Chapter 5

Independent Component Analysis

Part I: Introduction and applications

max planck institut
informatik

Motivation

Skillikorn chapter 7

Cocktail party problem

Cocktail party problem

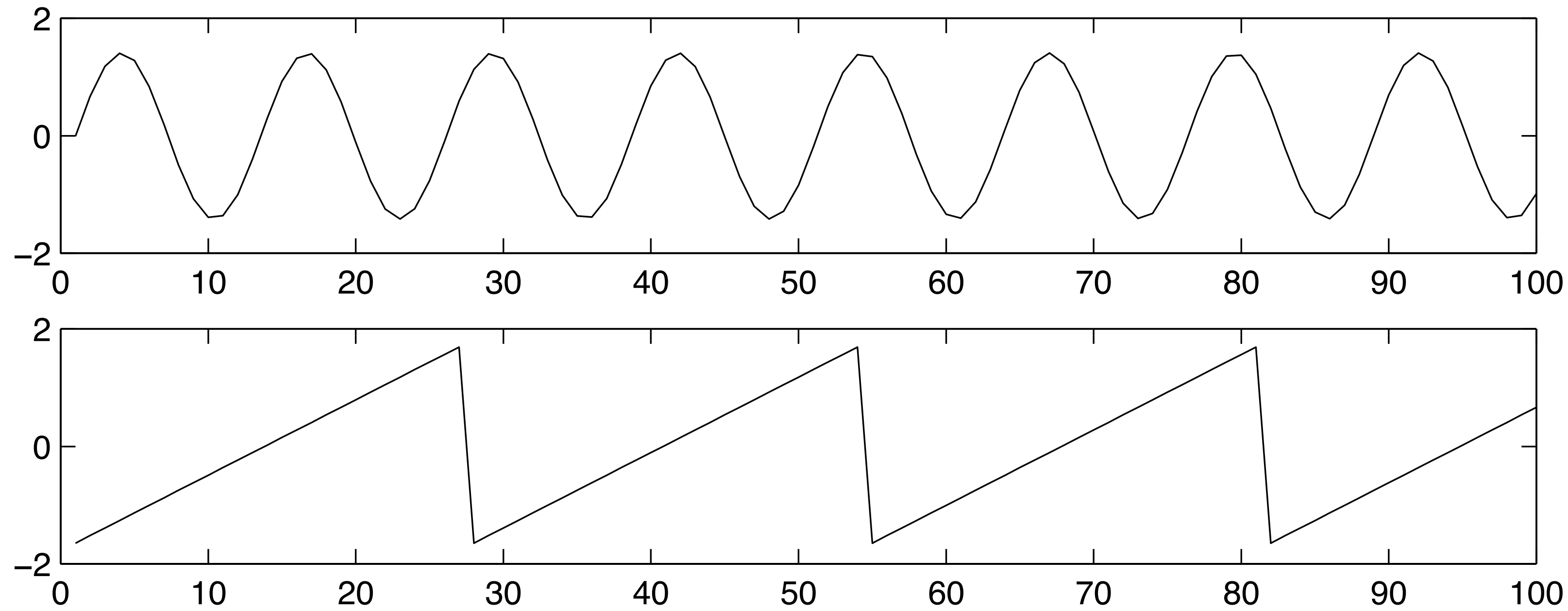
- Assume we have two microphones recording two speakers
  - We observe $x_1(t)$ and $x_2(t)$ where t is time
- Assume what the speakers say is statistically independent
  - Real signals are $s_1(t)$ and $s_2(t)$
  - $x_j(t) = a_{1j}s_1(t) + a_{2j}s_2(t) \Rightarrow \boldsymbol{x} = \boldsymbol{sA}$

Cocktail party examples

Original signals

s_1

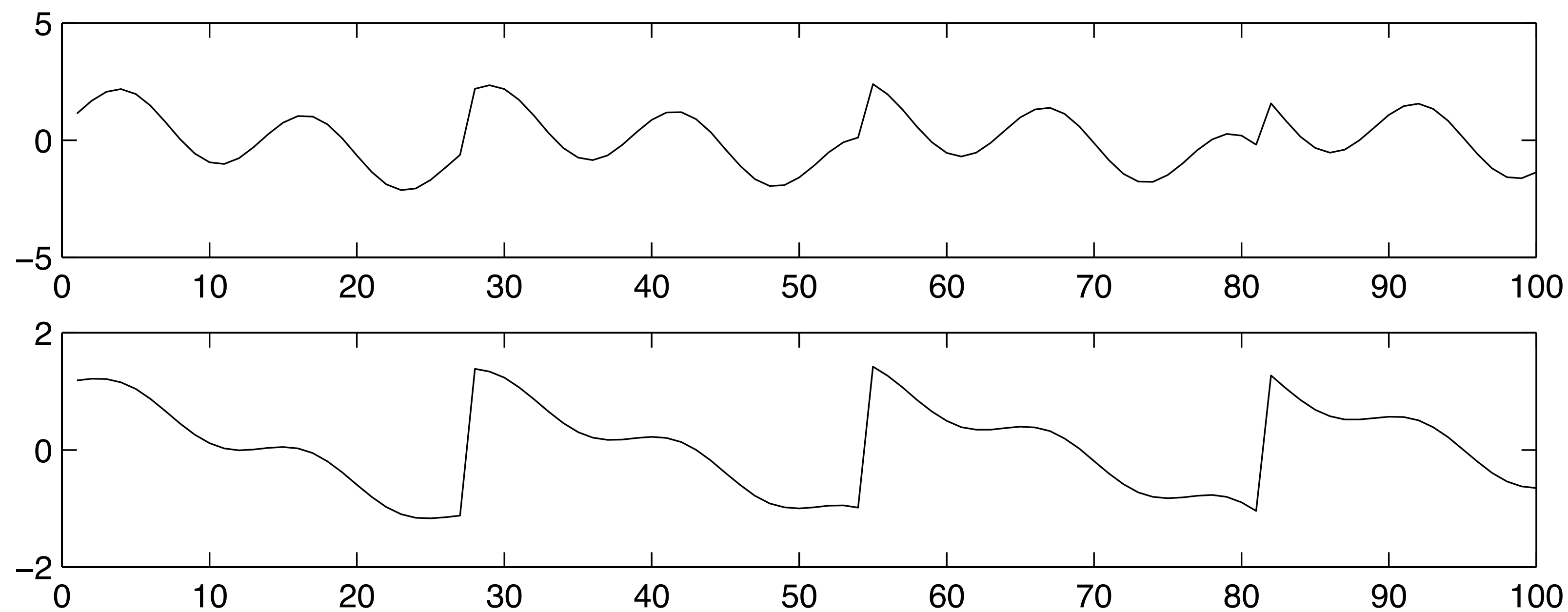
s_2



Observed signals

x_1

x_2



Hyvärinen & Oja 2000

Cocktail party question

- **Problem:** can we reconstruct the original signal and mixing coefficients knowing only the mixed signals?
  - I.e. can we build $\boldsymbol{A}$ and $\boldsymbol{s}$ knowing only $\boldsymbol{x}$?
- If we know $\boldsymbol{x}$ and $\boldsymbol{A}$, the problem is easy
  - But how to find $\boldsymbol{A}$?

The Definition

ICA definition

- **Setting**. Let $x_j \in \mathbb{R}$, $j=1,\ldots,n$ be observed random variables. Assume there exists n latent random variables $s_i \in \mathbb{R}$ and latent coefficients a_{ij} such that $x_j = \sum_i a_{ij} s_i$ for all j.
  - $\boldsymbol{x} = \boldsymbol{s}\boldsymbol{A}$ and for T observations, $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}$ where $\boldsymbol{X}$ and $\boldsymbol{S}$ have T rows
- **Problem**. Find $\boldsymbol{A}$ and $\boldsymbol{s}$ given $\boldsymbol{x}$

ICA assumptions (important slide!)

- Original signals s_i are mutually **statistically independent**
- **At most one** original signal s_i is **normally distributed**
- The mixing matrix $\boldsymbol{A}$ is **square** and **invertible**
  - This is not necessary but simplifies the theory

ICA is identifiable

- Under the above assumptions, we can estimate ***A*** and ***s*** up to
  - signs and scales of components
  - ordering of components
- In many applications this is good enough
  - And we can impose extra constraints for better stability

Constraints

- The input variables must have zero mean
  - Center the columns of $\boldsymbol{X}$ if needed
- Often, columns of $\boldsymbol{S}$ are fixed to unit variance
  - The factors are pushed to $\boldsymbol{A}$

Statistical independency

- Two random variables x and y are **uncorrelated** if $E[xy] = E[x]E[y]$

E[·] is expectation

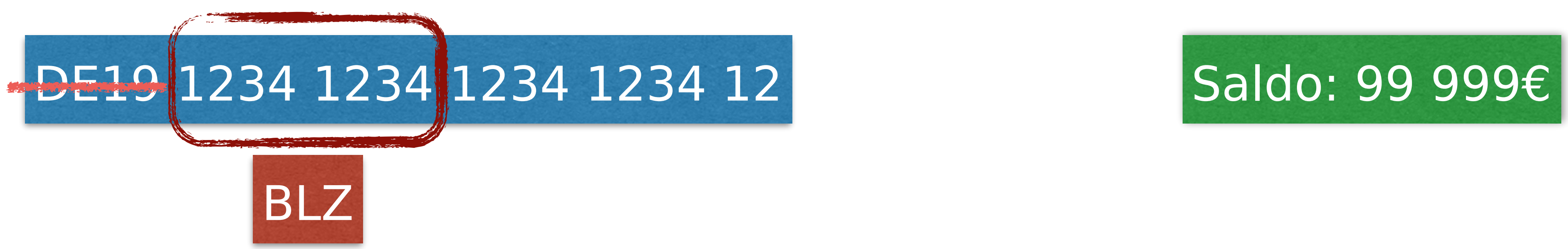
  - Knowing $E[x]$ tells us nothing of $E[xy]$
  - Recall: **covariance** $\text{cov}(x, y) = E[xy] – E[x]E[y]$
- R.v.'s x and y are **statistically independent** if for any transformation f_1 and f_2 $E[f_1(x)f_2(y)] = E[f_1(x)]E[f_2(y)]$

Example of independency

- Let x and y be s.t. $\Pr[(x, y) = (a, b)] = 1/4$ for $(a, b) \in \{(0,1), (0,-1), (1,0), (-1,0)\}$
- $\mathrm{cov}(x, y) = E[xy] - E[x]E[y] = 0 - 0 \cdot 0 = 0$
- Let $x \mapsto x^2$ and $y \mapsto y^2$
  - $E[x^2y^2] - E[x^2]E[y^2] = 0 - 0.5 \cdot 0.5 = -0.25$
    $\Rightarrow$ x and y are uncorrelated but not independent

Independency is strong

- IBAN account numbers and account holder's ability to pay bills are probably uncorrelated



- But they might still be dependent
  - First 8 numbers (after DE*xx*) are the bank and branch identifier

Whitening the signal

- **Whitening** is a transformation of random variables x_i to new variables y_i s.t. $E[y_i y_j] = 0$ if $i \neq j$ and $E[y_i y_i] = 1$
  - Zero mean is assumed
  - Thus, y_i are uncorrelated with unit variance
    - Compare to z-scores

Computing the whitening

- Decorrelation can be computed in many ways
  - ZCA whitening, Cholesky whitening, PCA whitening
- We will use the SVD
- Let $\boldsymbol{X}$ have $\boldsymbol{x}_i$ as its columns and observations as its rows and let $\boldsymbol{X} = \boldsymbol{U\Sigma V}^T$ be its SVD
  - Columns of $\boldsymbol{U}$ give the whitened variables

ICA and SVD

- SVD (or PCA) cannot solve ICA
  - Essentially: they find uncorrelated but not necessarily independent components
- Whitening gives us $\boldsymbol{X}\boldsymbol{V}\boldsymbol{\Sigma}^{-1} = \boldsymbol{S}\boldsymbol{A}\boldsymbol{V}\boldsymbol{\Sigma}^{-1} = \boldsymbol{S}\boldsymbol{B}$
  - $\boldsymbol{B}$ is new mixing matrix
  - Whitening is a standard pre-processing technique in ICA

Why Gaussians are forbidden?

- Let s_1 and s_2 be original independent components whose joint distribution is Gaussian
  $$p(s_1, s_2) = \frac{1}{2\pi} \exp\left\{-\frac{s_1^2+s_2^2}{2}\right\} = \frac{1}{2\pi} \exp\left\{-\frac{\|\boldsymbol{s}\|^2}{2}\right\}$$
- Let $\boldsymbol{A}$ be orthogonal
  - $\boldsymbol{x} = \boldsymbol{sA}$ is Gaussian with covariance matrix equal to identity and $p(x_1, x_2) = \frac{1}{2\pi} \exp\left\{-\frac{\|\boldsymbol{s}\|^2}{2}\right\}$
  - No $\boldsymbol{A}$ in the pdf, the original and mixed distributions are identical

More on Gaussians

- Two uncorrelated Gaussians are necessarily independent
  - With Gaussian distributions, we loose the strength of the independency
  - Equivalently, the joint distribution of independent Gaussians is rotationally invariant
- But we can do ICA with **at most one** Gaussian distribution

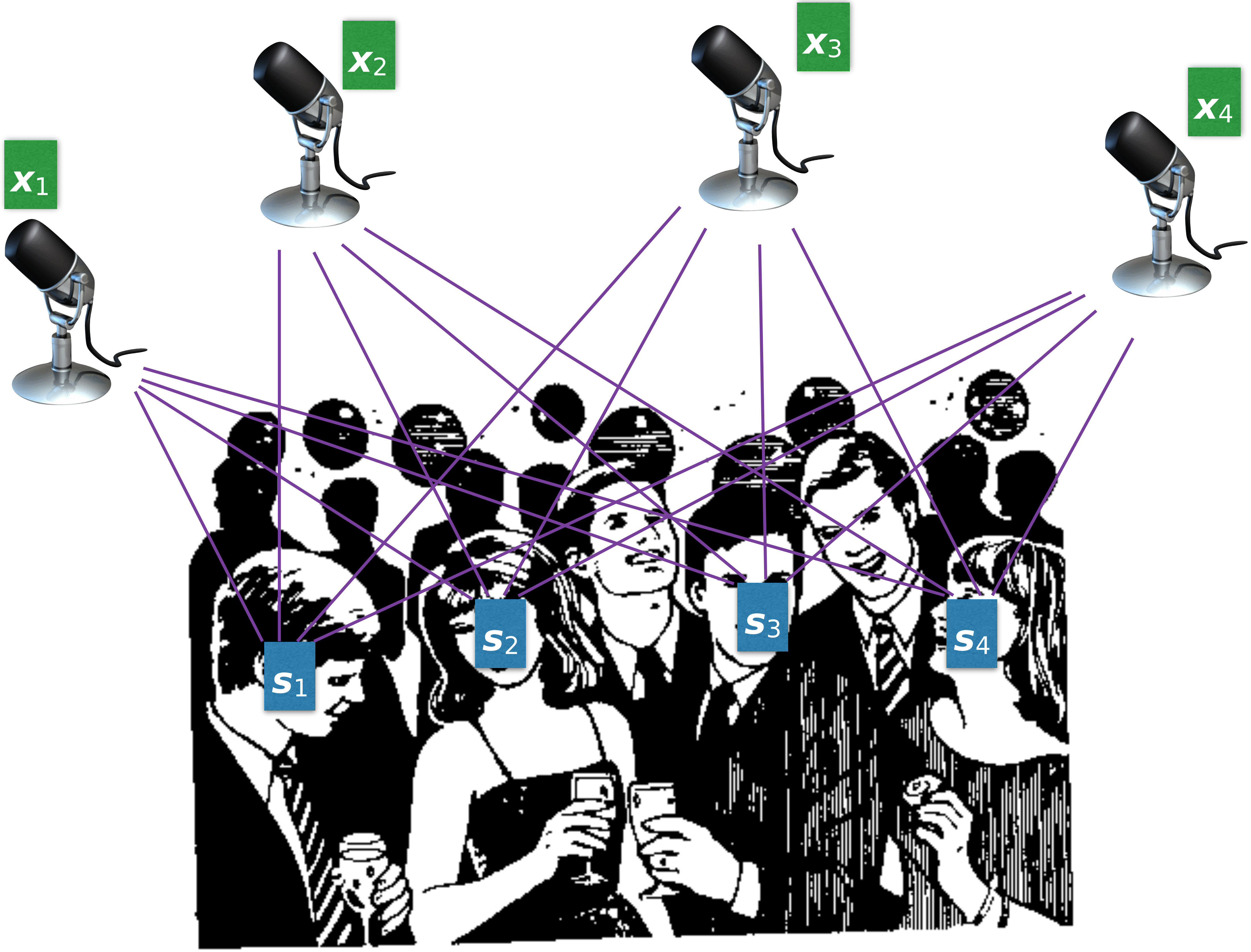
ICA and other matrix factorizations

- ICA does not reduce the rank of the matrix
  - But we can apply the whitening first
- ICA does not have noise in the model
  - Some components express noise (c.f. SVD)
  - Noise is often Gaussian, and hence, if one factor is Gaussian, it is considered the noise

Interpreting an ICA

Factor interpretation

- Most natural interpretation in many applications
- Columns of $\boldsymbol{S}$ give the independent components
  - People in cocktail party
- Rows of $\boldsymbol{A}$ explain how the components are mixed
  - Placement of the microphones



x_1
x_2
x_3
x_4
s_1
s_2
s_3
s_4

Geometric interpretation

- Independent components are not (necessarily) orthogonal
  - They are not axes, per se
- We can still treat the columns of $\boldsymbol{A}$ as coordinates in some space and plot the first two rows (say)
  - But two points that are close in the plot might not be close in reality

Component interpretation

- The rank-1 components can be studied to understand how the columns of $\boldsymbol{S}$ are used to create the data
  - But their ordering is not fixed
- If one column has Gaussian histogram, it can be considered to be noise
  - Columns of $\boldsymbol{S}$ can be ordered based on how non-Gaussian they are (more on that next week)

Applications of ICA

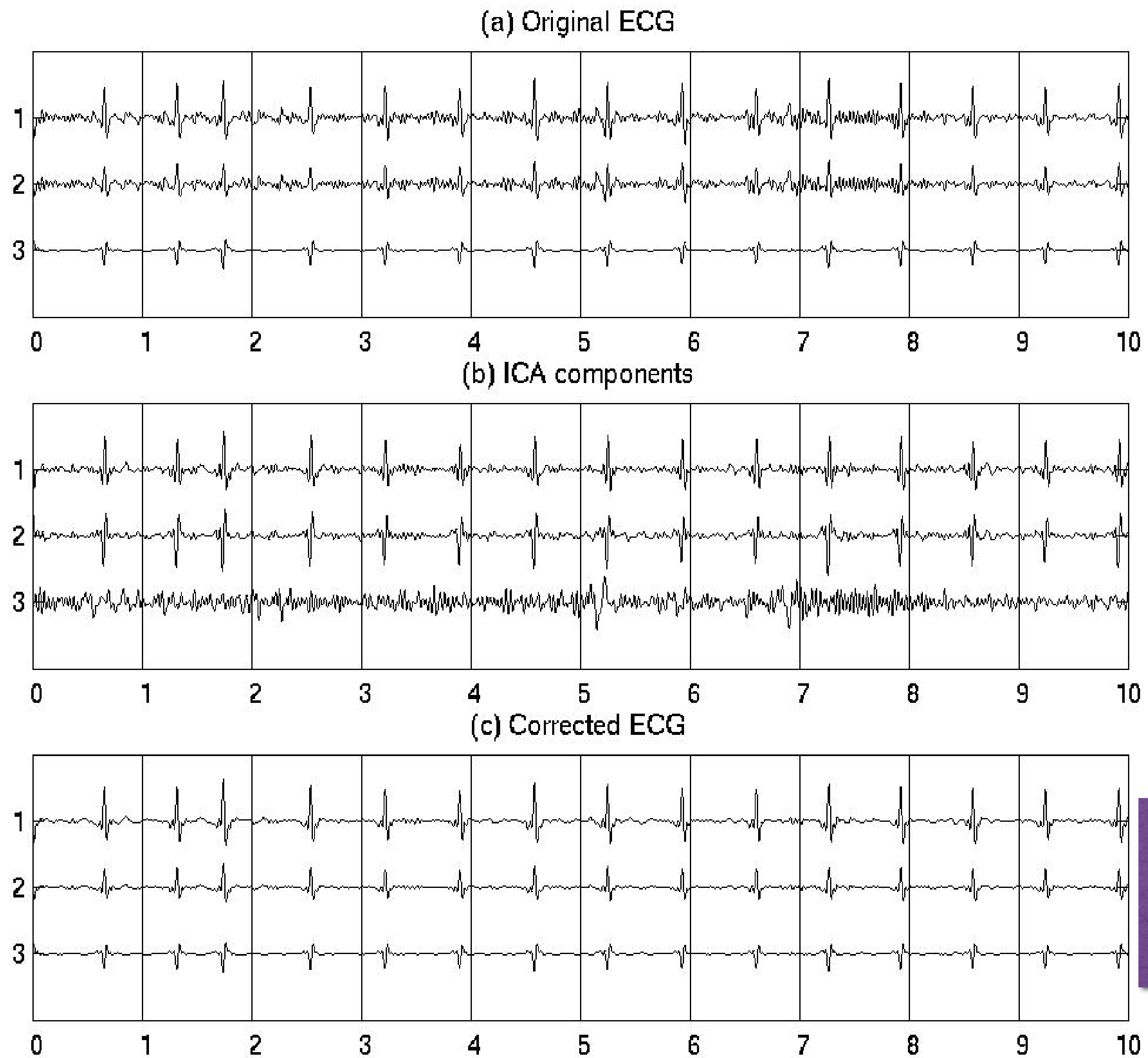
Blind source separation from ECG data

- Electrocardiograms (ECG) have many types of noise and artefacts
  - Electrode movement, muscle movement, etc.
  - Might confuse the interpretation
- ICA can be used to clean the data

He, Clifford & Tarassenko 2006

ECG example #1

Noise



Noise

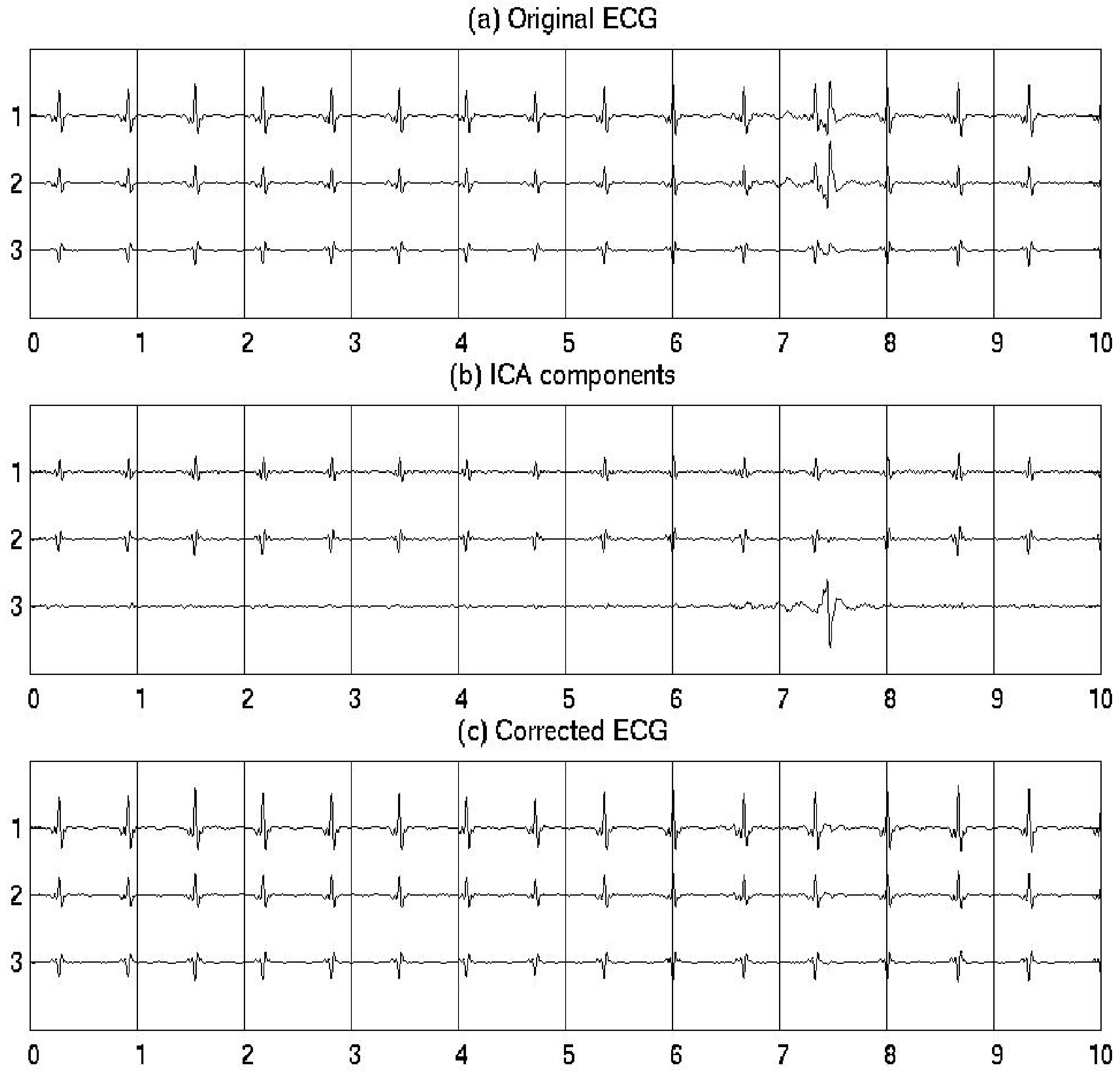
3 input signals

3 independent components

Sum of first two rank-1 components

He, Clifford & Tarassenko 2006

ECG example #2



He, Clifford & Tarassenko 2006

Detecting suspicious messages

- Assume an inmate tries to communicate with criminals outside the prison
  - His communication is statistically monitored and he wants to “fly under the radar”
    - No encryption and no “hot” terms
- Plan: replace hot terms with random terms
  - “Put the file inside the cake” $\rightsquigarrow$ “Put the asparagus inside the cake”
    - These discussions can be identified as they have anomalous term frequencies

Skillicorn chapter 7.5.1

Suspicious message example #1

Scatterplot of first three rows of ***A***

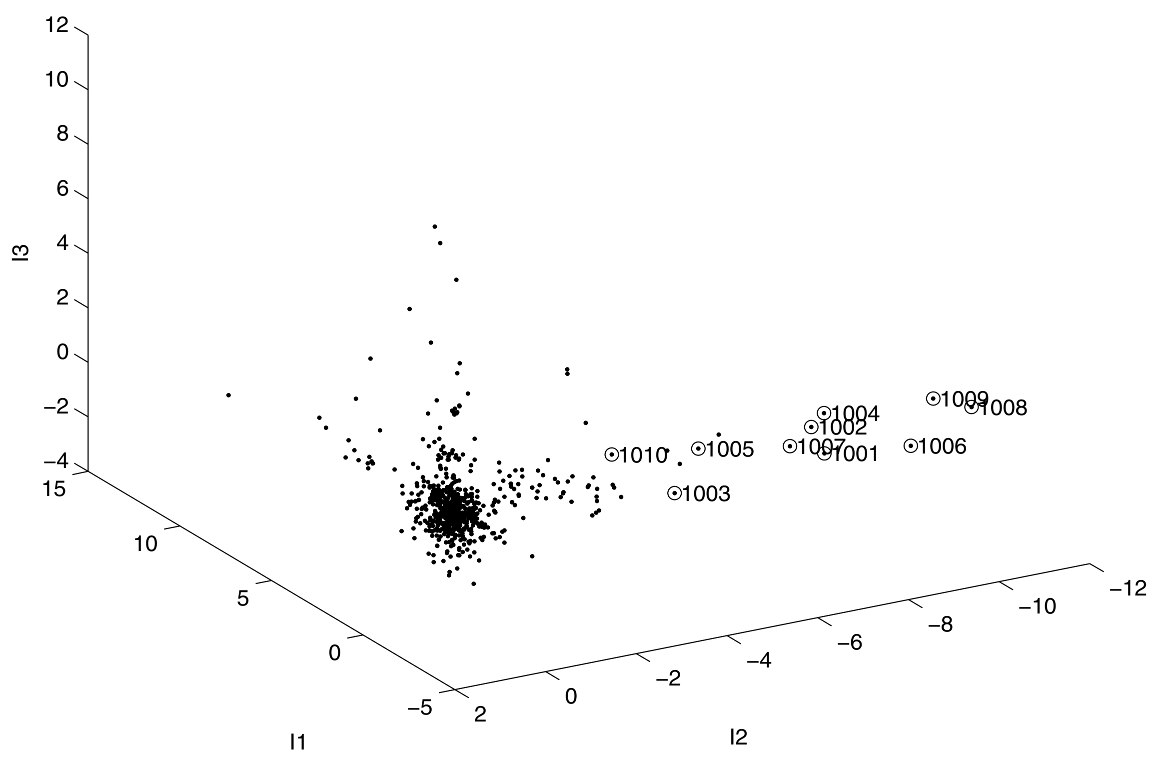


Figure 7.1. *3-dimensional plot from an ICA of messages with correlated unusual word use. The messages of interest are circled.*

ICA finds messages with correlated unusual word use

Skillicorn chapter 7.5.1

Suspicious message example #2

Scatterplot of first three rows of ***A***

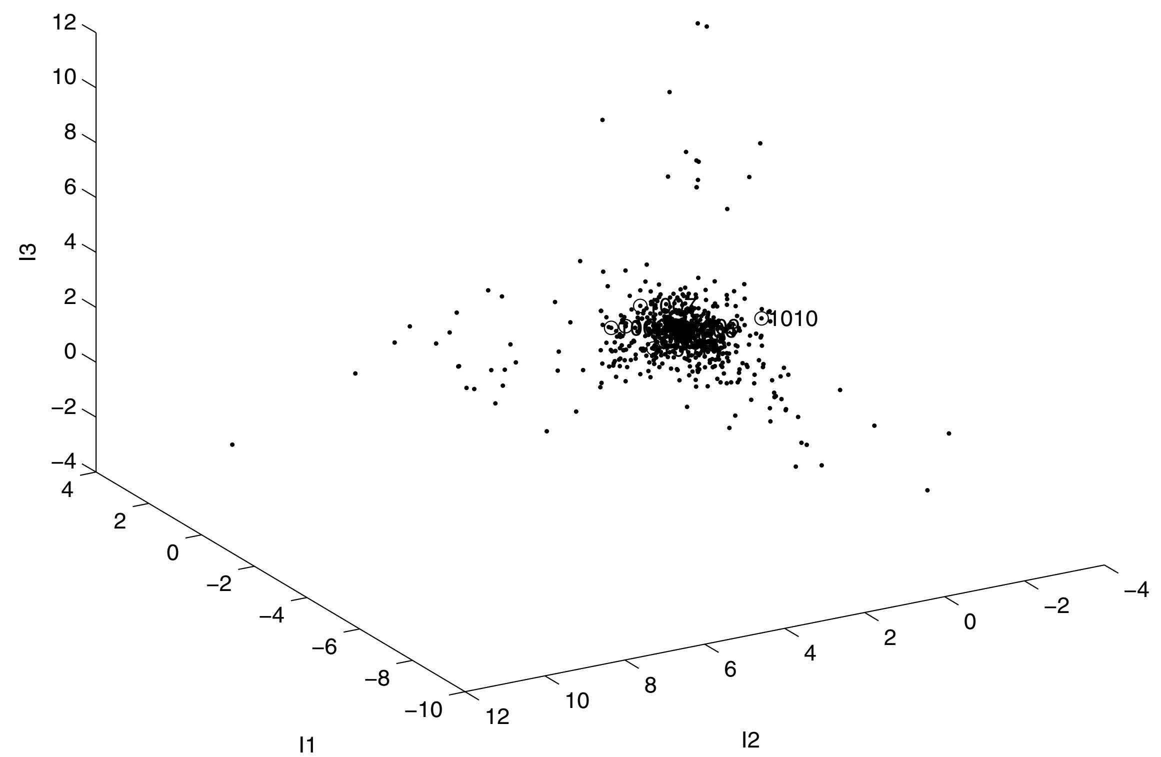


Figure 7.2. *3-dimensional plot from an ICA of messages with correlated ordinary word use. The messages of interest are circled.*

ICA doesn't identify messages with usual word use

Skillicorn chapter 7.5.1

Suspicious message example #3

Scatterplot of first three rows of $\boldsymbol{A}$

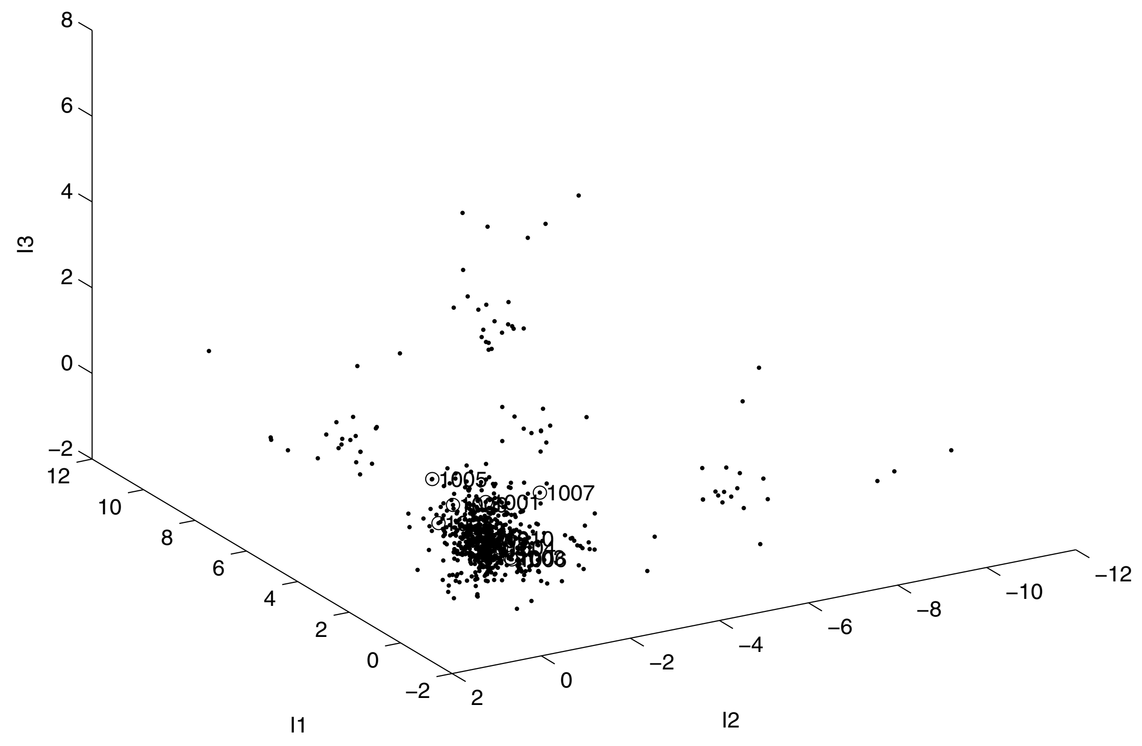


Figure 7.3. *3-dimensional plot from an ICA of messages with unusual word use. The messages of interest are circled.*

ICA doesn't identify messages with uncorrelated unusual word use

Skillicorn chapter 7.5.1

Direction of causality

- Assume we observe x_1 and x_2 and we know one is the cause and the other the effect
    - Which one is which?
- Assume linear regression model
    - Either $x_2 = b_1x_1 + e_1$ or $x_1 = b_2x_2 + e_2$
- If x_1 and x_2 are Gaussian, both models will be equally good

Hyvärinen 2012

Causality and ICA

- If x_1 and x_2 are non-Gaussian, we have
  - Model 1: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ b_1 & 1 \end{pmatrix} \begin{pmatrix} s \\ e_1 \end{pmatrix}$
  - Model 2: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} s \\ e_2 \end{pmatrix}$
- We can solve ICA on the data and decide if mixing matrix $\boldsymbol{A}$ is closer to model 1 or model 2

Hyvärinen 2012

Summary (so far)

- ICA lets us to separate independent, non-Gaussian factors
  - Will not do noise removal or dimensionality reduction (or feature selection) per se
- Orthogonal to SVD
  - Pun perhaps intended
- Next week: How to compute ICA? *Stay tuned!*