boldsymbol{x}}_i = \boldsymbol{C}^{-\frac{1}{2}} \boldsymbol{x}_i$



Orthogonal Matrix

- $\widetilde{\boldsymbol{M}}$ is an orthogonal matrix since

$$\widetilde{\boldsymbol{C}} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^\top = \boldsymbol{I}_d$$

$$\widetilde{\boldsymbol{C}} = \widetilde{\boldsymbol{M}}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{s}_i\boldsymbol{s}_i^\top\right)\widetilde{\boldsymbol{M}}^\top = \widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{M}}^\top$$

- Therefore,

$$\widehat{\boldsymbol{s}}_i = \boldsymbol{W}\widetilde{\boldsymbol{x}}$$

$$\boldsymbol{W} = \widetilde{\boldsymbol{M}}^{-1} = \widetilde{\boldsymbol{M}}^\top \equiv (\boldsymbol{w}^{(1)}|\boldsymbol{w}^{(2)}|\cdots|\boldsymbol{w}^{(d)})^\top$$

$\{\boldsymbol{w}^{(j)}\}_{j=1}^{d}$:Orthonormal basis

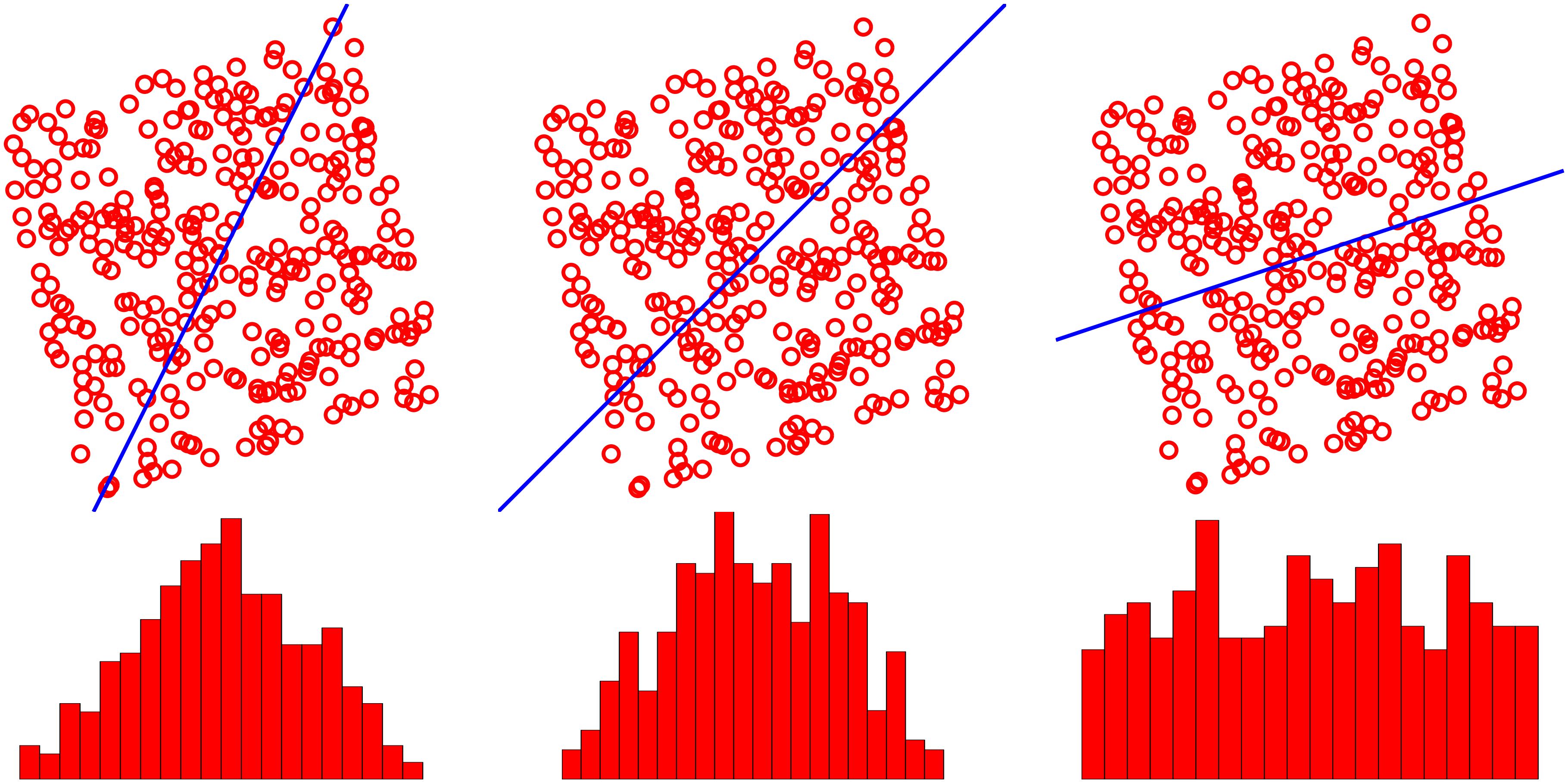
$$\widehat{s}_i^{(j)} = \langle \boldsymbol{w}^{(j)}, \widetilde{\boldsymbol{x}}_i \rangle$$

Non-Gaussian Is Independent

- Now we want to find an ONB $\{\boldsymbol{w}^{(j)}\}_{j=1}^{d}$ such that $\{\widehat{s}^{(j)}\}_{j=1}^{d}$ are independent.
- Central limit theorem: Sum of independent variables tends to be Gaussian.
- Conversely, non-Gaussian variables are independent.
- We find non-Gaussian directions in $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^{n}$.

Example (cont.)

- Non-Gaussian direction is independent.

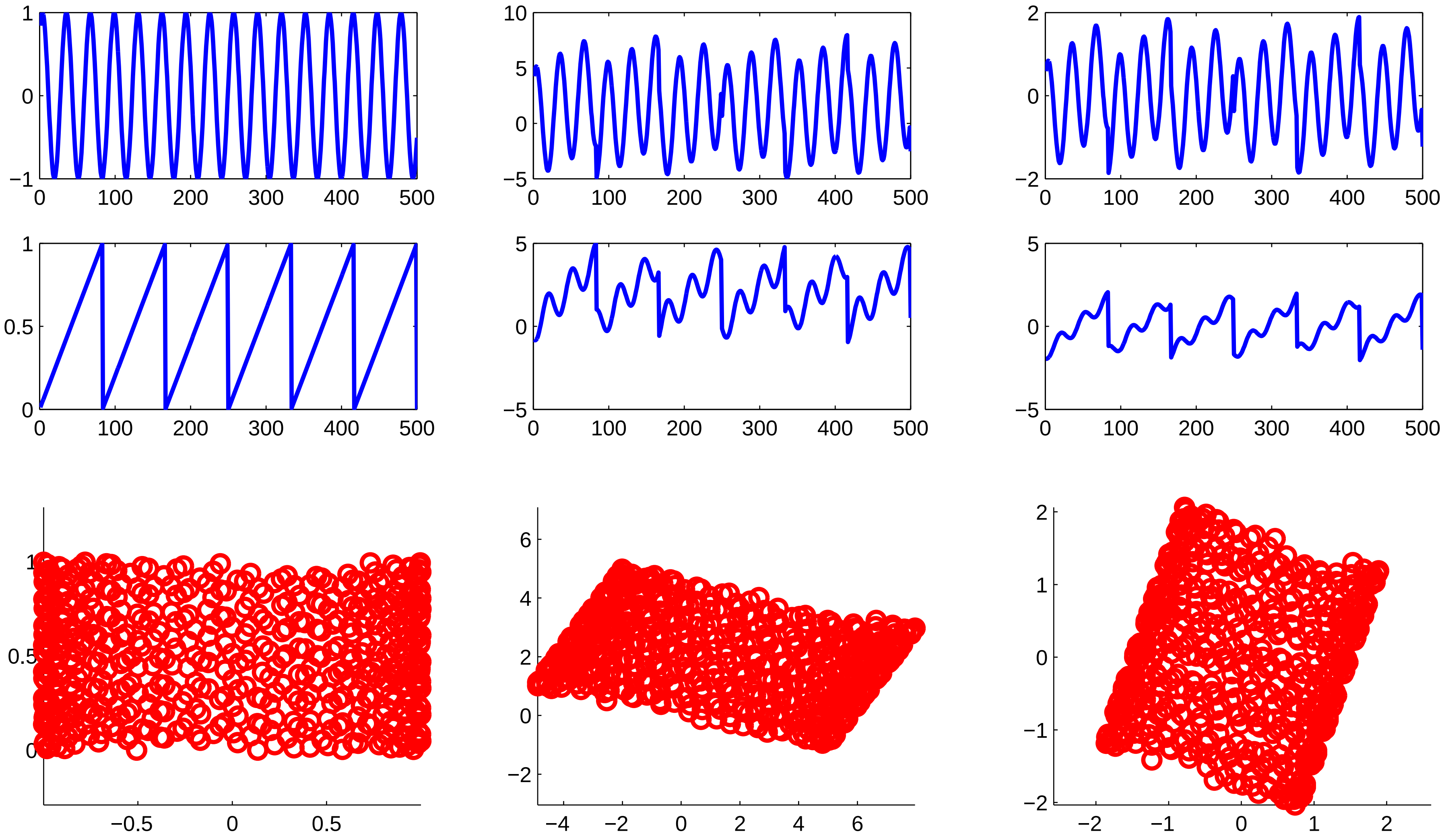


ICA by Projection Pursuit

- Finding non-Gaussian directions can be achieved by projection pursuit algorithms!
  - Center and sphere the data.
  - Find non-Gaussian directions by PP.
- We may use an approximate Newton-based PP method, which is called FastICA.

Example 2

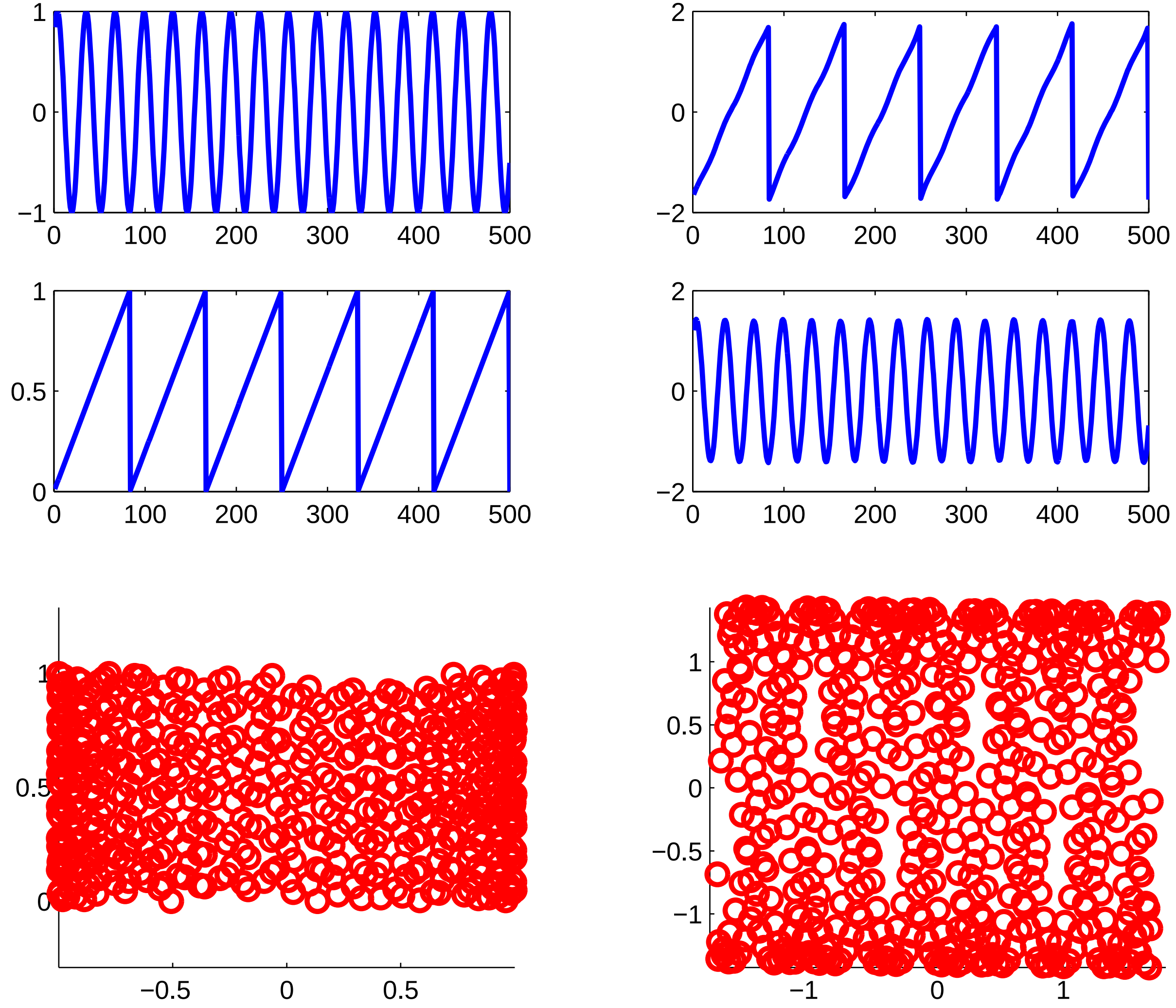
Source | Mixed | Sphered



Example 2 (cont.)

Source

Separated



- Original signals are recovered up to permutation and scaling.

Notification of Final Assignment

- Data Analysis: Apply dimensionality reduction or clustering techniques to your own data set and “mine” something interesting!

- Deadline: July 31st (Wed) 17:00
  - Bring your printed report to W8E-406.
  - E-mail submission is also possible (though not recommended).

Mini-Conference on Data Analysis

- Program
- Presentation:
  - 7 mins (+ 2 mins for Q&A)
  - Description of your data
  - Methods to be used
  - Outcome
- Slides should be in English.

July 16th	July 23rd
Ikko Yamane	Yisha Sun
Tomoya Sakai	Tran Hai Dang
Kiung Park	Hao Zhang
Janya Sainui	Biriukova Nataliia
Zhuolin Liang	Fumito Nakamura
Sagong Sun	Juuti Mika
Kishan Wimalawarne	Vektor Dewanto
Daniel Louw	Luis Cardona
Duong Nguyen	Song Yang
Voot Tangkaratt	Mengxi Lin

Schedule

- July 9th: Preparation for Mini-Conference
- July 16th: Mini-Conference Day 1
- July 23rd: Mini-Conference Day 2