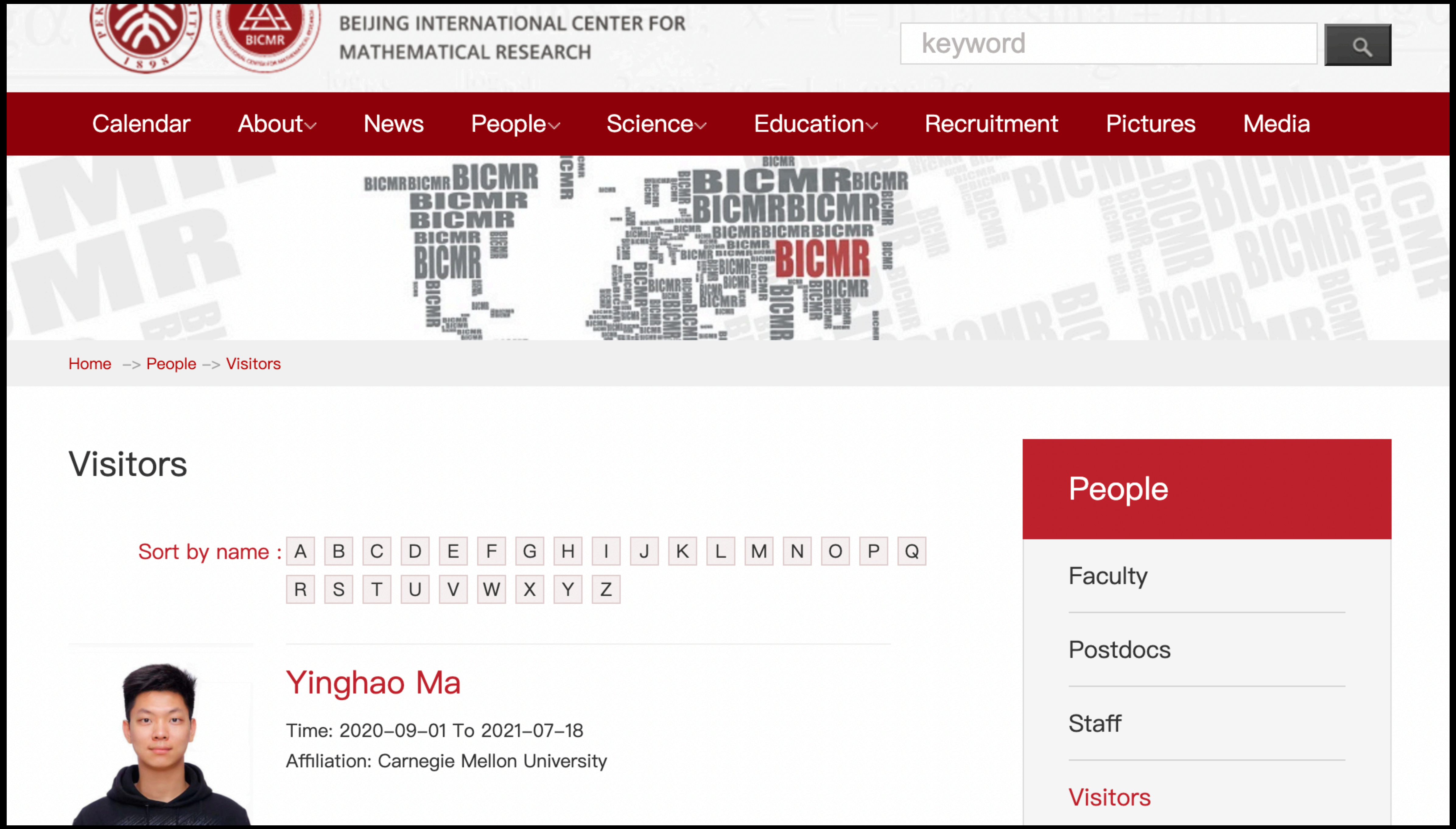
Machine Learning for Signal Processing

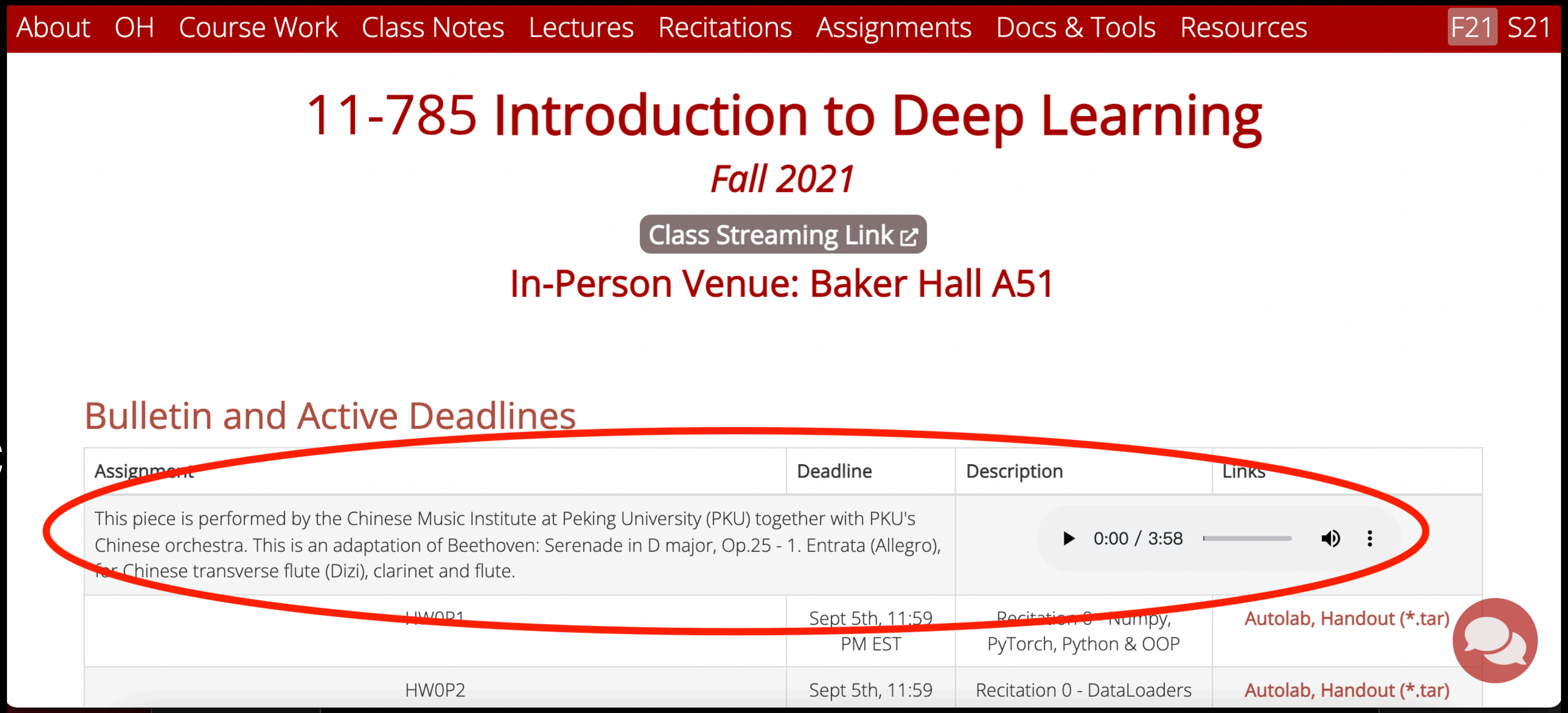
Independent Component Analysis

Oct. 4th & 6th, 2021

Self Introduction

- Yinghao Ma, 马英浩
- https://nicolaus625.github.io
- ~~Failed to be a yound mathematician~~
- A Master Student at Music & Tech, supervised by prof. Richard Stern
- email: yinghaom@andrew.cmu.edu
- Facebook manager of Chinese Music Institute, Peking University



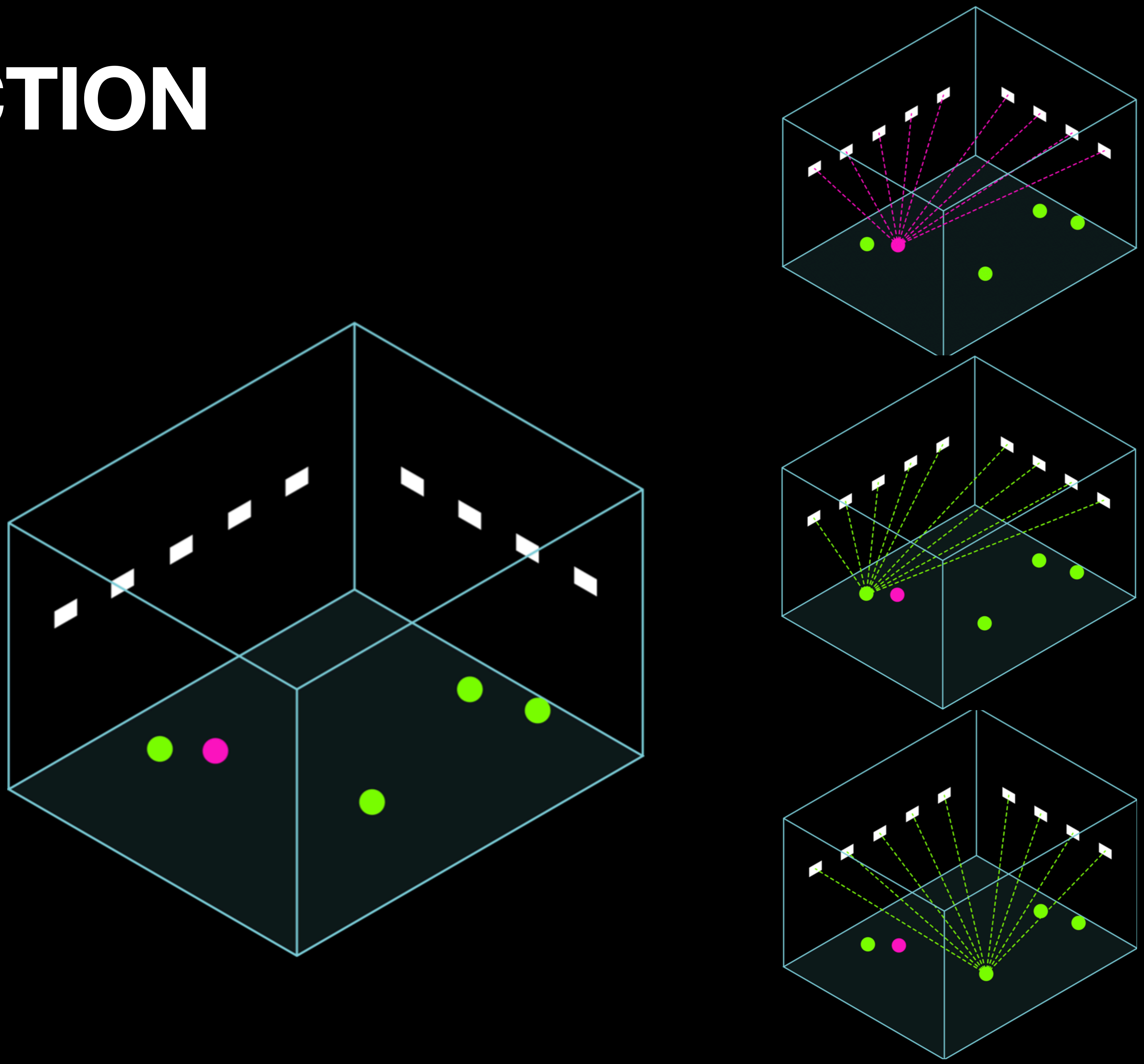


Outline

- INTRODUCTION
  - Independent, non-Gaussian source separation, 2 ideas
- FIRST ICA IMPLEMENTATION: Fourth order blind identification (FBOI)
- MEASURE OF GAUSSIAN
  - Kurtosis divergence, Neg-entropy
- SECOND ICA IMPLEMENTATION: Fast-ICA
- APPLICATION
- COMPARED ICA WITH PCA
- DISADVANTAGE WITH REFINEMENT

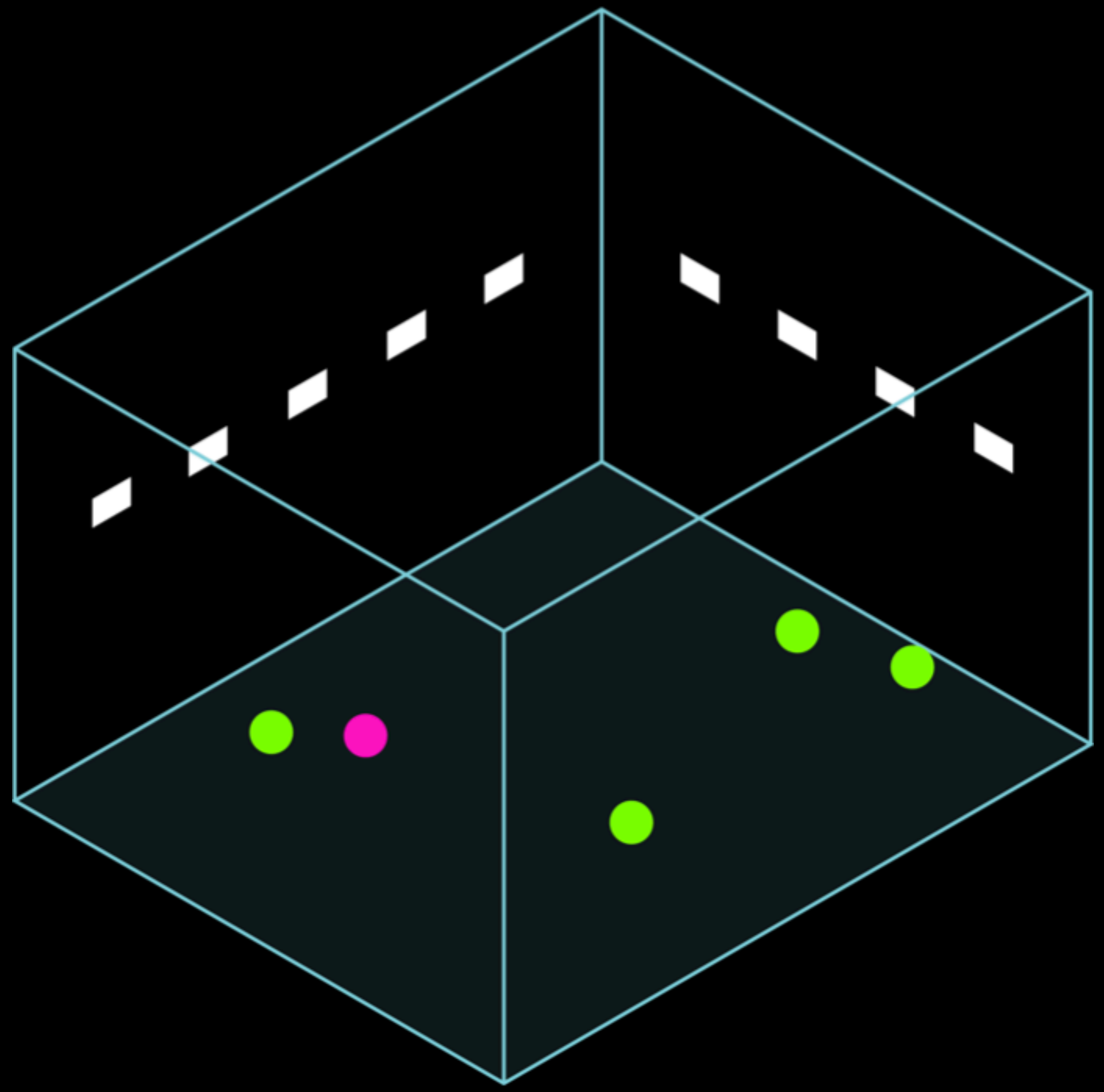
• 1 INTRODUCTION

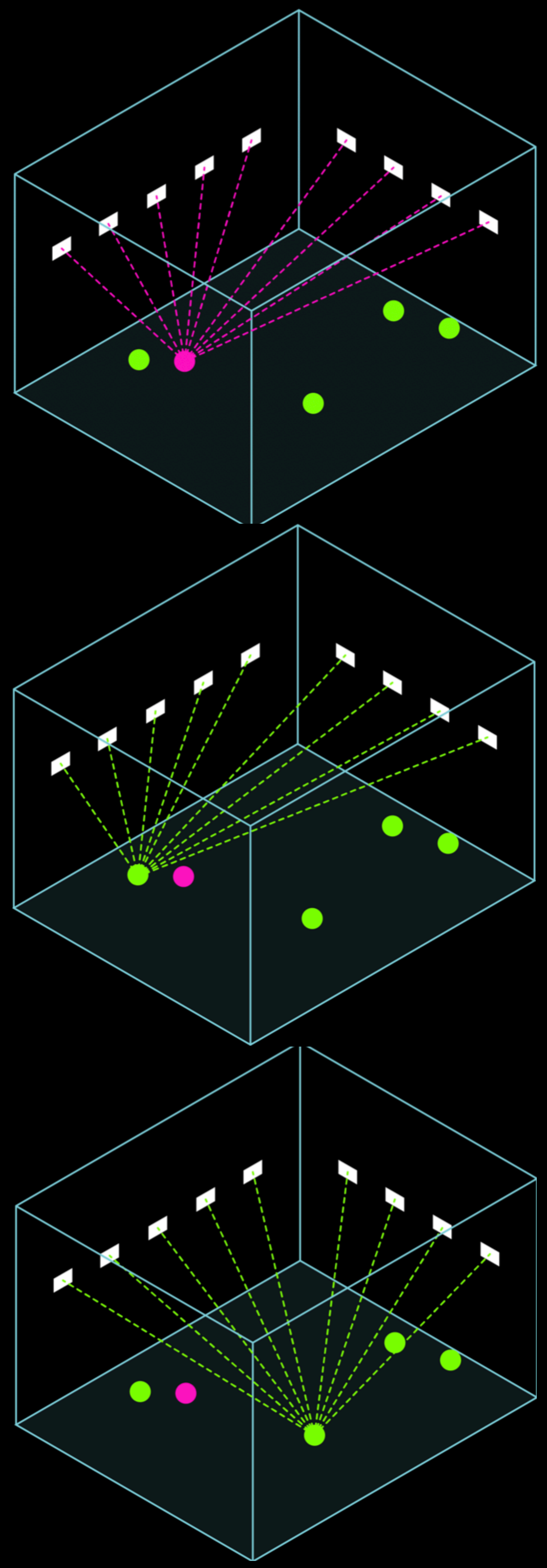
- Source separation
  - discussing project
  - delivering lecture
  - microphones



- # 1 INTRODUCTION

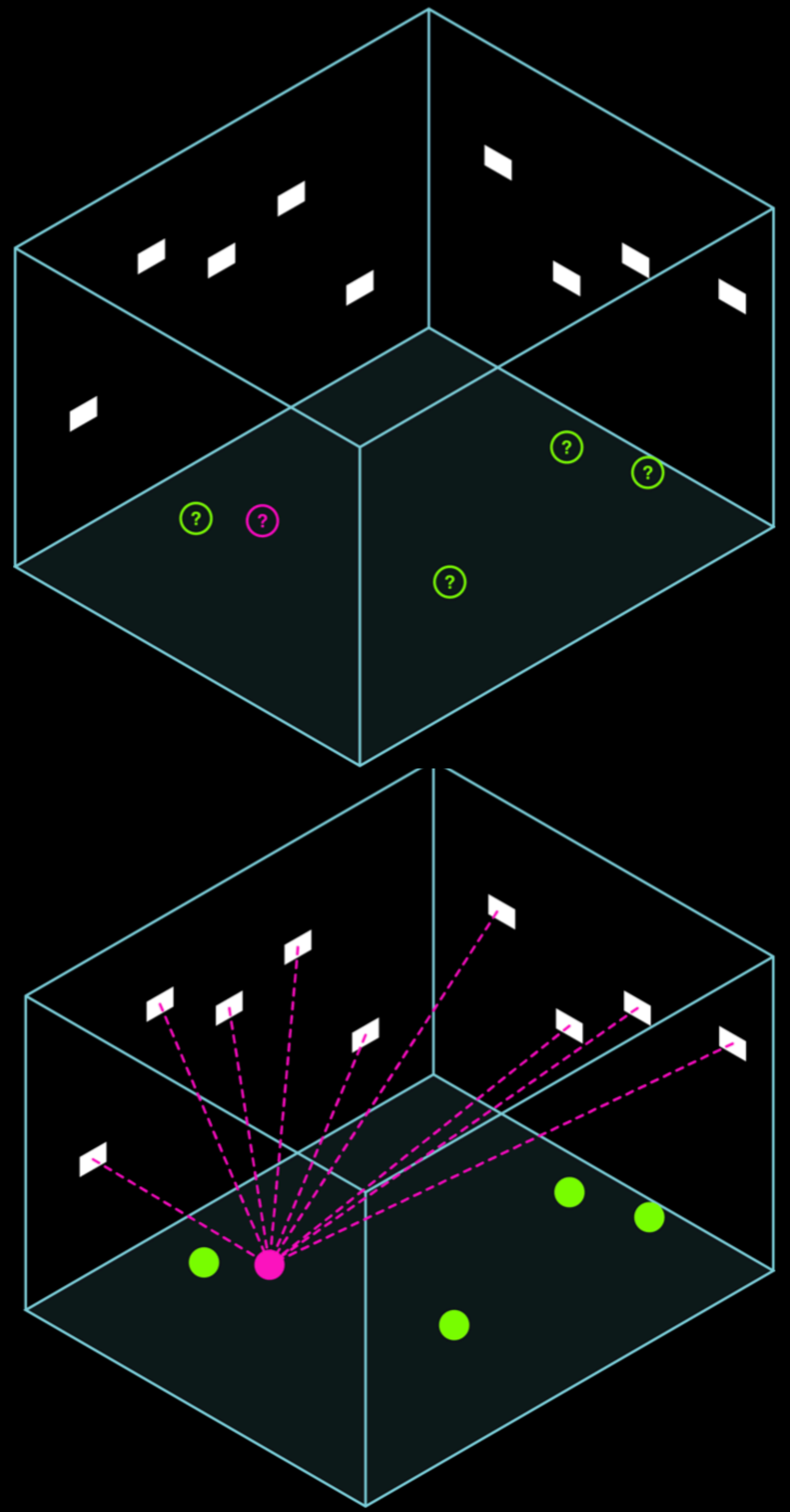
- could students in SV hear us clearly?
  - Beamforming?
  - Adaptive arrays?



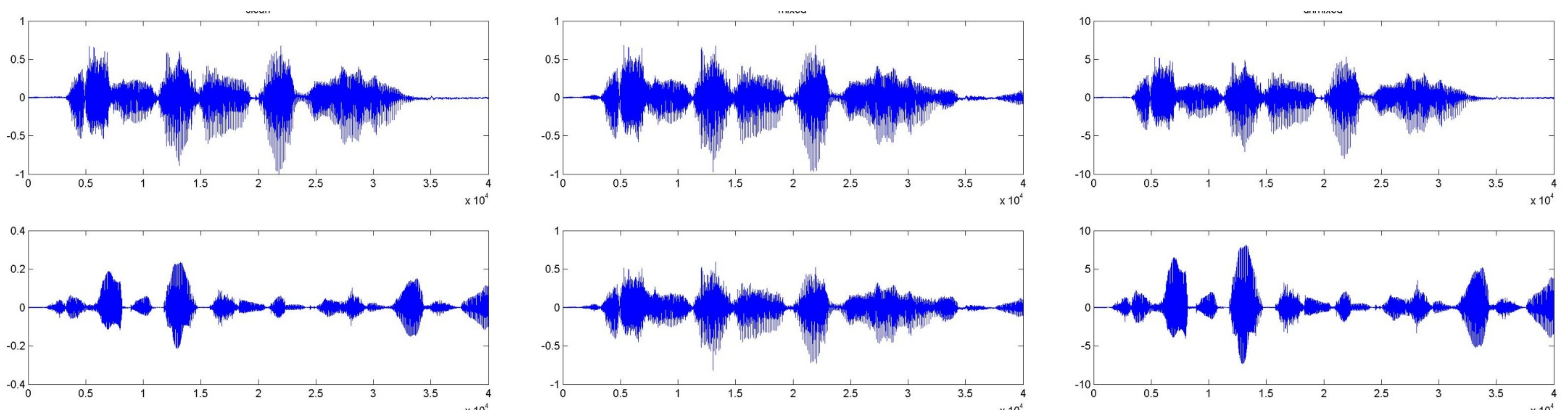


• 1 INTRODUCTION

- Should have known something on mixing processing and observation
  - Arrangements of microphones array
  - Direction of speaker
  - Time delay should be significant
- Take 18-792 Advanced Digital Signal Processing :-)



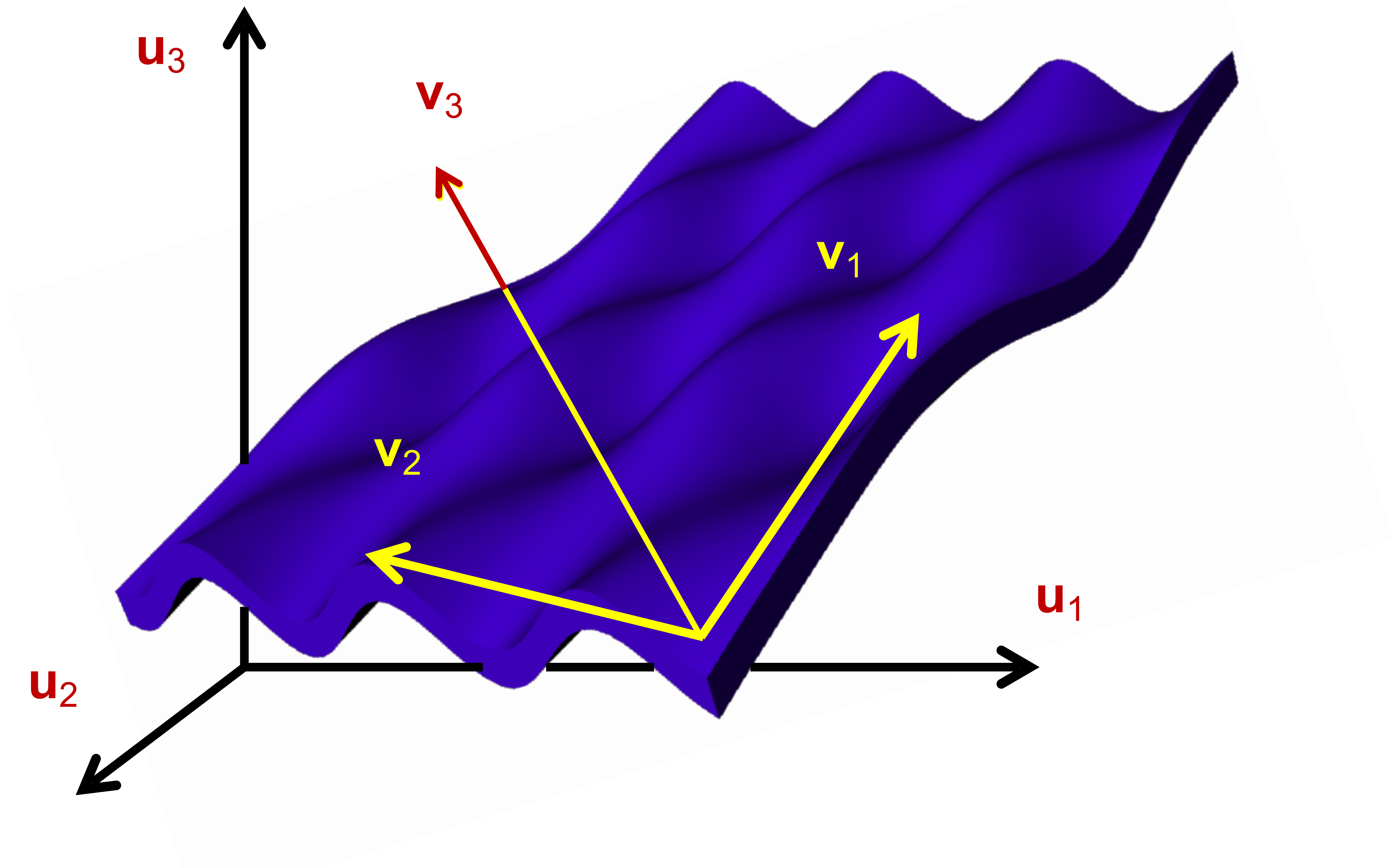
So how does other algorithm work?



- Example with instantaneous mixture of two speakers
- Natural gradient update
- Works very well!

Story so far
(and ahead)

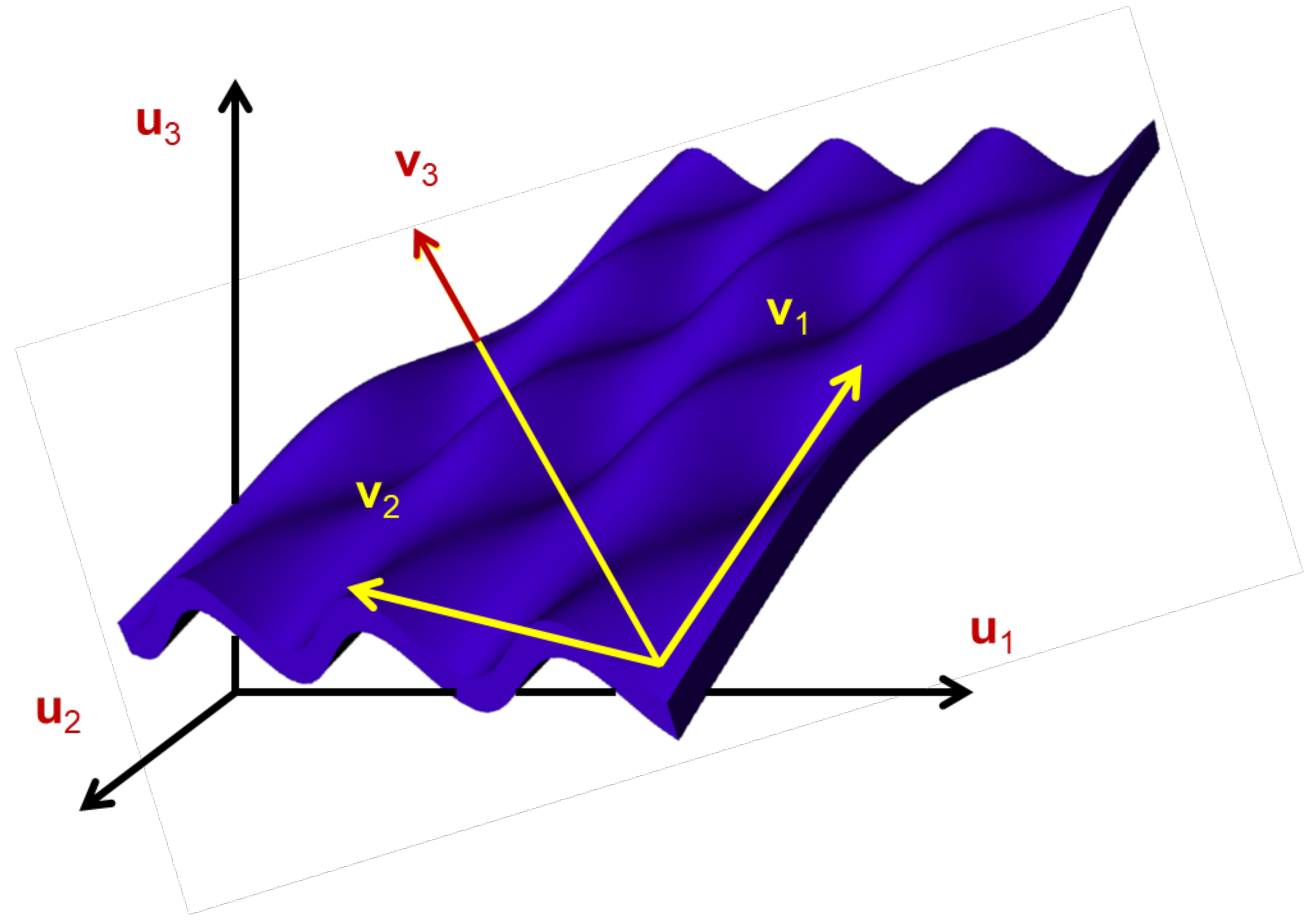
Capturing the structure in data

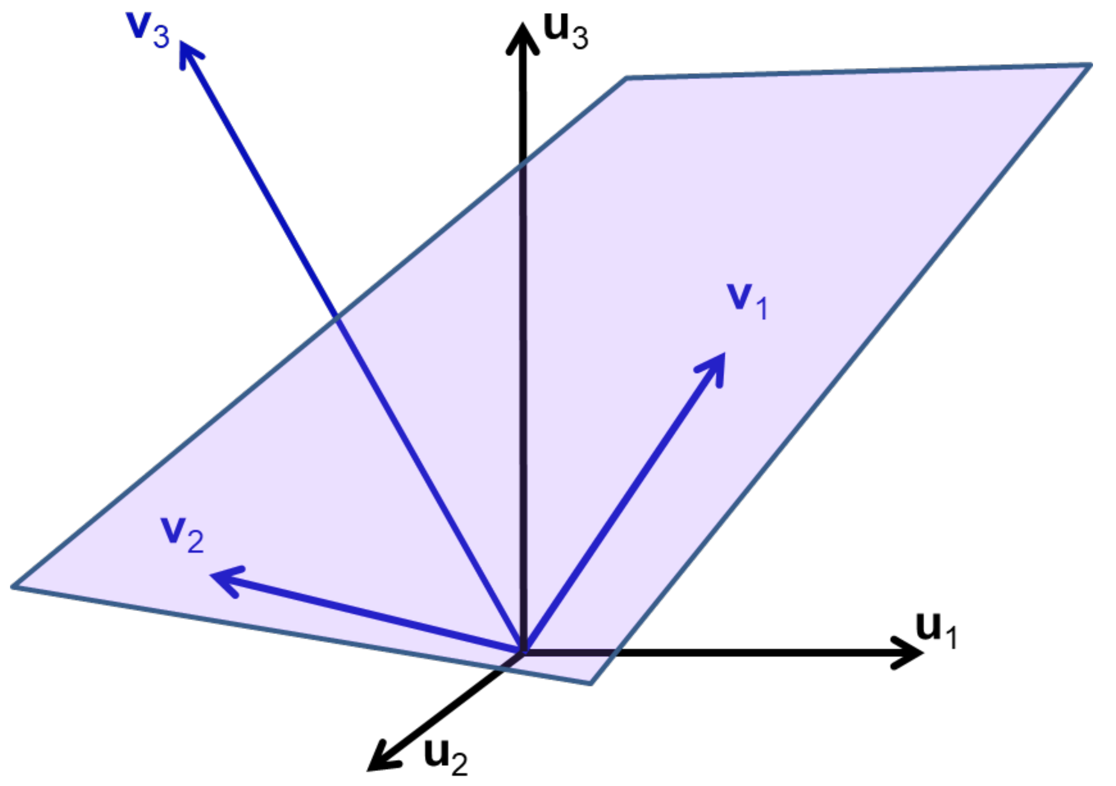


- *The most important challenge* in ML: Find the best set of bases for a given data set

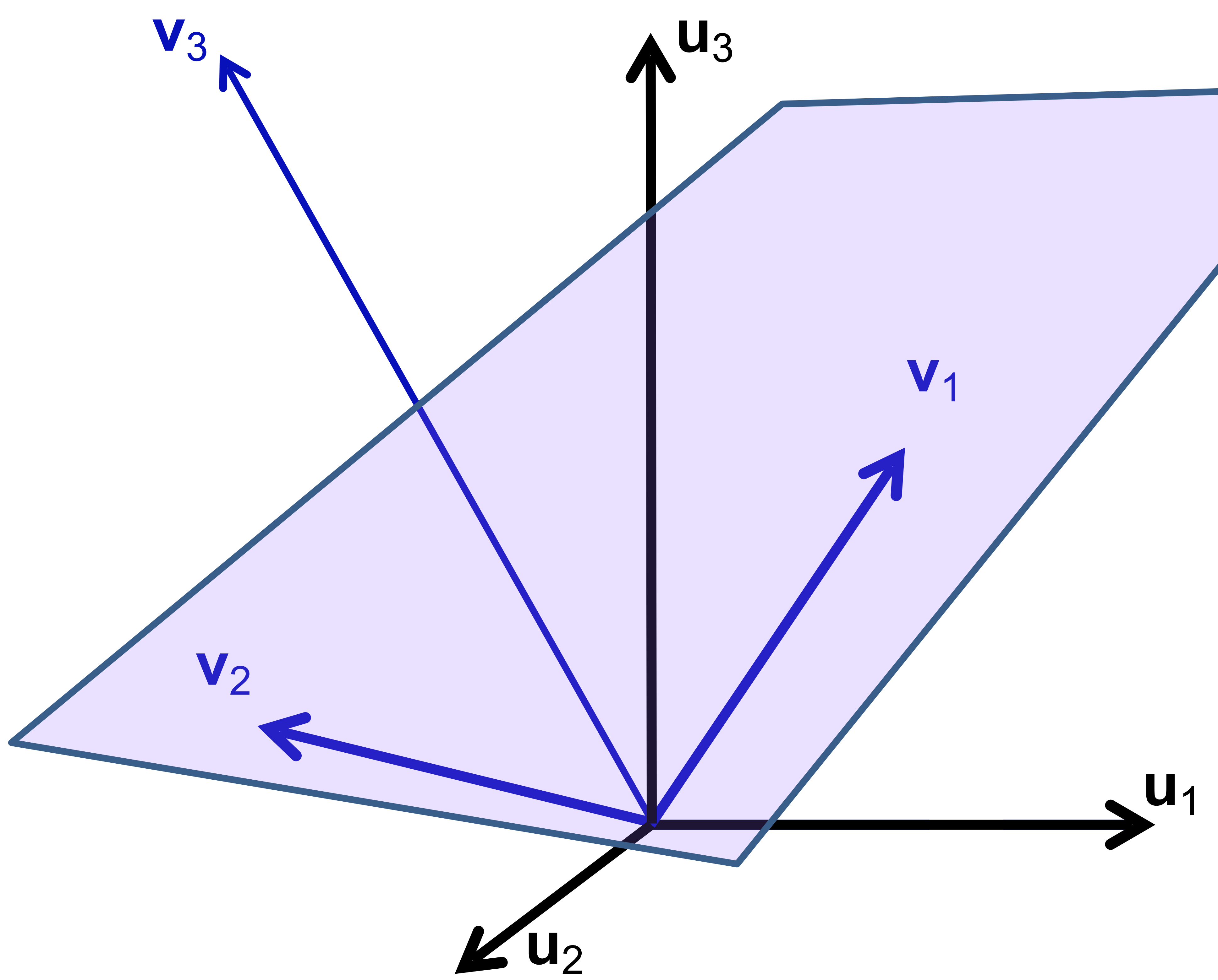
Capturing data structure

- Much of what we've done has attempted to find the underlying structure of the data
  - By analyzing the data itself
- We have assumed a linear structure
  - The data lie primarily on a linear subspace or manifold
  - Variations off the manifold are fine detail that may just be noise
- Linear models get 90% of the way
  - But the math is extendable to non-linear manifolds, though we won't go into it much, in this class





The Linear Model



- Find the best set of bases $v_1, v_2, \ldots v_D$ plane
  - Given only a collection of data points x_1, x_2, … x_N

The linear model

- Given the bases $\boldsymbol{B} = [b_1 b_2 \ldots b_D]$, for any vector $\boldsymbol{x}$: $\boldsymbol{x} \approx \boldsymbol{B}\boldsymbol{w}$
- Our actual problem: Given a collection of N vectors $\boldsymbol{x}_1,\ \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, find $\boldsymbol{B}$ and $\boldsymbol{w}_1,\ \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N$ such that

$$\left[\boldsymbol{x}_1,\ \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D\right] \approx \boldsymbol{B}[\boldsymbol{x}_1,\ \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$$

$$\boldsymbol{X} \approx \boldsymbol{B}\boldsymbol{W}$$

- So, algebraic problem, given $\boldsymbol{X}$, find $\boldsymbol{B}$ and $\boldsymbol{W}$ such that $\boldsymbol{X} \approx \boldsymbol{B}\boldsymbol{W}$ as closely as possible

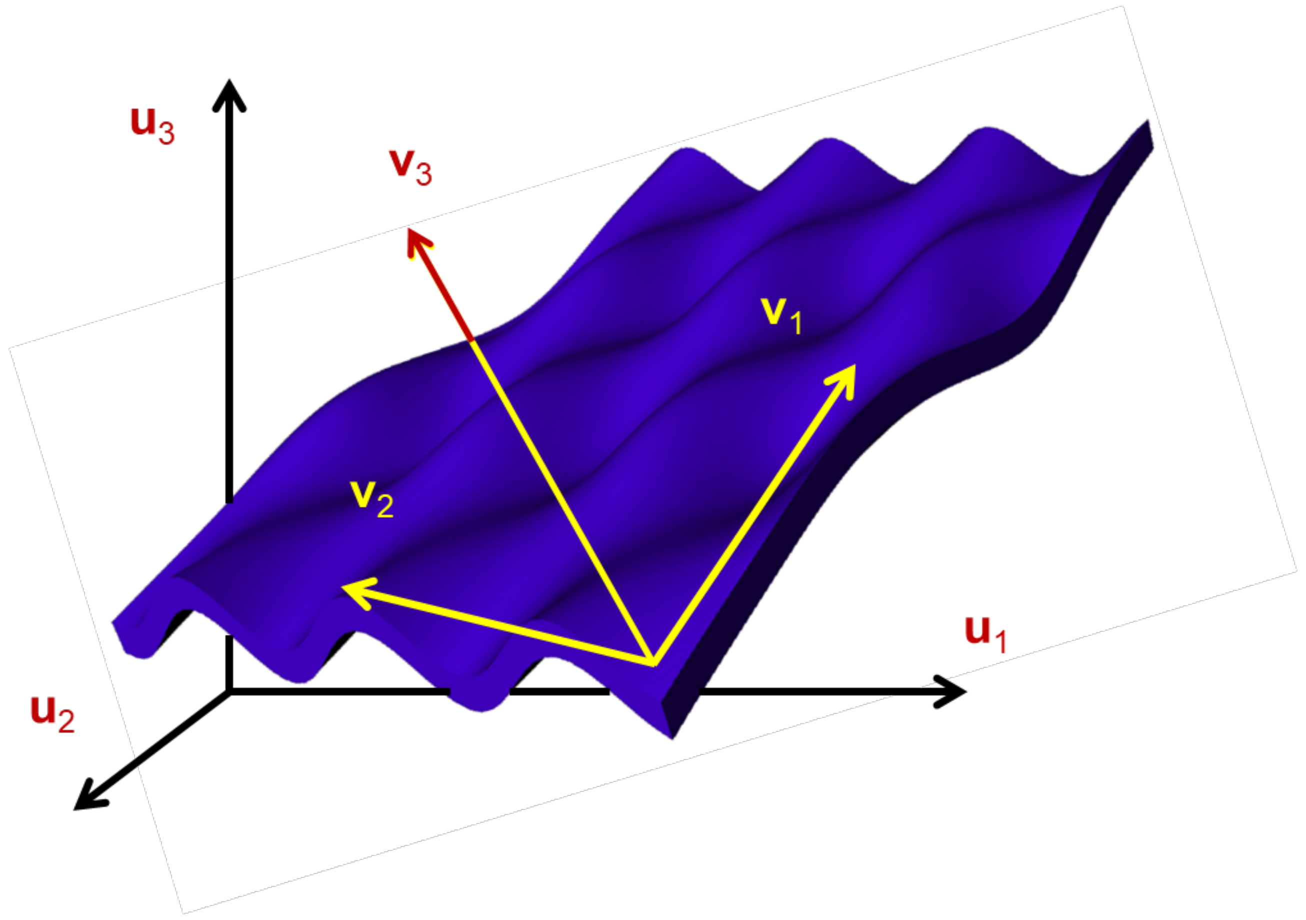
Views of the decomposition

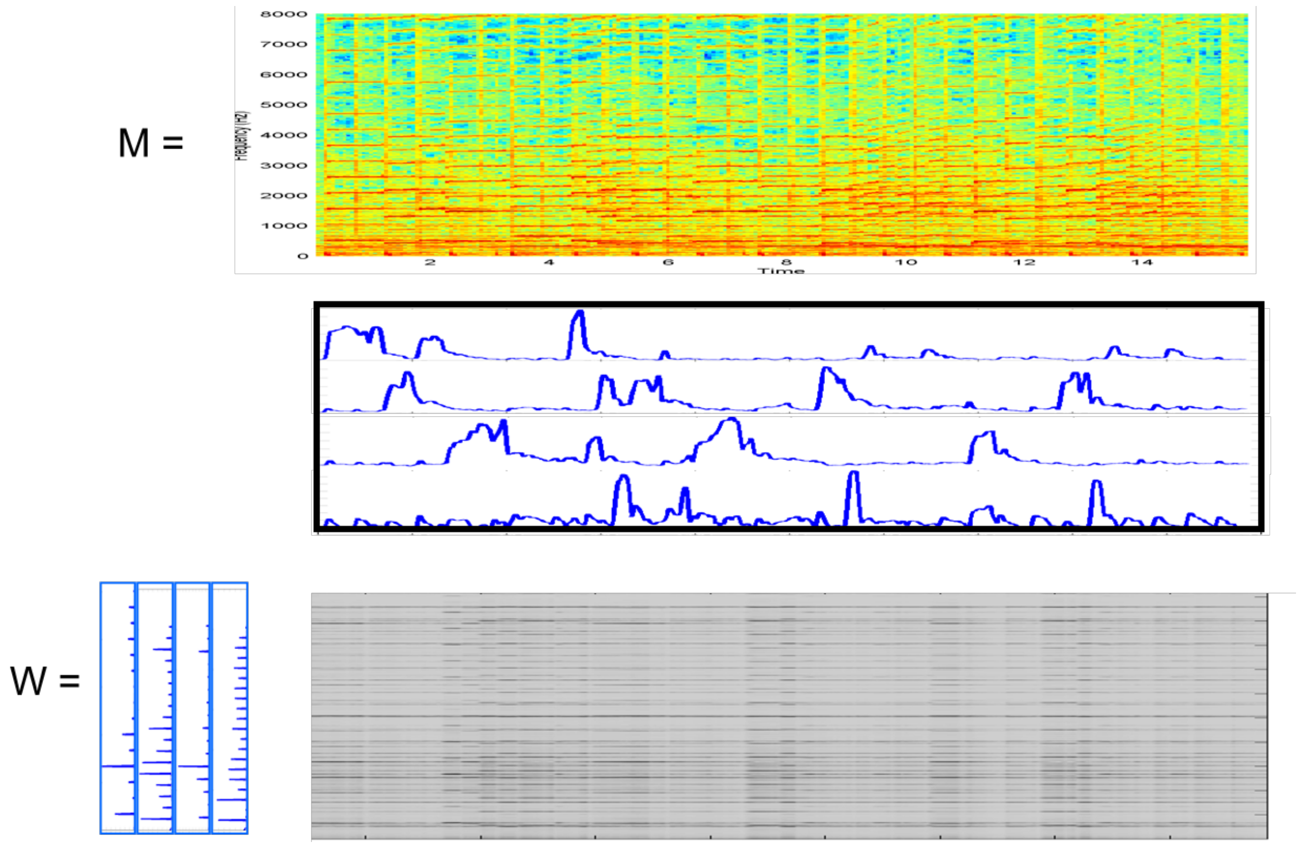
- $x \approx Bw \; (X \approx BW)$

- B are the *bases* of the subspace where most of the information about the data lies
  - w are the coordinates of the instance in these bases

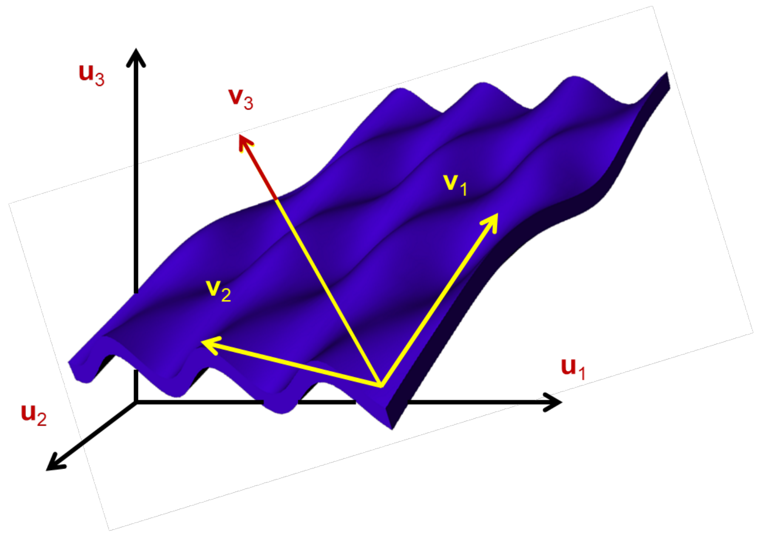
- B are the *building blocks* that compose the data
  - w are the *mixing weights* with which the building blocks are combined to compose the instance

- Believe it or not, the two perspectives are interchangeable

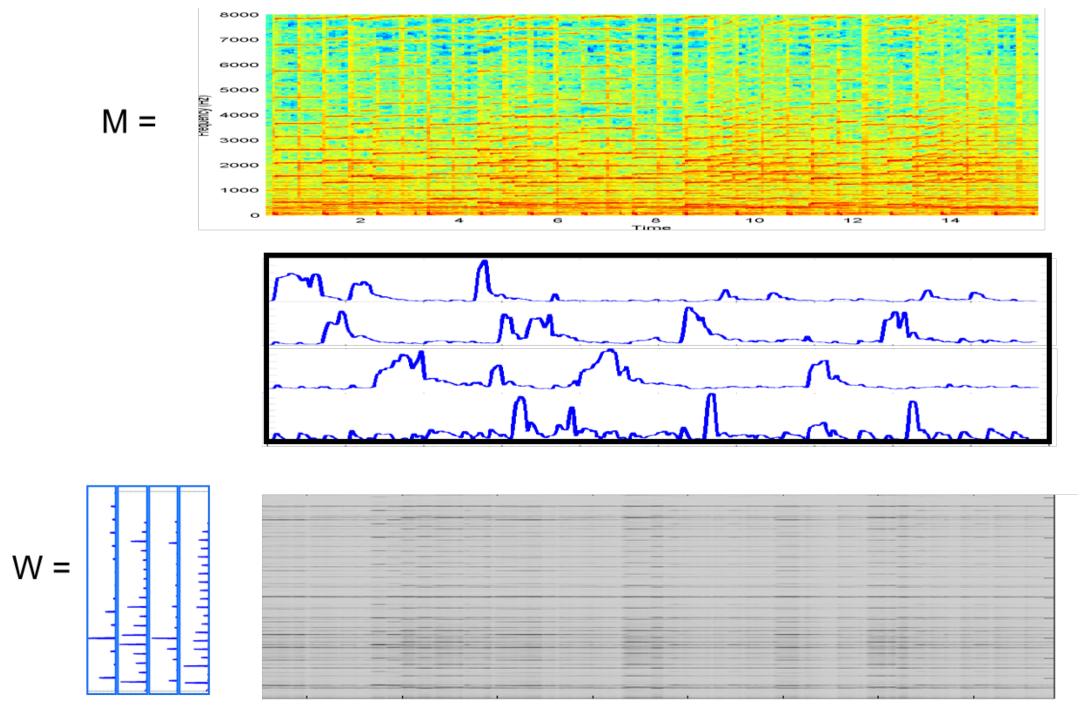




But what is a good decomposition?



$$X \approx BW$$



- Minimum error criterion: The error between X and BW is minimized
  - KLT
- Statistical criteria:
  - The rows of W (the components of w) are uncorrelated
    - PCA
  - The rows of W are statistically independent
    - ICA
- Physics-motivated:
  - The building blocks combine in a purely constructive way
    - NMF
  - The number of building blocks can be very large (much larger than the dimensionality of the data)
    - Dictionary-based representations

• 1 INTRODUCTION

Source sepatation as finding independent components

- Some notations
  - Sources: $S = \left\{s_1, s_2, \cdots, s_N\right\}$
  - Observations: $X = \left\{x_1, x_2, \cdots, x_N\right\}$
  - Given $X = A\left(S\right)$, where A represents mixing process
- Find a inverse function $W \approx A^{-1}$ such that $S \approx A^{-1}\left(X\right)$

• 1 INTRODUCTION

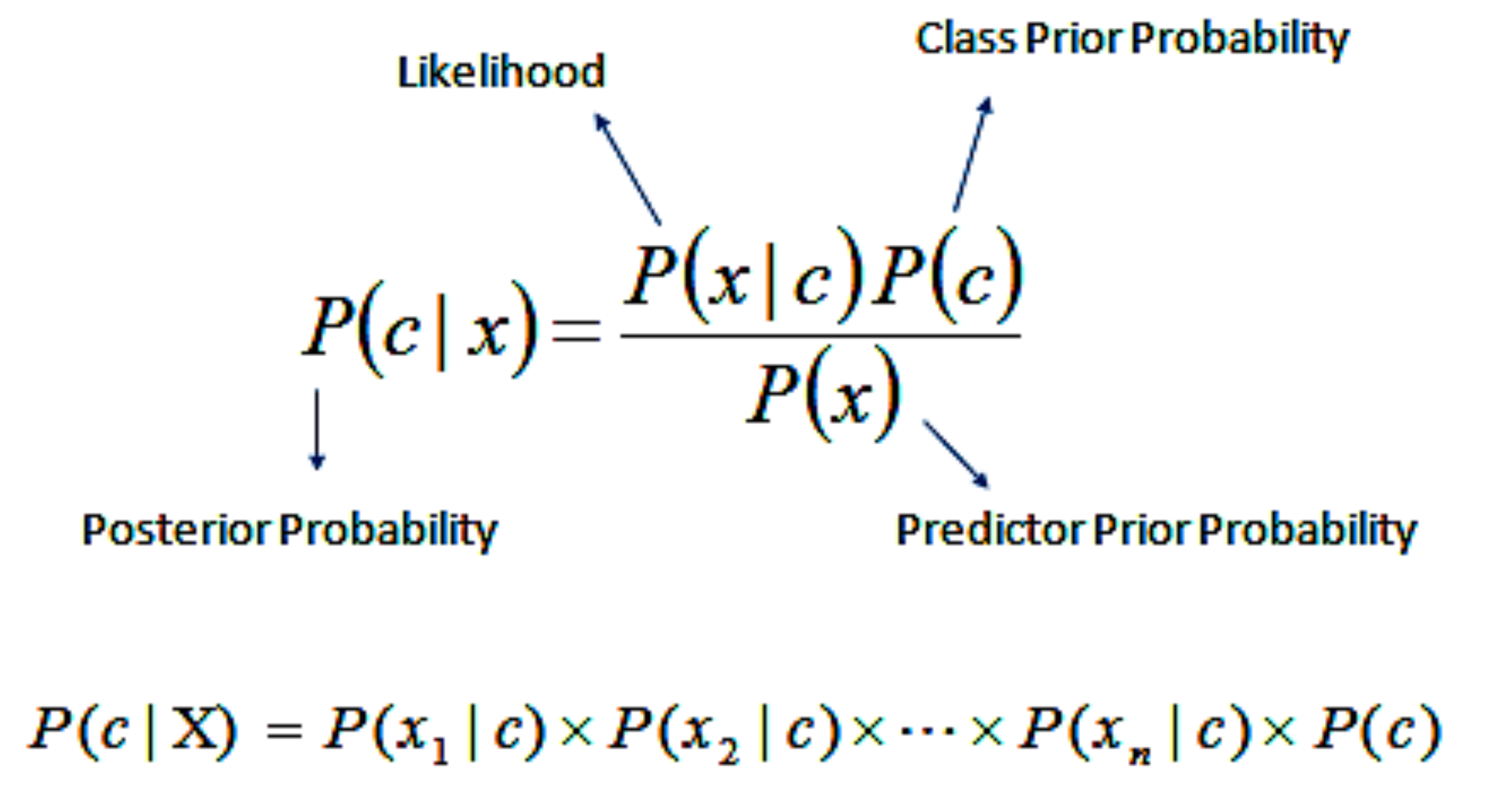
Source sepatation as finding independent components

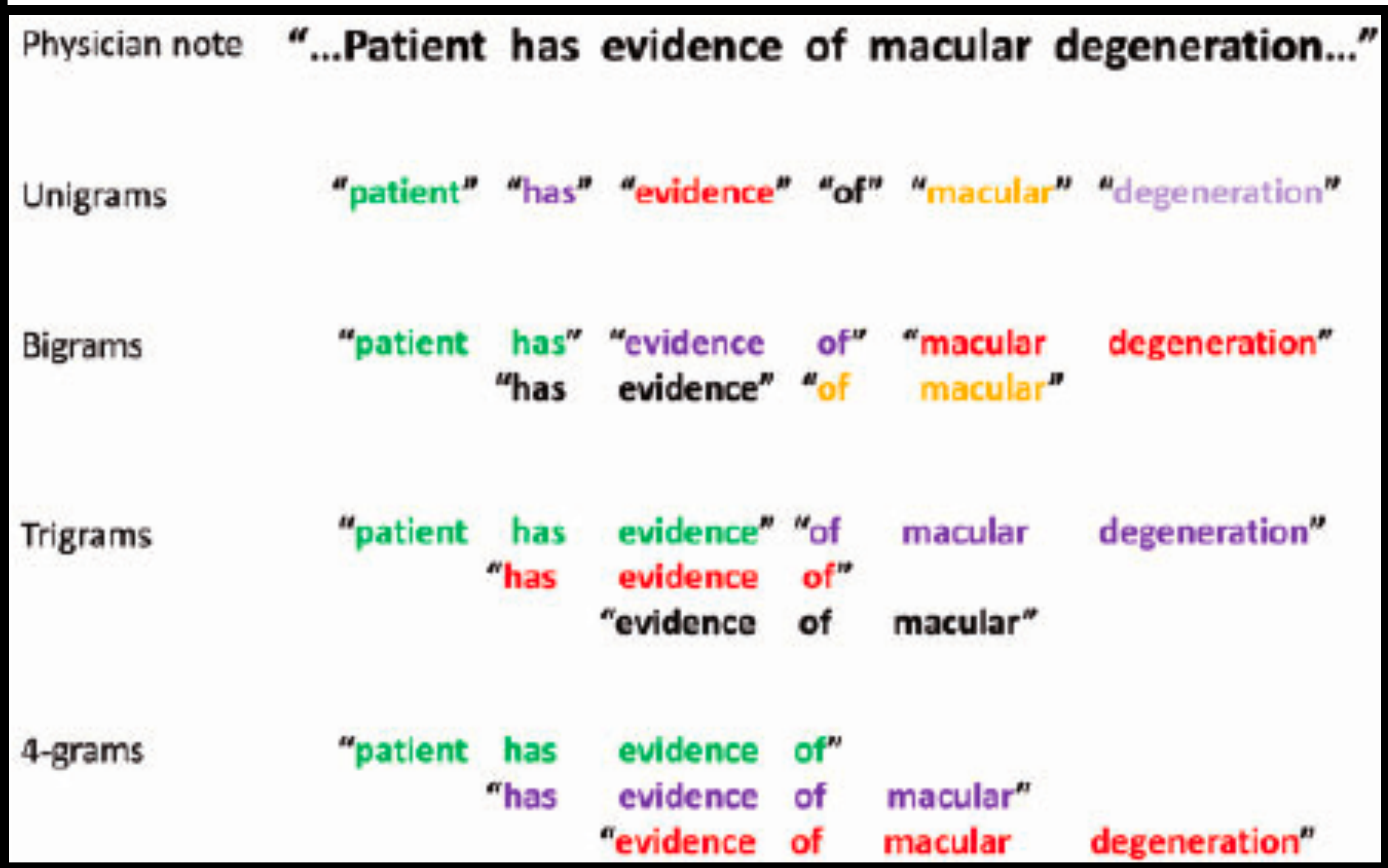
- $X = A(S)$ is BLIND source separation
    - BLIND: Know nothing about the mixing procedure A
- Quite difficult, need some assumptions on S and A, to make life easier
- For example, assumption on S is uncorrelated, what will happen?

• 1 INTRODUCTION

Recall the advantage of independence

- Uncorrelation of variables is generally considered desirable for modelling and analyses
  - Sometimes it can reduce the number of model parameters
  - Sometimes it is not practical to assume independence / uncorrelation
- We could transform correlated variables to make them uncorrelated in some cases





• 1 INTRODUCTION

Assumption for Blind Source sepatation

- (independent) source separation tasks aims to demix the observation to independent components
- (linear) $\begin{cases} x_1 = a_{11}s_1 + a_{12}s_2 \cdots + a_{1N}s_N \\ x_2 = a_{21}s_1 + a_{22}s_2 \cdots + a_{2N}s_N \\ \quad \vdots \\ x_N = a_{N1}s_1 + a_{N2}s_2 \cdots + a_{NN}s_N \end{cases}$ or $X = AS$

- # 1 INTRODUCTION

How to measure independence?

- What is tue virtue (specific excellence) of independent variable?

• 1 INTRODUCTION

How to measure independence

- Source should have higher-order statistics properties instead of only $E\left[S_1 S_2\right] = E\left[s_1\right] E\left[s_2\right]$ like PCA on tensorial decompositions
    - FOBI-ICA algorithm, JASE-ICA algotithm
- Source should be less Gaussian compared with observation
    - Fast-ICA algotithm

- # 1 INTRODUCTION

Using higher-order statistics properties to measure independence

- Source should have higher-order statistics properties instead of only $E\left[S_1 S_2\right] = E\left[s_1\right] E\left[s_2\right]$ like PCA on tensorial decompositions
  - $E\left[s_1 s_2 s_3 s_4\right] = E\left[s_1\right] E\left[s_2\right] E\left[s_3\right] E\left[s_4\right]$
  - $E\left[s_1^2 s_2 s_3\right] = E\left[s_1^2\right] E\left[s_2\right] E\left[s_3\right]$
  - $E\left[s_1^2 s_2^2\right] = E\left[s_1^2\right] E\left[s_2^2\right]$
  - $E\left[s_1^3 s_2\right] = E\left[s_1^3\right] E\left[s_2\right]$

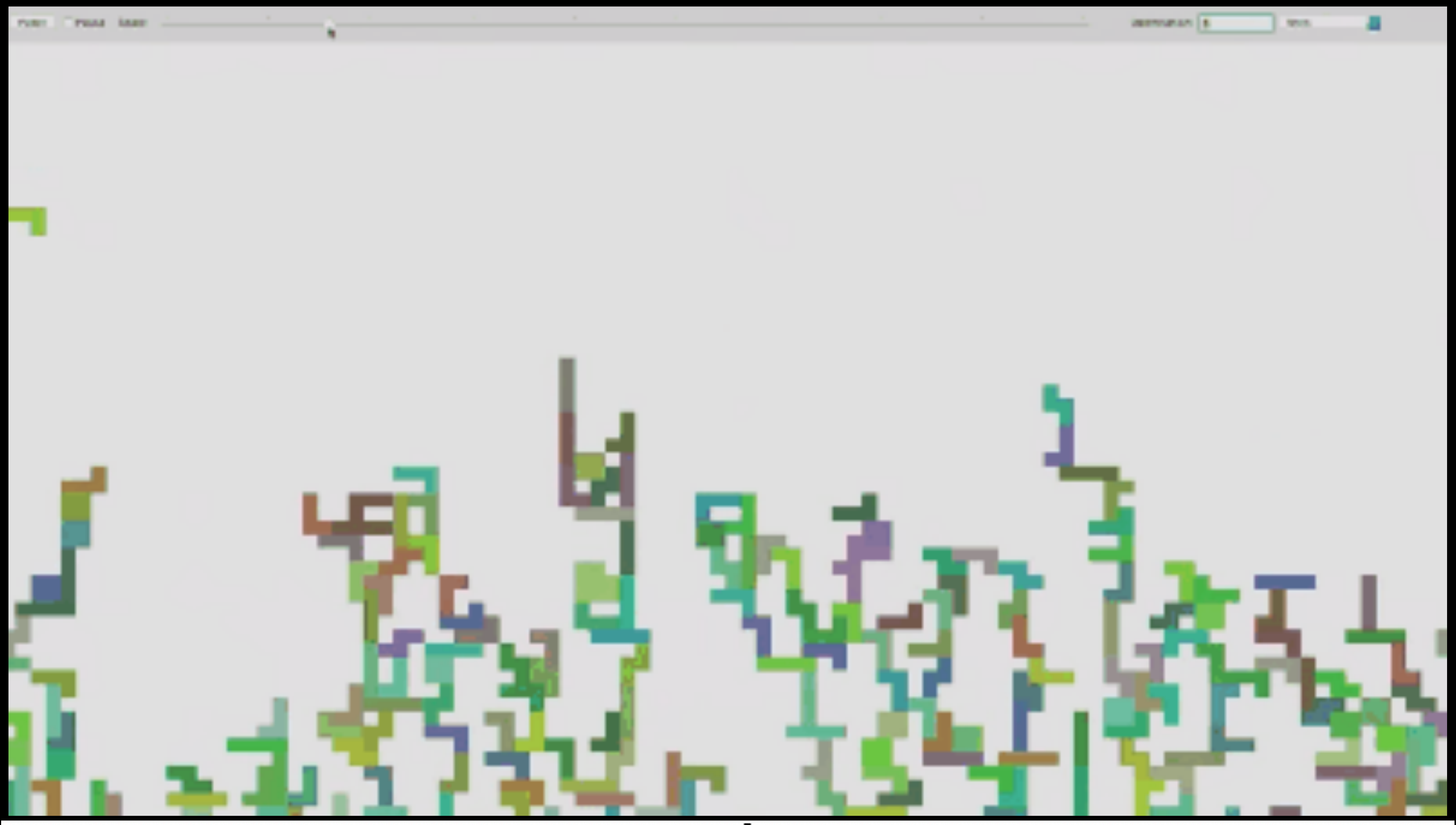
• 1 INTRODUCTION

Using higher-order statistics properties to measure independence

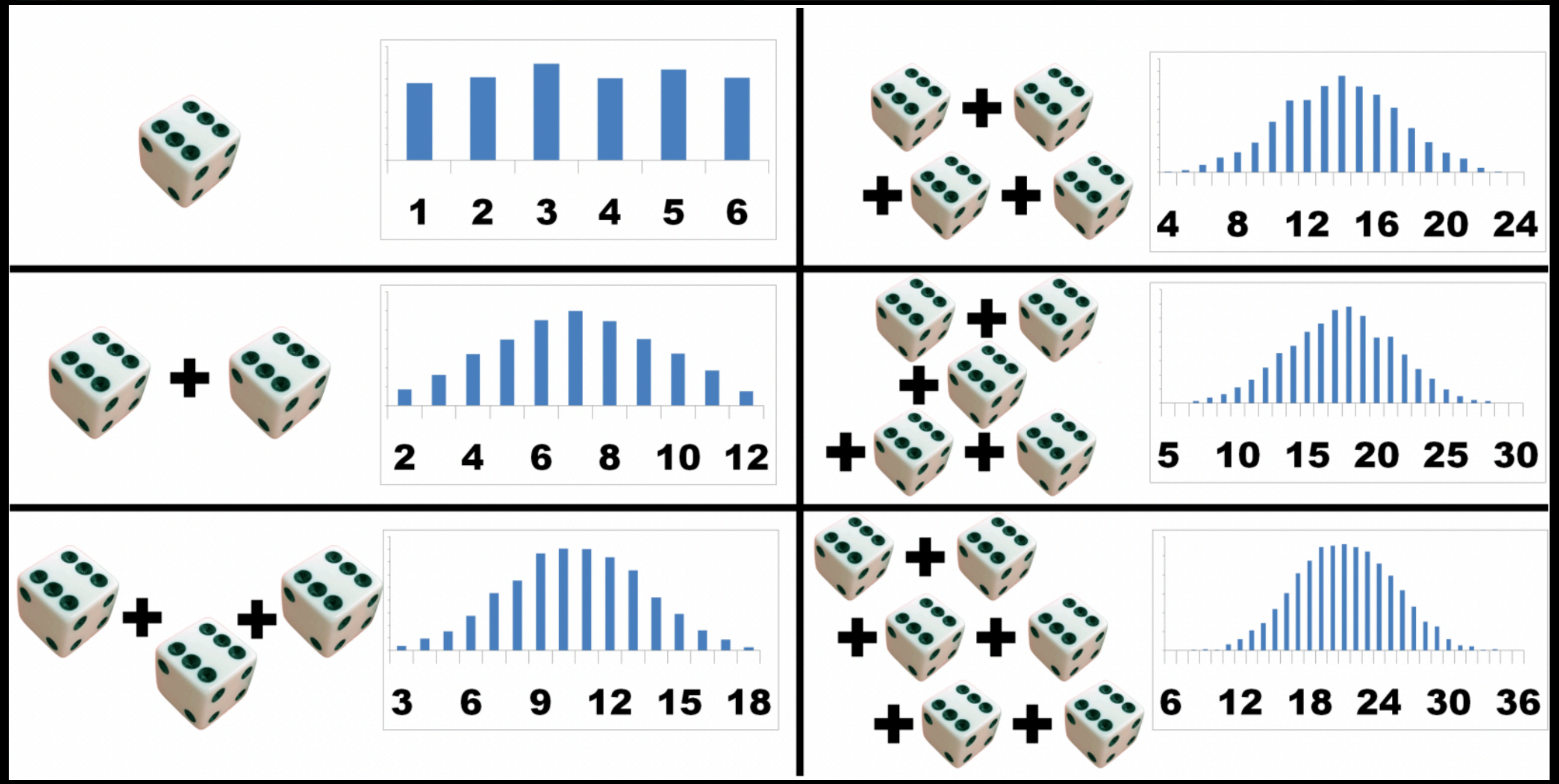
- Source should have higher-order statistics properties instead of only $E\left[S_1 S_2\right] = E\left[s_1\right] E\left[s_2\right]$ like PCA on tensorial decompositions
- We will use this high order moment to solve linear ICA
- While, let's see another measure of independence at first

• 1 INTRODUCTION

Difference between independent components and their mix



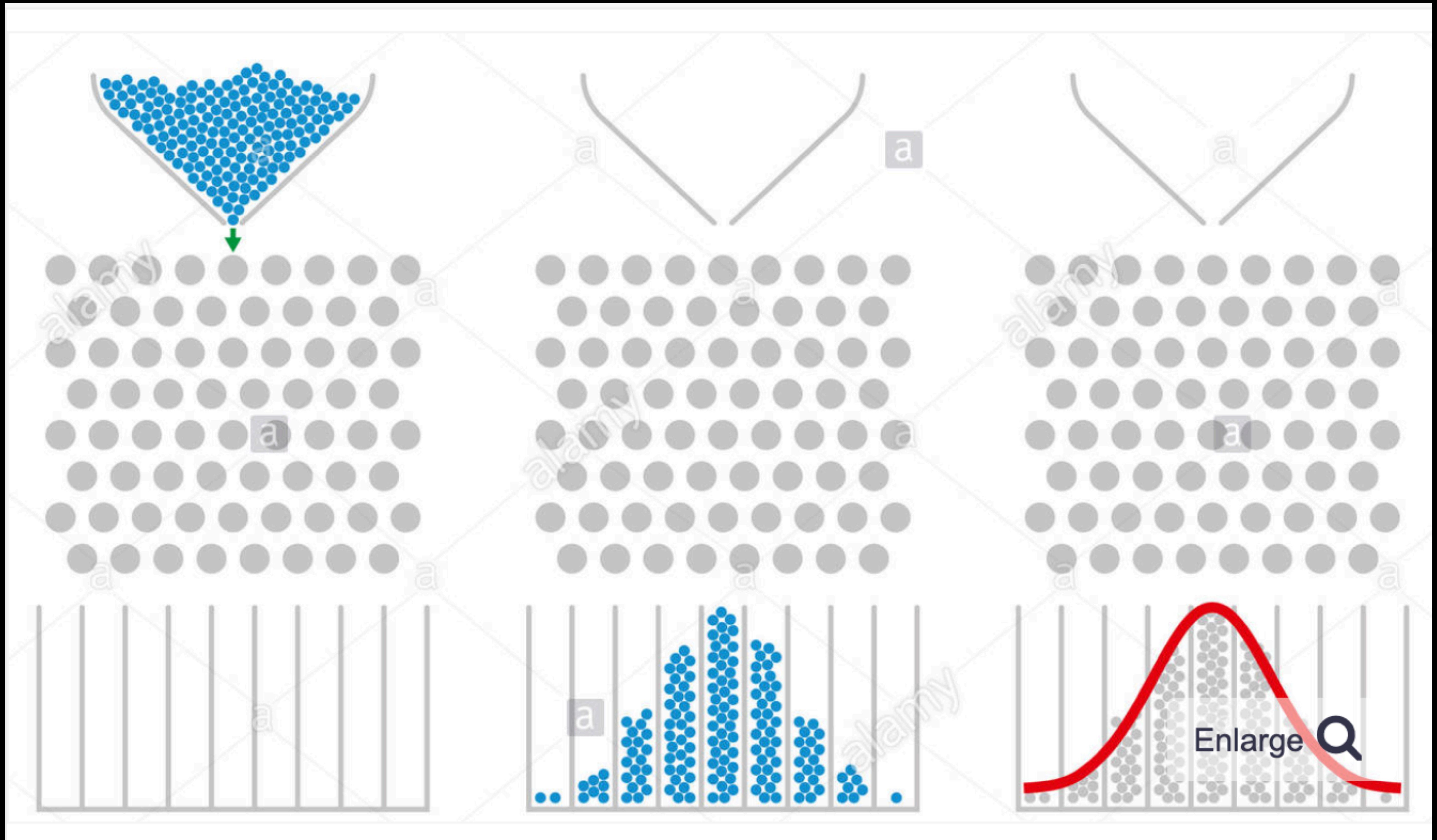
- UNIVERSALITY behind micro independent components
  - example: KPZ function behind tetris
  - example: Center Limit Theorem implies Gaussian distribution behiend any set of “not bad” independent random variable



Center Limit Theorem

- # 1 INTRODUCTION

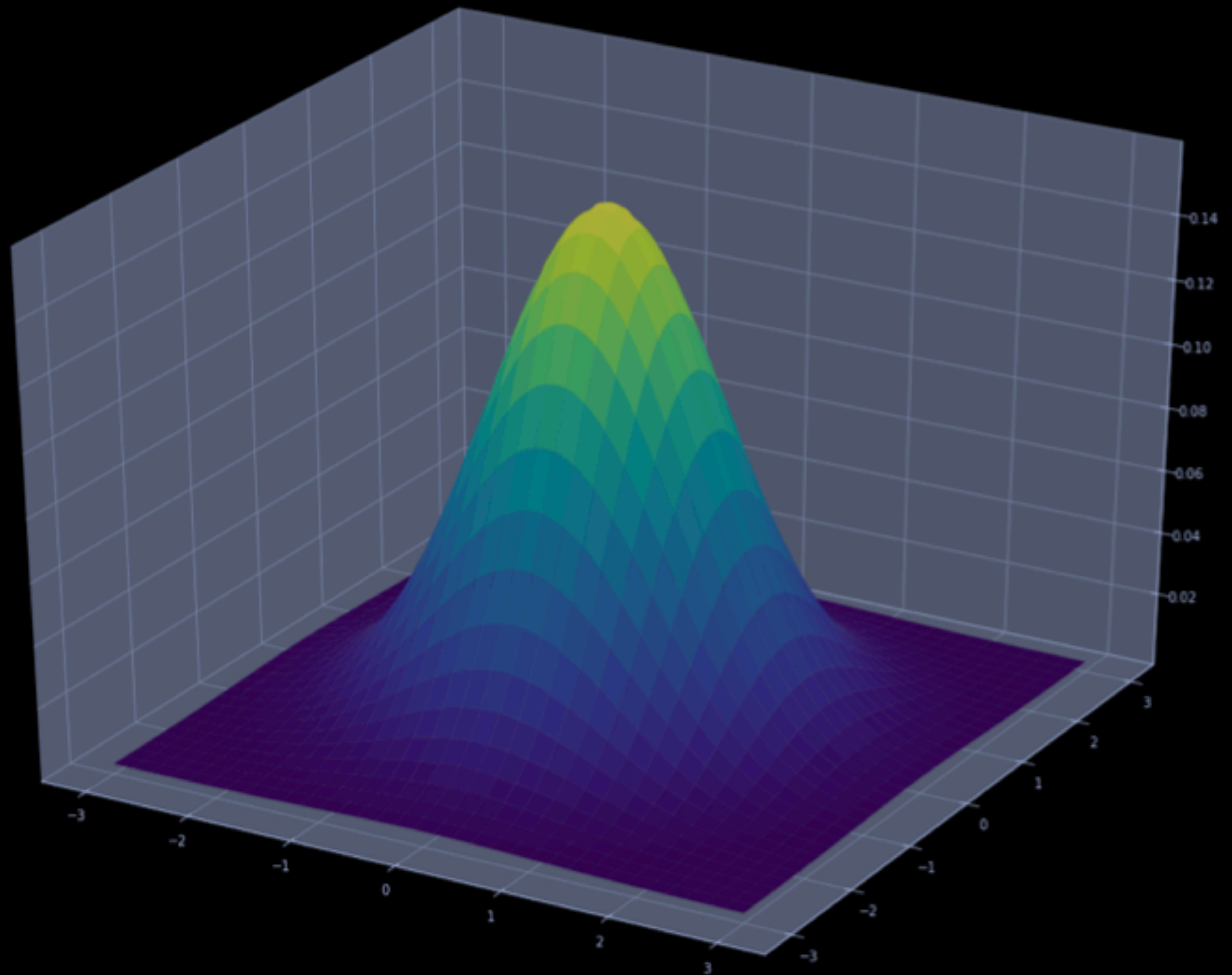
Intuition: source should be “less Gaussian” than mixed signal



• 1 INTRODUCTION

Intuition: source should be “less Gaussian” than mixed signal

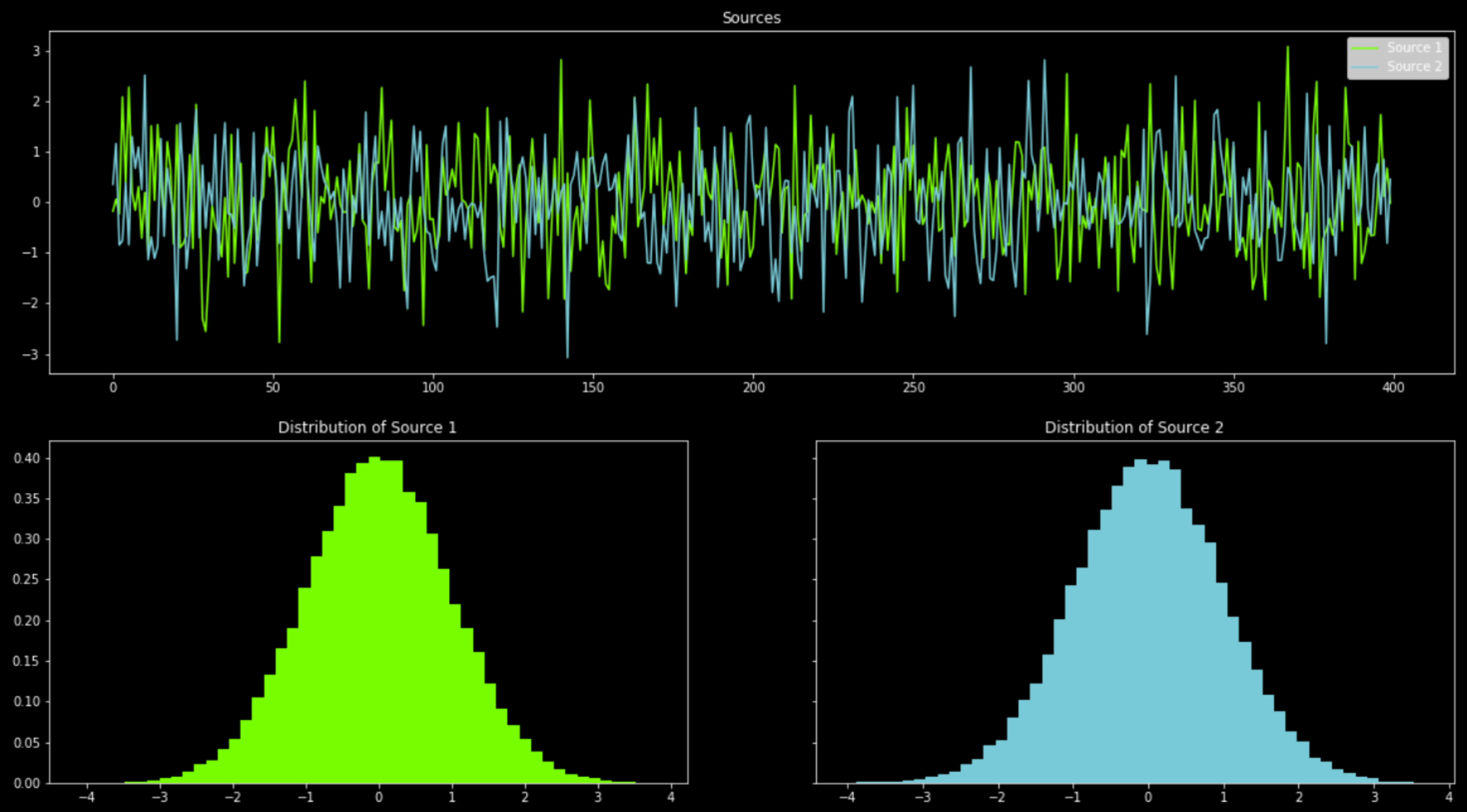
- S is “less Gaussian” and $X = AS$ could be “more Gaussian”



• 1 INTRODUCTION

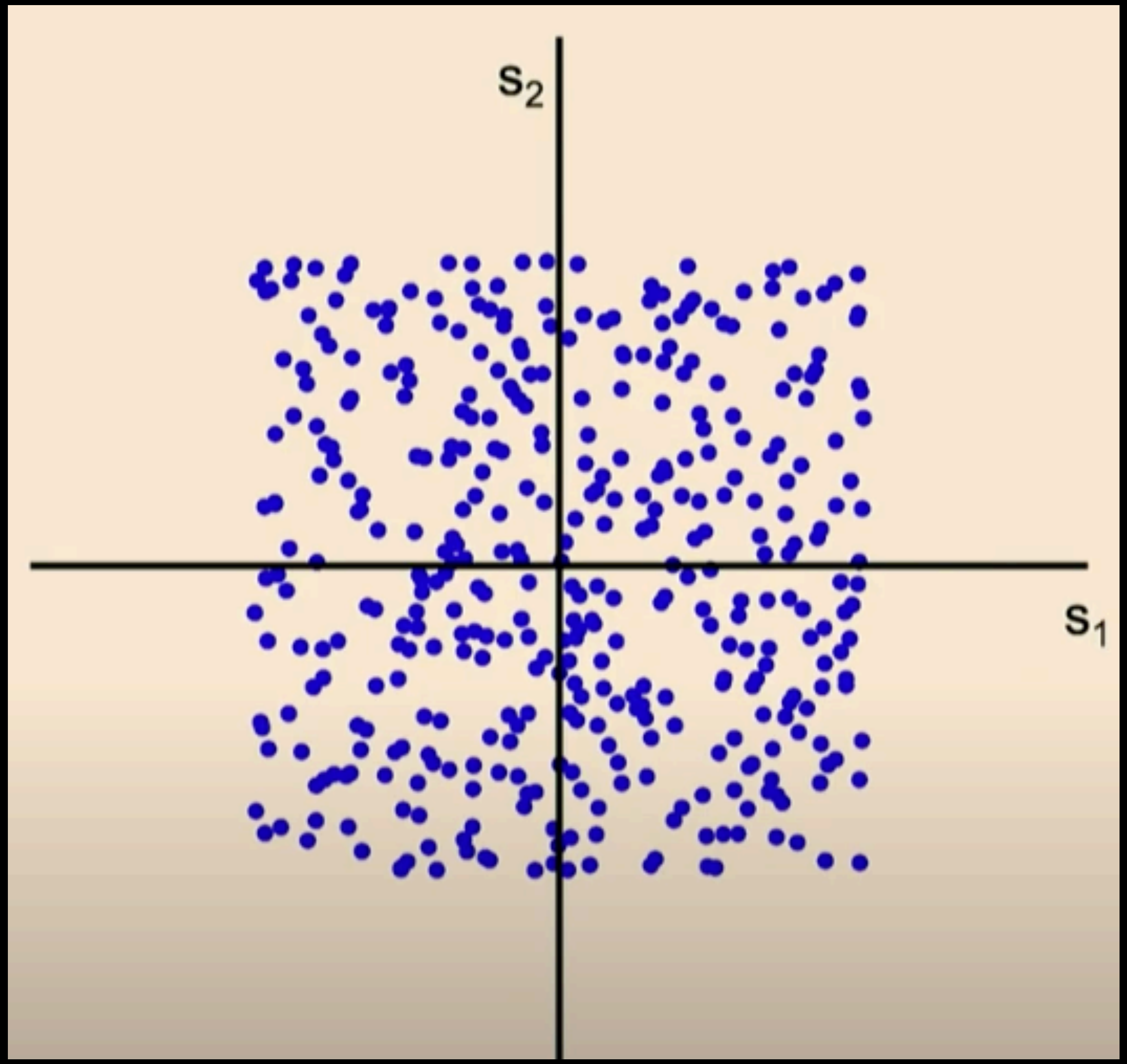
Intuition: source should be “less Gaussian” than mixed signal

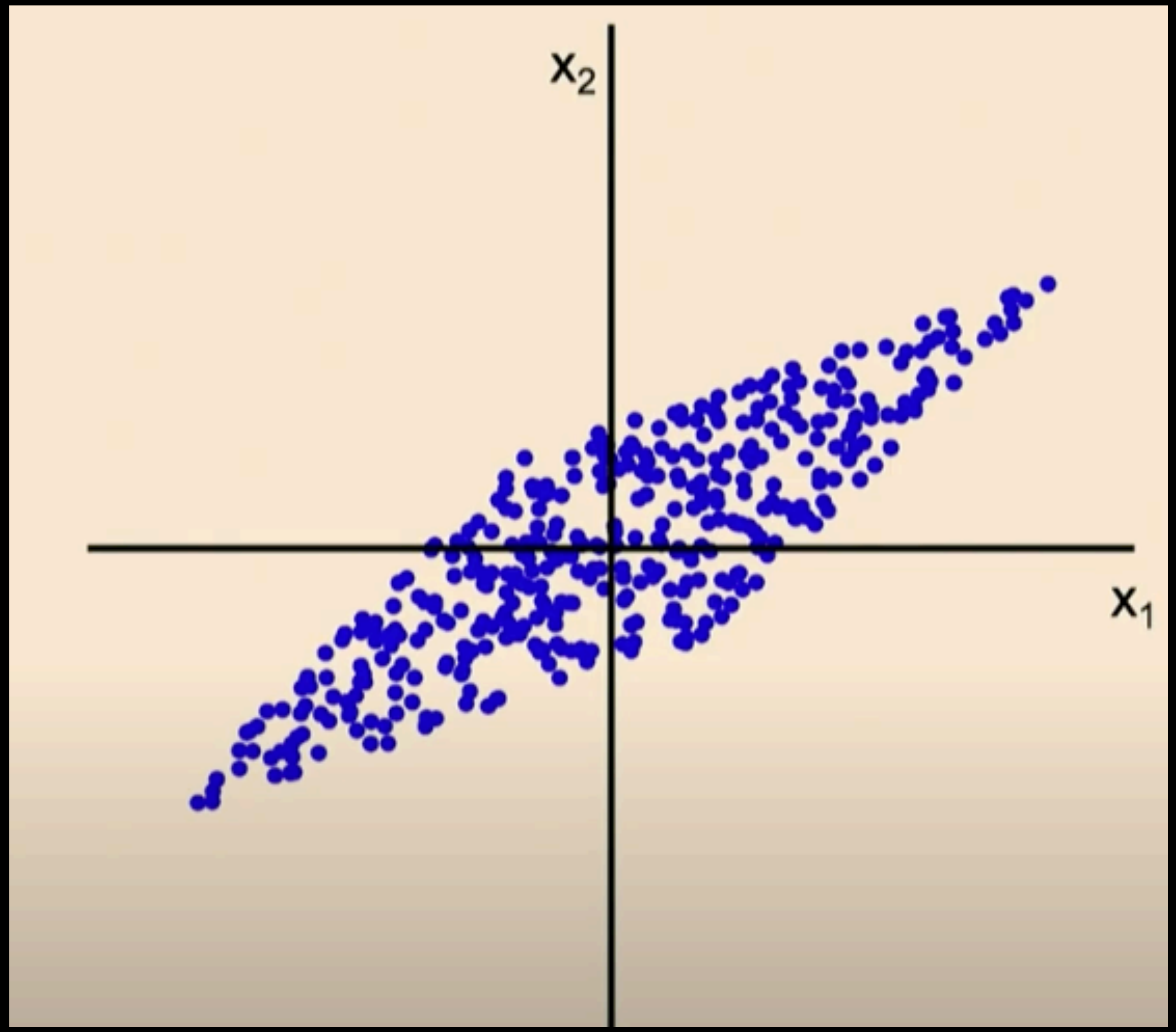
- What if S itself is Gaussian?



• 1 INTRODUCTION

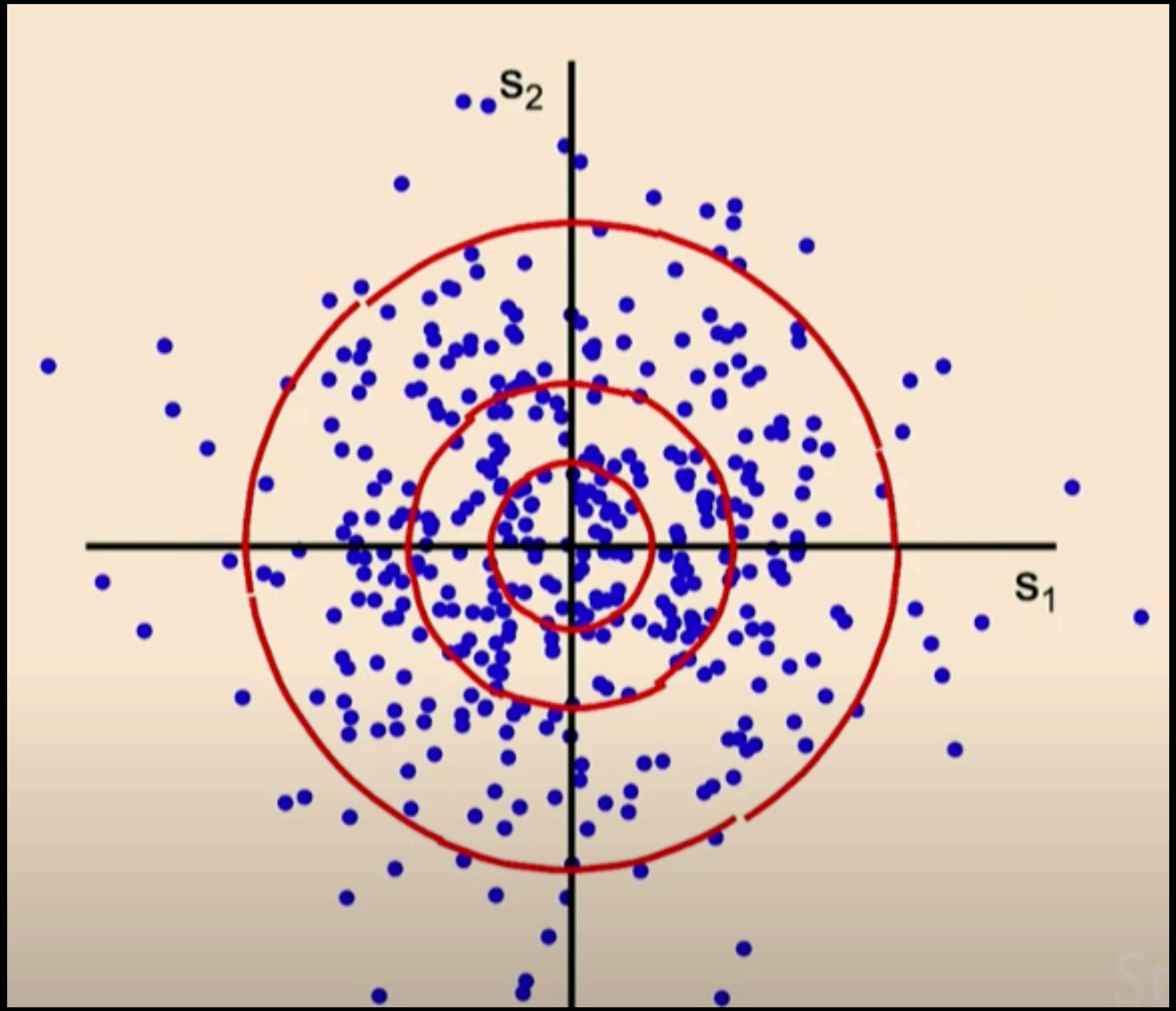
- in geometry, non-Gaussian source





- # 1 INTRODUCTION

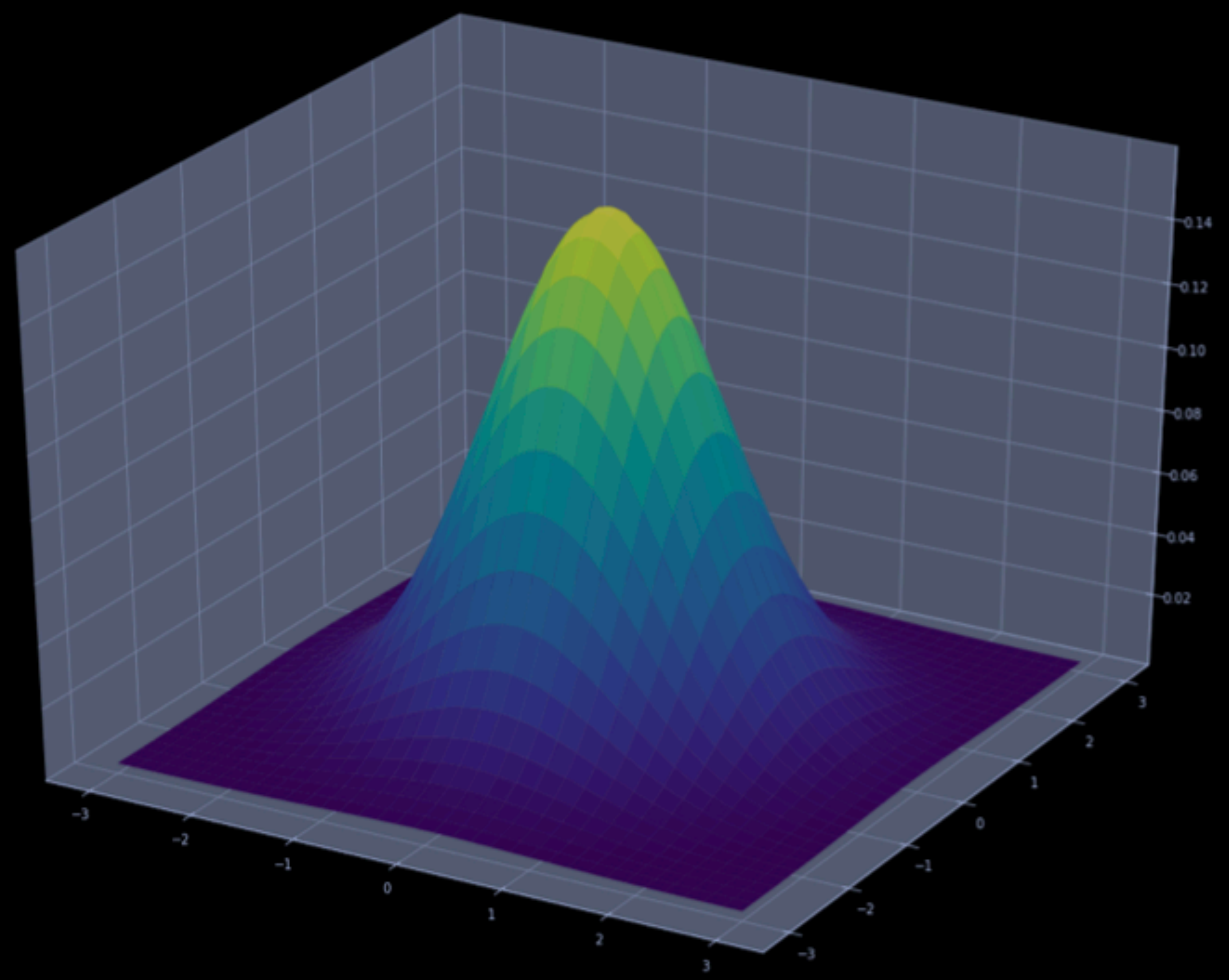
- in geometry, Gaussian source



• 1 INTRODUCTION

Case of Gaussian source shall be omitted

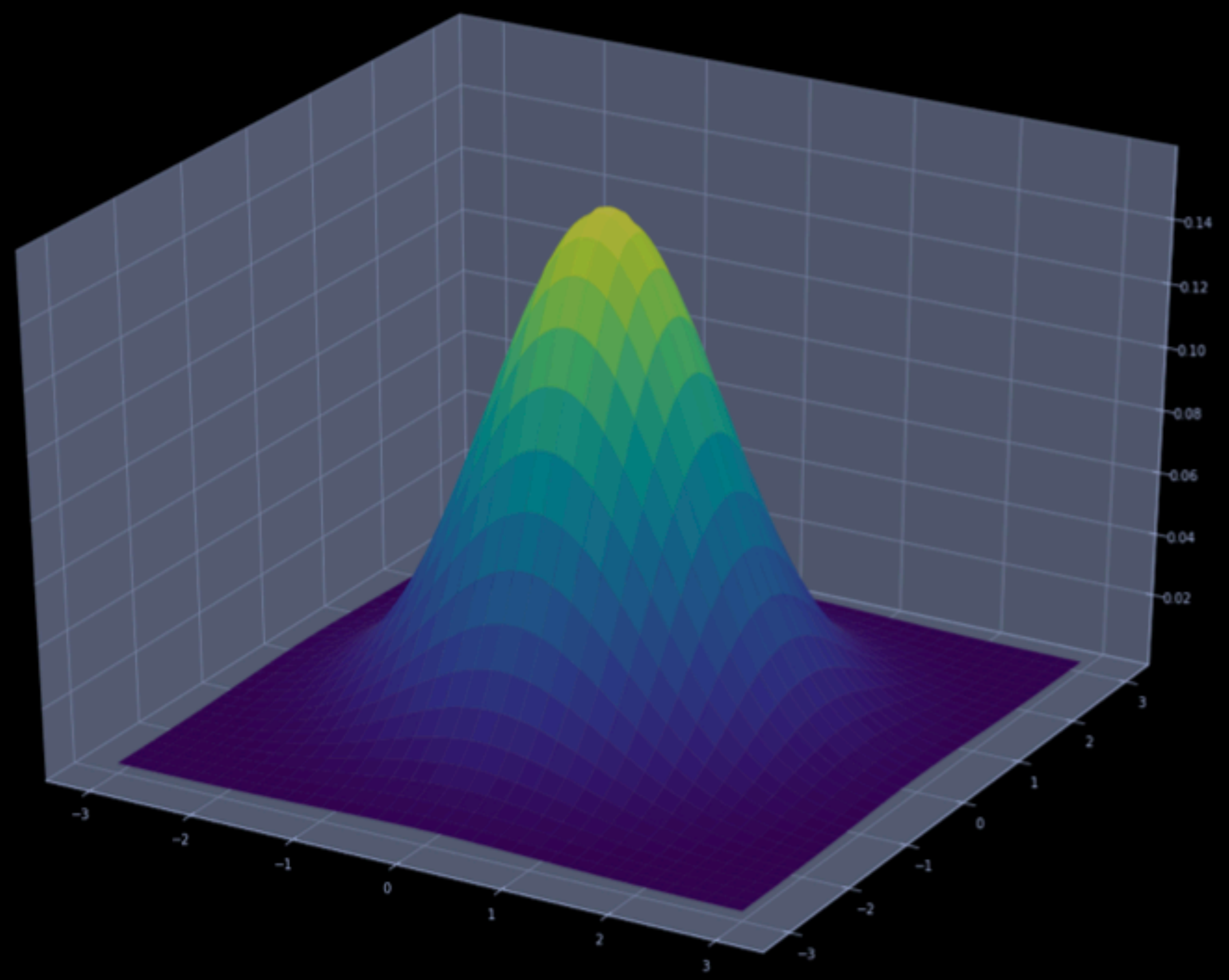
- $X = AS$
  - Let A be an mixing matrix with full rank
- $S \sim N(0, I)$
  - Each source s_i is Gaussian with mean 0
  - The vector S with N dimension is jointly Gaussian and covariance matrix I
- then what will X look like?



- # 1 INTRODUCTION

Case of Gaussian source shall be omitted

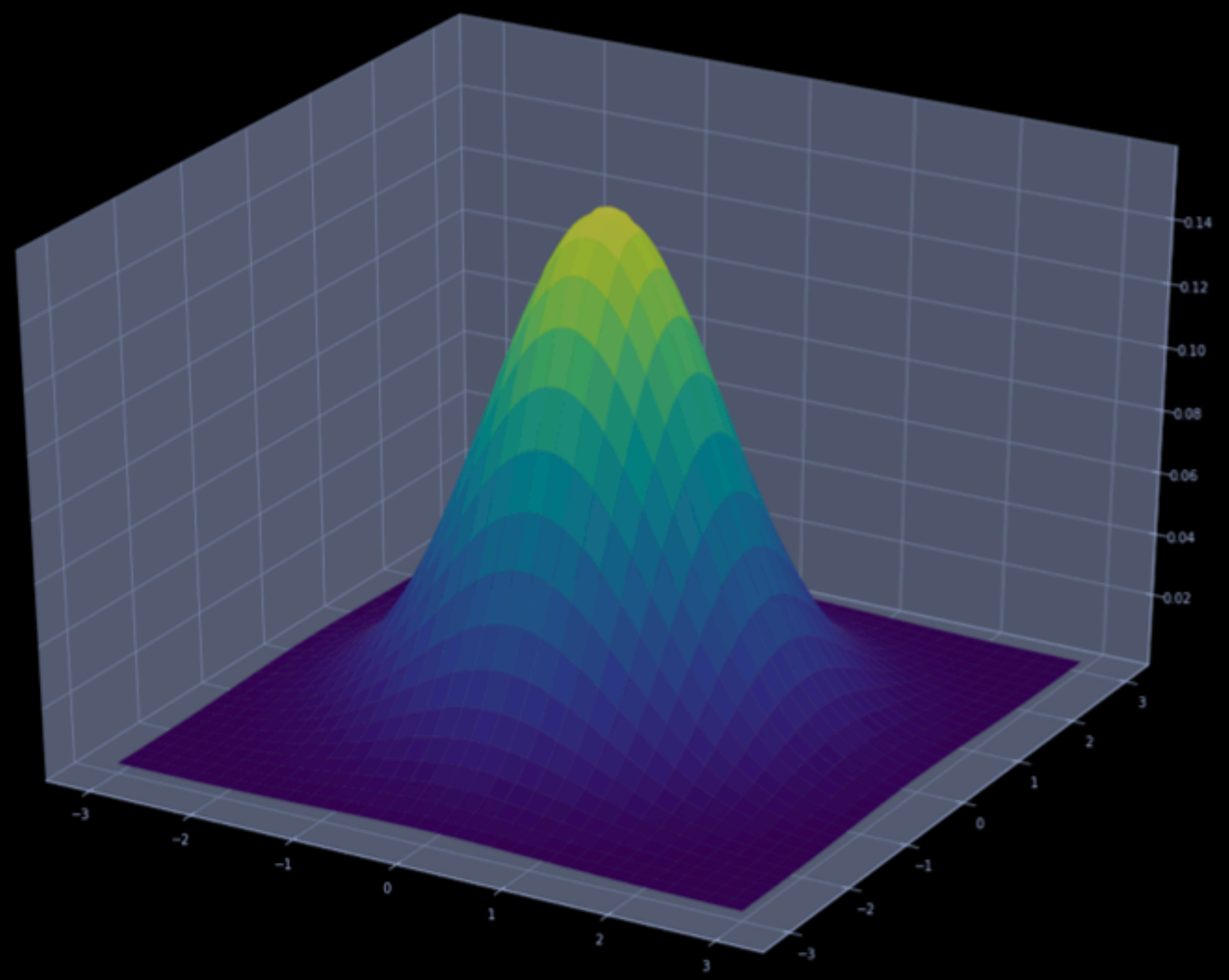
- It's still Gaussian distribution
- What is the two essential components to describe a Gaussian distribution?



• 1 INTRODUCTION

Case of Gaussian source shall be omitted

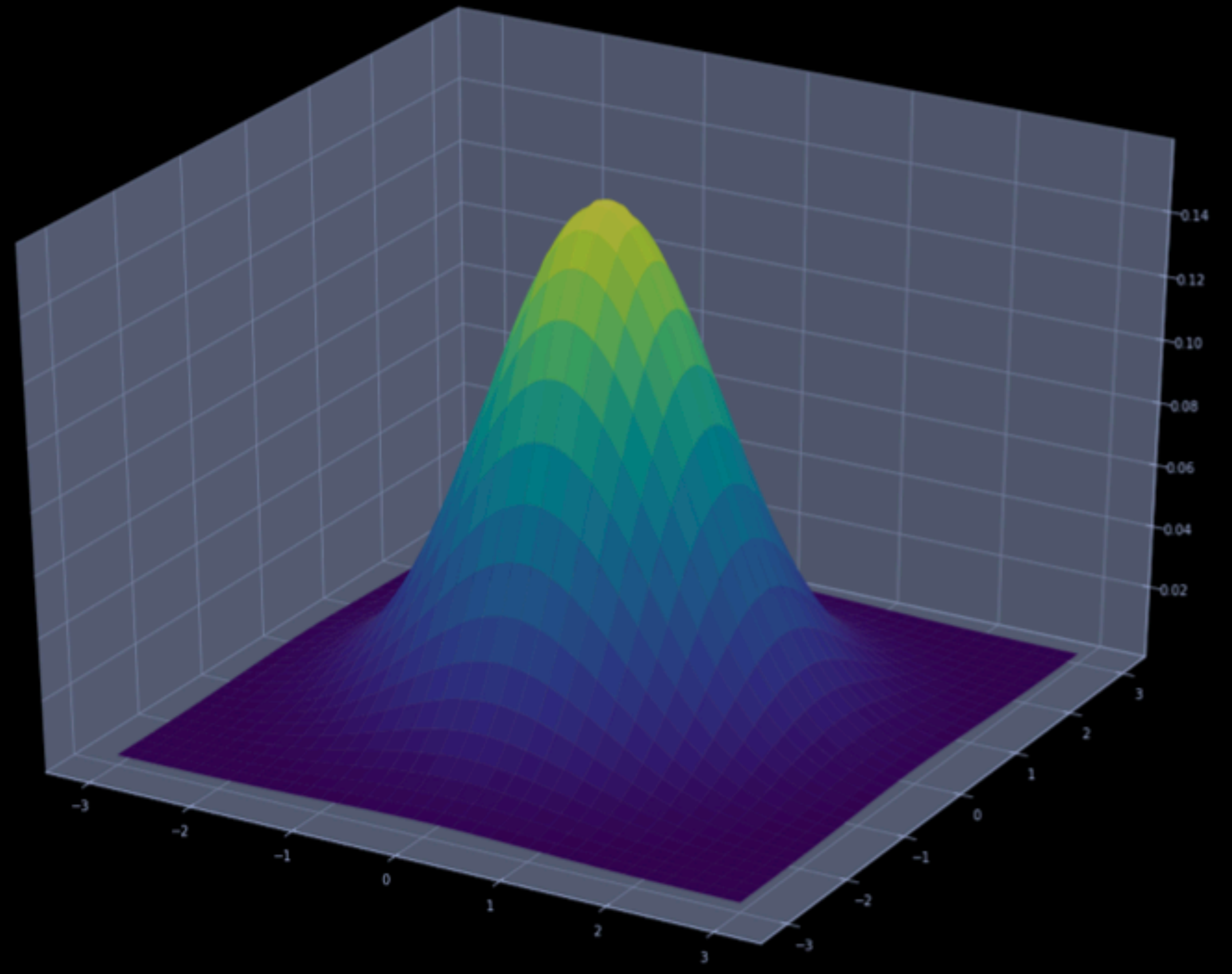
- $X = AS$ is a Gaussian distribution with mean 0 and covariance matrix $E[XX^t] = E[ASS^tA^t] = AA^t$
- Let B be an orthogonal mixing matrix
- $X' = ABS$ is also Gaussian
- X' has mean 0 and covariance matrix $E[X'X'^t] = E[ABSS^tB^tA^t] = AA^t$
- What does that mean?



• 1 INTRODUCTION

Case of Gaussian source shall be omitted

- $X = AS$ is a Gaussian distribution with mean 0 and covariance matrix AA^t
- $X' = A\,(BS)$ is a Gaussian distribution with mean 0 and covariance matrix AA^t
- S and BS are both solution



• 1 INTRODUCTION

Summary: ssumption for (basic) ICA algorithm

- Time delay is not significant in all microphone / observation
- The mixing function is linear
- The sources are not (joint) Gaussian distribution
- The sources EITHER is less Gaussian, OR has properties of higher-order statistics properties on tensorial decompositions

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_n(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \\ \vdots \\ s_n(t) \end{bmatrix},$$

• Poll 1

- Choose all the true statements
    - ICA can handle all kind of independent sources including Gaussian source and square wave source
    - The independence condition is only an assumption and may not be true for some tasks
    - Due to universality, the observation (mixed source) should be more Gaussian for all cases (all the sources and all the mixing procesure)
    - Thanks to universality, the observations are often more Gaussian especially when the number of (independent) source is large
    - Signal processing techniques including beamforming and adaptive filtering is preferred on source separation for the cases time delay is significant

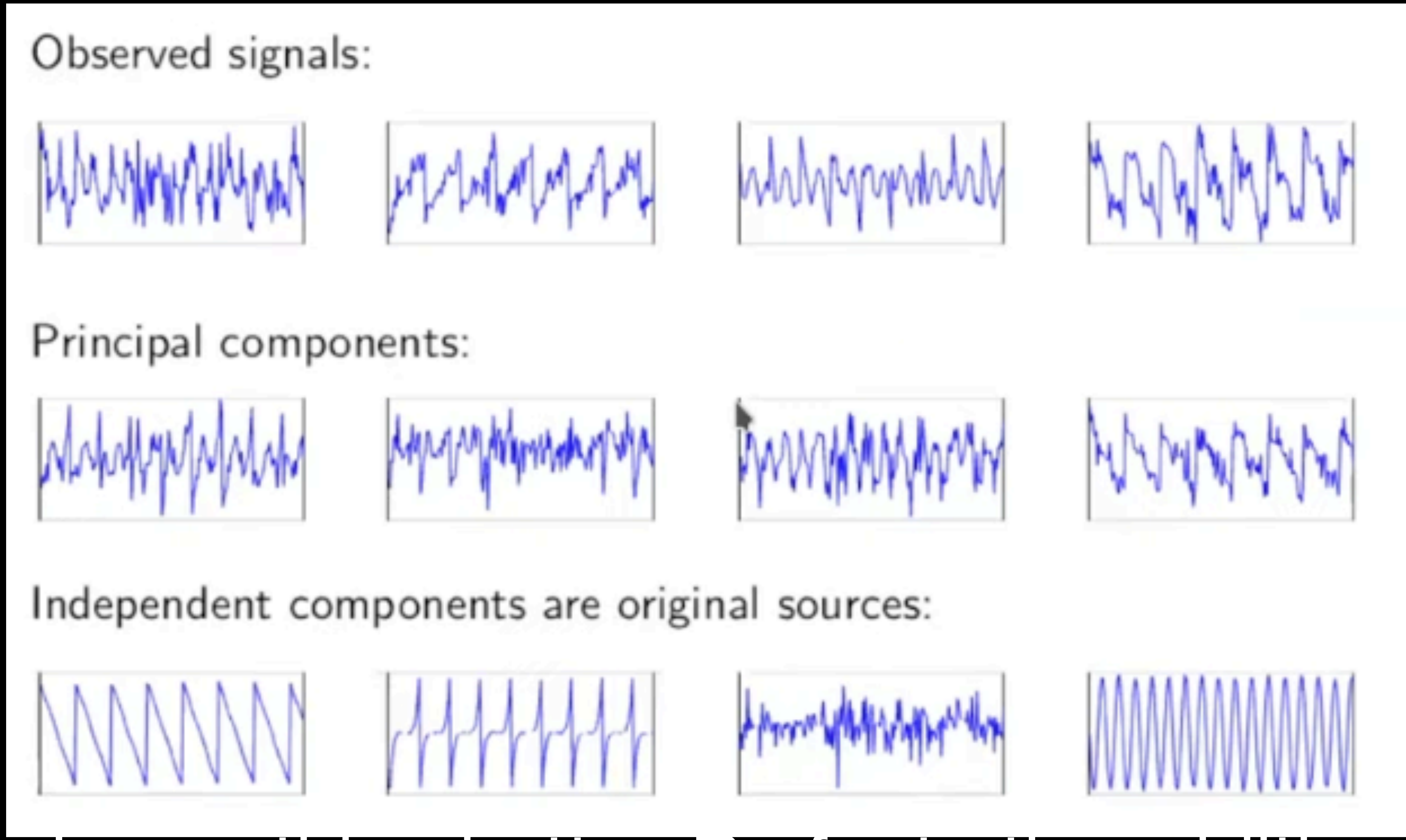
- # 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- First talk about this for homework2
- Main idea: higher order decompositing properties

- **2 FIRST ICA IMPLEMENTATION: FBOI**

PCA: second order moment decomposition, not good enough !!!



- Then, trying to find something better. Rrefer to the additional notes on piazza https://piazza.com/class/ksset4cralds5?cid=150 or what I (will) write on the blackboard

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- PCA: given X, optimize W such that $E\left[SS^t\right] = E\left[WXX^tW^t\right] = I$

- FOBI: given X, optimize W such that
$E\left[S^tSSS^t\right] = E\left[X^tW^tWXWXX^tW^t\right] = I$

- It is recommended to read the additional notes on piazza https://piazza.com/class/ksset4cralds5?cid=150 instead of teh following pages to get to know the intuition of FOBI, although I omit the proof on piazza

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- Main idea: higher order decompositing properties
- FOBI:
  - $E\left[s_1 s_2 s_3 s_4\right] = E\left[s_1\right] E\left[s_2\right] E\left[s_3\right] E\left[s_4\right]$
  - $E\left[s_1^2 s_2 s_3\right] = E\left[s_1^2\right] E\left[s_2\right] E\left[s_3\right]$
  - $E\left[s_1^2 s_2^2\right] = E\left[s_1^2\right] E\left[s_2^2\right]$
  - $E\left[s_1^3 s_2\right] = E\left[s_1^3\right] E\left[s_2\right]$

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- How to evaluate the “independence” with forth order?
  - For any random vector $a = \left(a_1, a_2, \cdots, a_N\right)^T$ with zero mean, defined the fourth order indicator
  - $D_a = E\left[\, \| \, a \, \|^{\,2} aa^t \right]$
- D_a is diagonal if and only if a_i are pairwise independent

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- $D_a = E\left[\| a \|^2 aa^t \right]$
  - D_a is diagonal if and only if a_i are pairwise independent
- For sources S, the indicator matrix D_S should be diagonal

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- $S = WX$, where $X := X - \mu_X$ has zero mean
- $D_S = E\left[S^t SSS^t\right]$
- $D_S = E\left[\left(X^t W^t\right)(WX)(WX)\left(X^t W^t\right)\right]$
- Quite complex
- If only $W^t W = I$ or $XX^t = I$

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- We claim $S = WX$, then $WW^t = I \iff E\left[XX^t\right] = I$
  - Because the covariance matrix of S is identity matrix, so W is a unitary matrix if and only if $E\left[XX^t\right]$ is also identity matrix
- Could we make it identity matrix :-)
- Whiten data !!!

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- Whiten data
  - Orthogonal diagonalization: $E\left[XX^t\right] = P\Lambda P^t$
  - $\hat{X} = \Lambda^{-\frac{1}{2}}P^t \cdot X$
  - Then, $E\left[\hat{X}\hat{X}^t\right] = E\left[\Lambda^{-\frac{1}{2}}P^tXX^tP\Lambda^{-\frac{1}{2}}\right] = I$
- $S = W\hat{X}$ where W is a unitary matrix

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- $W^tW = I$, what will happens?
- $D_S = E\left[S^tSSS^t\right] = E\left[\left(\hat{X}^tW^t\right)\left(W\hat{X}\right)\left(W\hat{X}\right)\left(\hat{X}^tW^t\right)\right]$
- $D_S = E\left[\hat{X}^t\hat{X}W\hat{X}\hat{X}^tW^t\right] = W \cdot E\left[\hat{X}^t\hat{X}\hat{X}\hat{X}^t\right] \cdot W^t = W \cdot D_{\hat{X}} \cdot W^t$
- $W^tD_SW = D_{\hat{X}}$

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- $W^tW = I$, what will happens?
  - $W^tD_SW = D_{\hat{X}}$
  - What's your observation for the equation?

- # 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- What's your observation for the equation?
- Recall that $D_{\hat{X}}$ is symmetric and can be diagnosis with unitary matrix W
- Apply eigen decomposition to $D_{\hat{X}}$

• 2 FIRST ICA IMPLEMENTATION: FBOI

Fourth order blind identification (FOBI)

- Procedure of FOBI
  - (0) let the observation be zero mean $X := X - \mu_X$
  - (1) whiten data $\hat{X} = \Lambda^{-\frac{1}{2}} P^t \cdot X$, where $E\left[XX^t\right] = P\Lambda P^t$
  - (2) Compute weighted fourth order correlation $D_{\hat{X}} = E\left[\hat{X}^t \hat{X} \hat{X} \hat{X}^t\right]$
  - (3) Eigen decomposition: $D_{\hat{X}} = U\Lambda_{\hat{X}} U^t$ and let $W = U^t$
  - (4) Obtain sources: $S = W\hat{X}$

• 2 FIRST ICA IMPLEMENTATION: FBOI

- One last thing for the FOBI Procedure

- What is $E\left[\left(X^t X\right) X X^t\right]$?

- Samples is not random variables !!!

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_n(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \\ \vdots \\ s_n(t) \end{bmatrix},$$

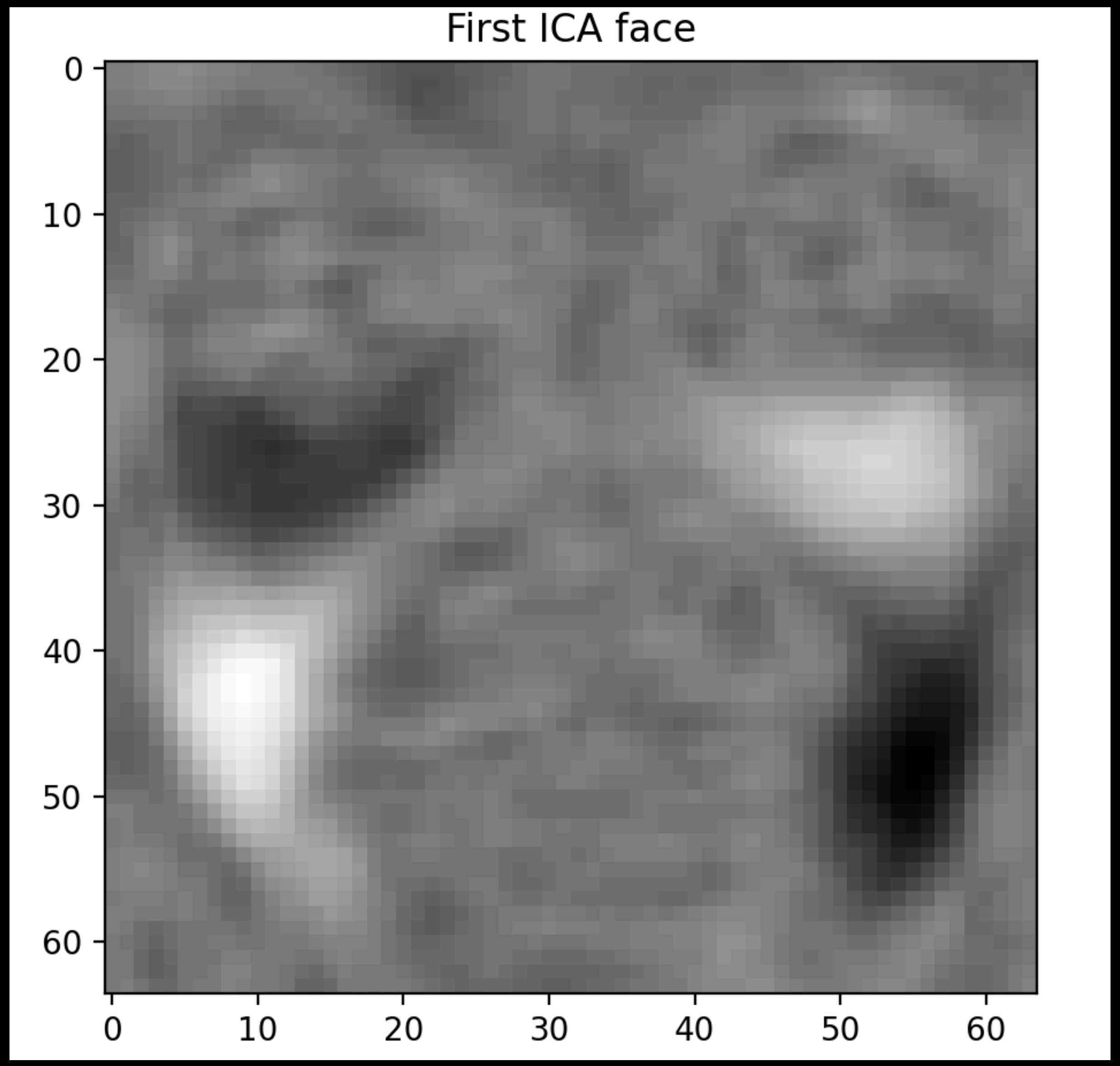
• 2 FIRST ICA IMPLEMENTATION: FBOI

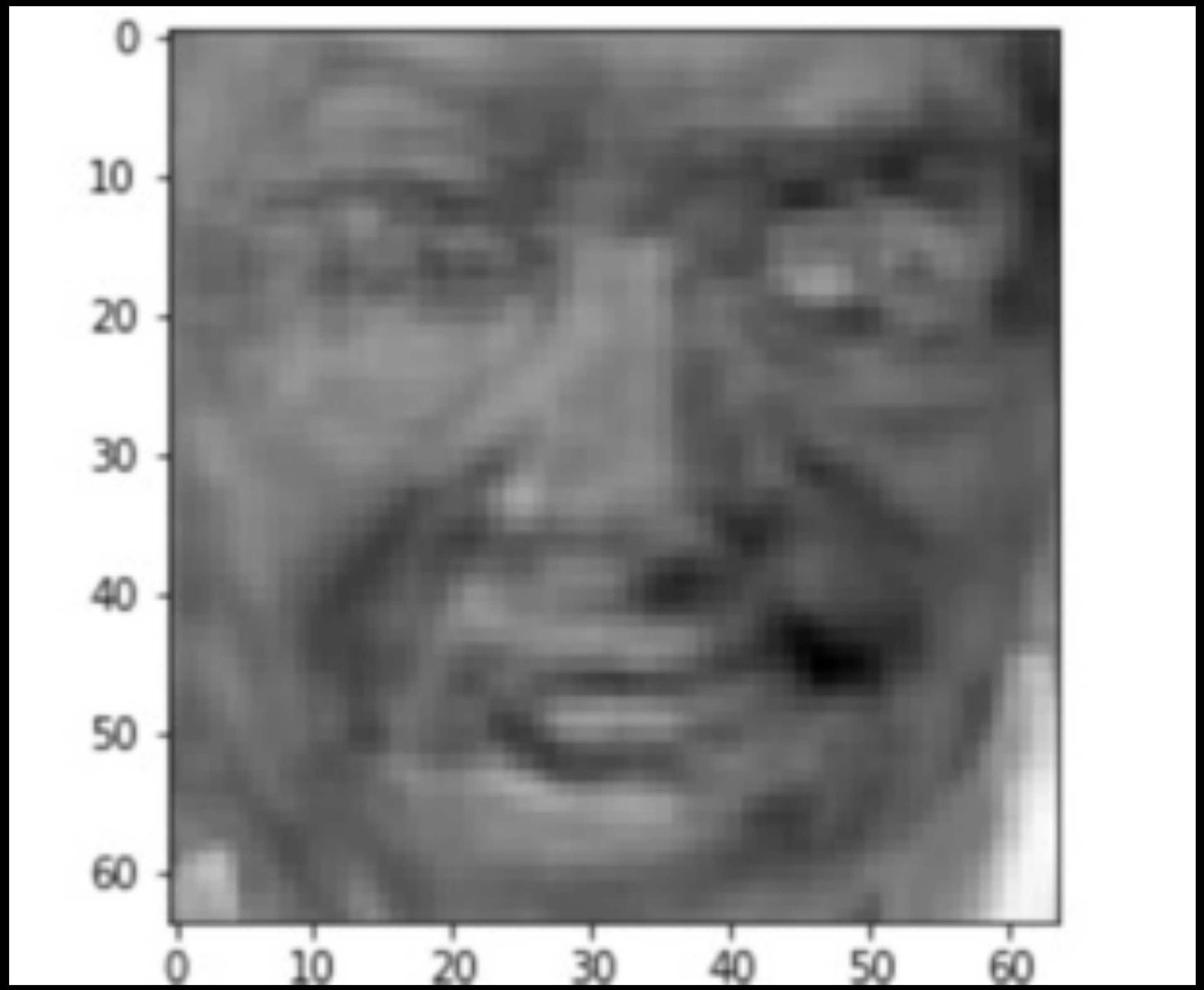
Some Remarks on FOBI

- FOBI is one of the first and most simple ICA methods
- Whiten data can reduce the freedom dimension of W and fasten the convergence
- FastICA based on Gaussian measure generally performances better in case of high-dimensional data
- The most notable drawback of FOBI require all the sources have quite distant in their fourth order moment values, implicating the failure in case of having several mechanisms characterized with the same distribution

• No poll for FOBI

- Best wishes to your homework2 :-)
- Pay attention that you could get different result with FOBI and Fast-ICA (in ski-learn)





• 3 MEASURE OF GAUSSIAN

- Besides, using fourth order moment
- Independent sources have less Gaussian compated to the observation
  - What is “less Gaussian”?

• 3 MEASURE OF GAUSSIAN

divergence = contrast function

- Contrast function, also known as divergence, is a function which establishes the “distance” of one probability distribution to the other on a statistical manifold. ——wikipedia

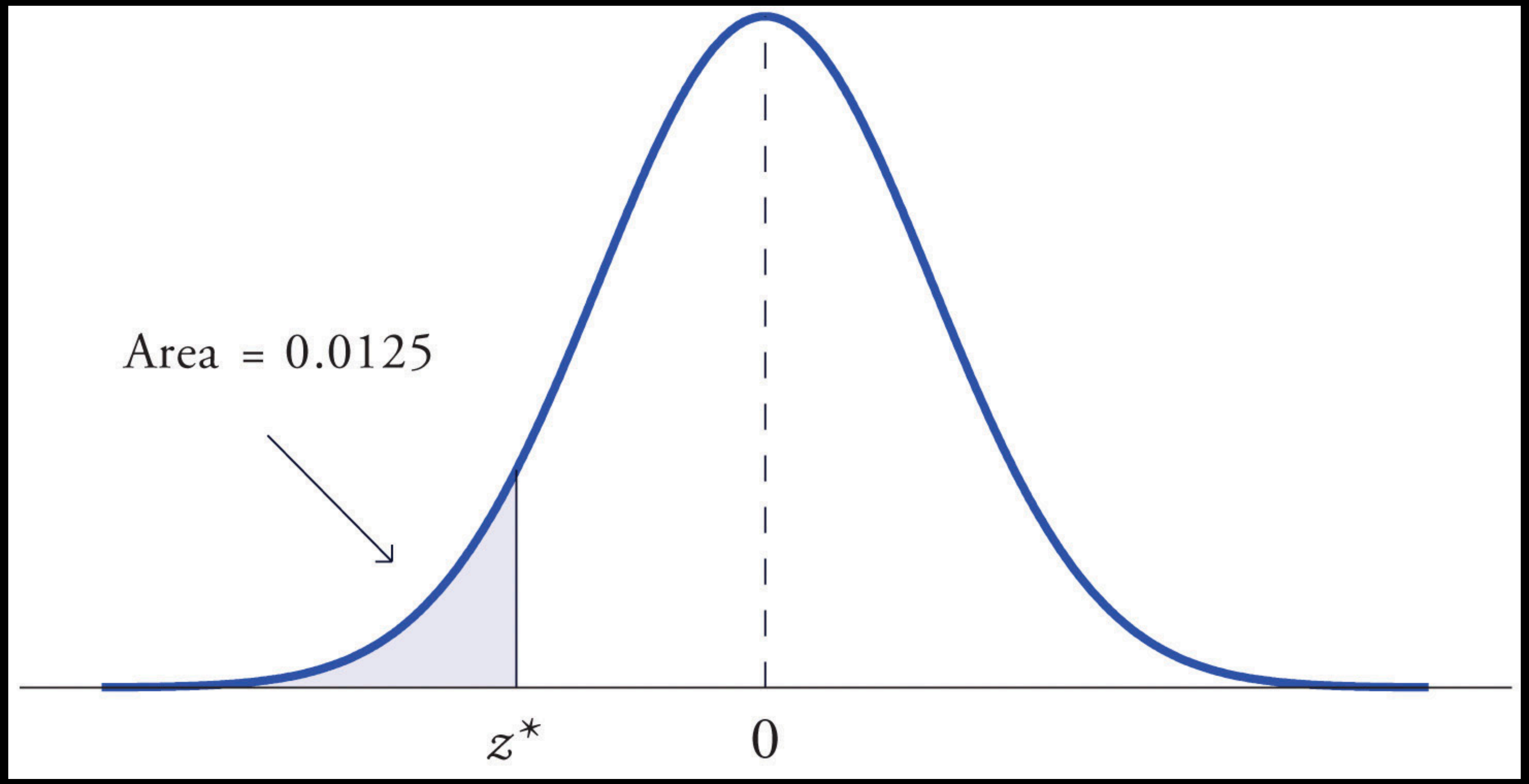
- For exxample: KL-divergence

$$
\begin{aligned}
D_{\mathrm{KL}}(P \parallel Q) &= \int_{x_a}^{x_b} P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx \\
&= \int_{y_a}^{y_b} P(y) \log\left(\frac{P(y)\,\frac{dy}{dx}}{Q(y)\,\frac{dy}{dx}}\right) dy = \int_{y_a}^{y_b} P(y) \log\left(\frac{P(y)}{Q(y)}\right) dy
\end{aligned}
$$

• 3 MEASURE OF GAUSSIAN

3.1 Kurtosis divergence

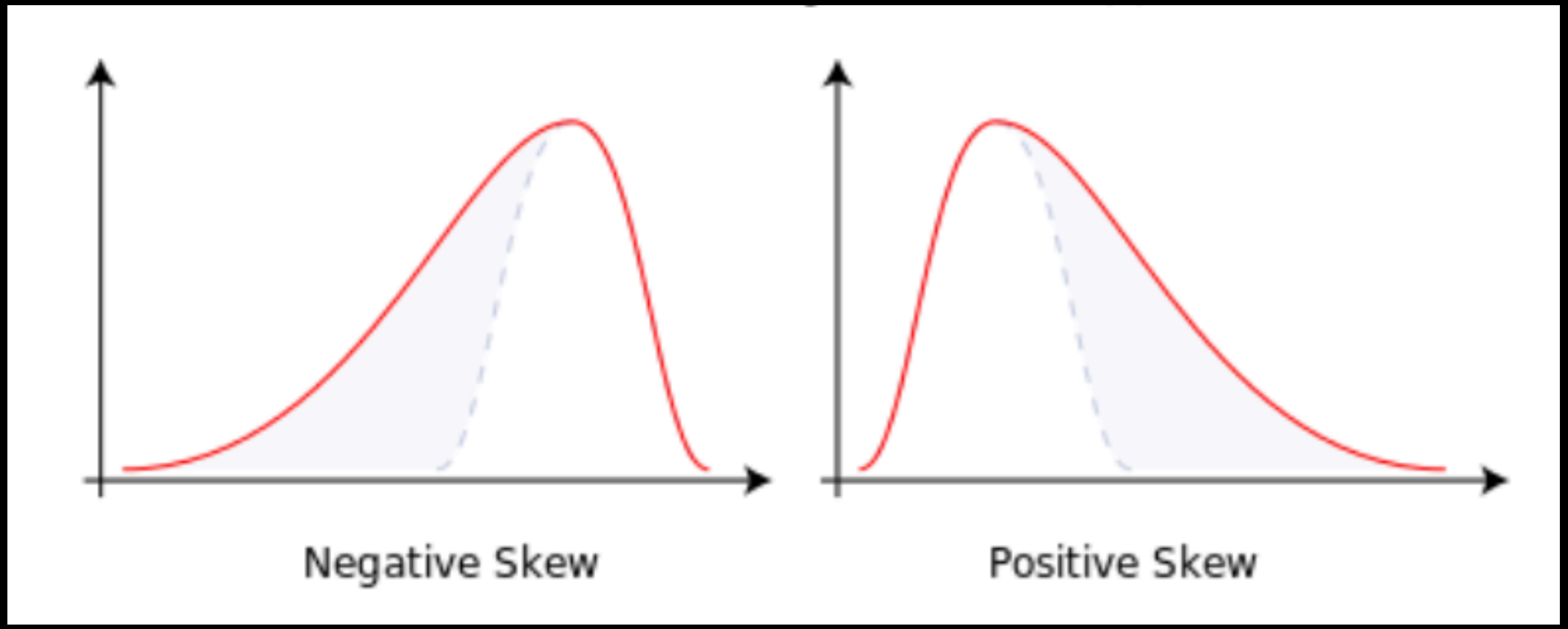
- Gaussian has little tail probability
  - Kurtosis is a scale of forth central moment —— a measure of how heavy the tails of a distribution are



• 3 MEASURE OF GAUSSIAN

3.1 Kurtosis divergence

- Third central moment (skewness) may not be good enough ?
- Third central moment (skewness) may not be good enough ?
- Every symmetric distribution has zero skewness.



- **3 MEASURE OF GAUSSIAN**

3.1 Kurtosis divergence

- Definition: X is a random variable with mean μ and variance σ^2, then

$$Kurt[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{E\left[(x-\mu)^4\right]}{E\left[(x-\mu)^2\right]^2} = \frac{\mu_4}{\sigma^4}$$

- Scale of fourth central moment

- # 3 MEASURE OF GAUSSIAN

3.1 Kurtosis divergence

- For random variable $X \sim N\left(\mu, \sigma^2\right), Kurt[X] = \dfrac{3\sigma^4}{\left(\sigma^2\right)^2} = 3$

- Optimize the Kurtosis to 3 with gradient descent / increase?

• 3 MEASURE OF GAUSSIAN

3.1 Kurtosis divergence

- refined version $Kurt[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3 = \frac{\mu_4}{\sigma^4} - 3$
- Or when X has mean 0 and variance 1,
- $Kurt[X]] = E\left[X^4\right] - 3\left(E\left[X^2\right]\right)^2$

- # 3 MEASURE OF GAUSSIAN

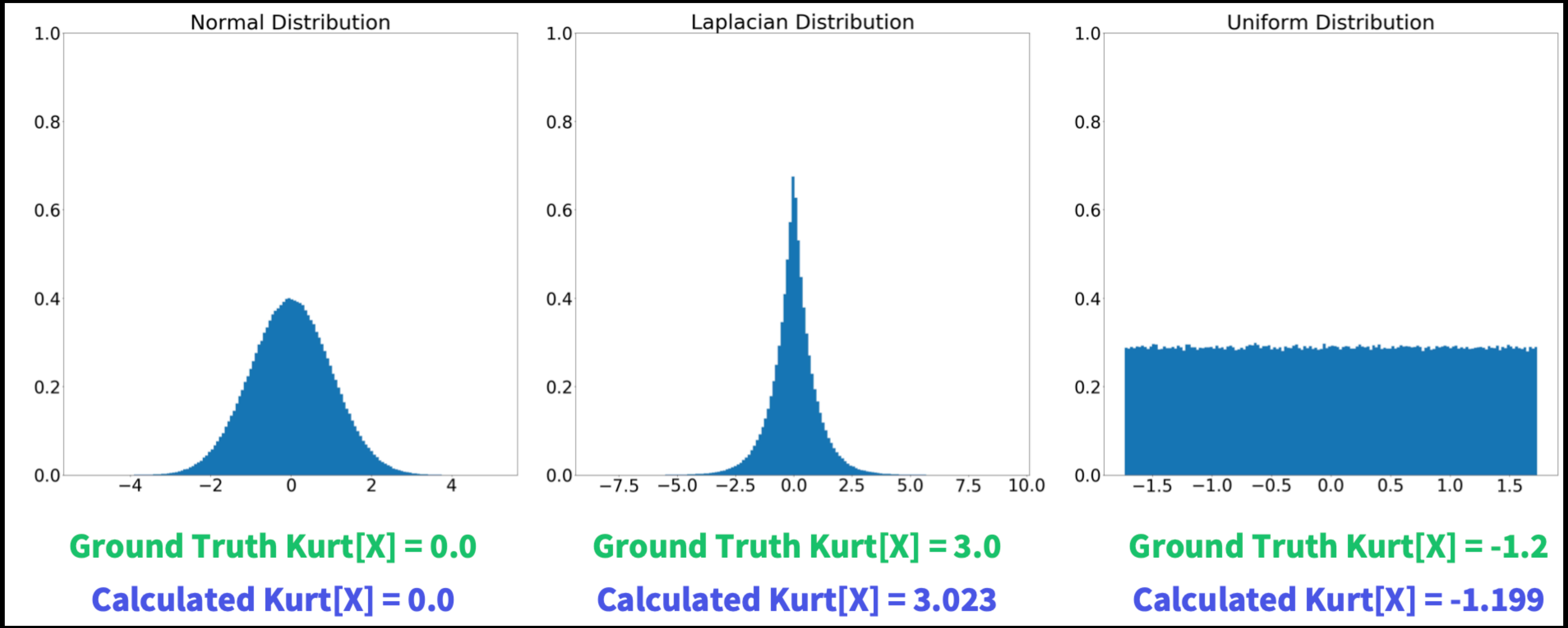
3.1 Kurtosis divergence

- Advantage: easy to compute & optimize
  - For a Gaussian R.V., its (refined) kurtosis is 0
  - Use the absolute value of kurtosis
  - Therefore, we want to maximize the kurtosis of the distribution

• 3 MEASURE OF GAUSSIAN

3.1 Kurtosis divergence

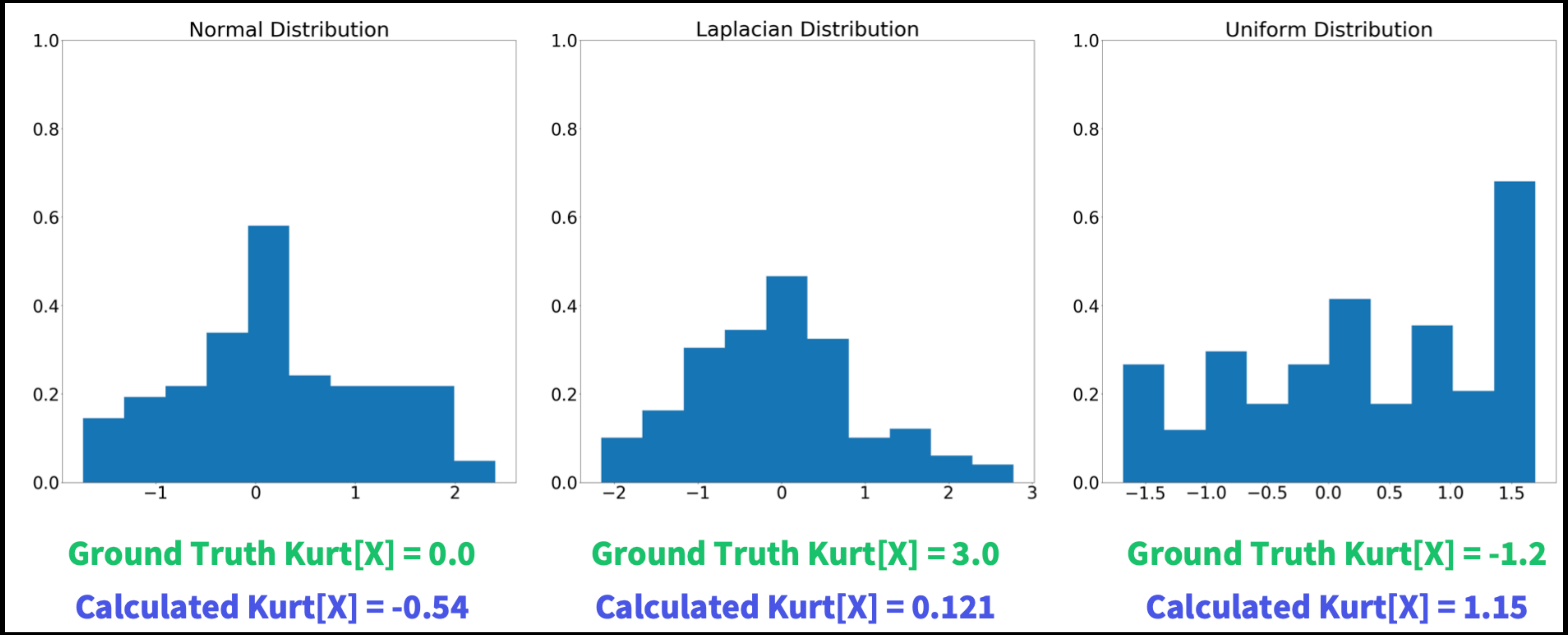
- You can only evaluate with data (sample of R.V.) instead of the R.v. itself
- Generate with 1000000 examples



• 3 MEASURE OF GAUSSIAN

3.1 Kurtosis divergence

- What if less samples?
- Generate with 100 examples



• 3 MEASURE OF GAUSSIAN
3.1 Kurtosis divergence

- Benefits
  - computationally easy
  - widely used!
- Disadvantages
  - Susceptible to outliers
  - Few data points leads to bad estimate
  - Not a robust measure of Gaussianity!

• 3 MEASURE OF GAUSSIAN

3.2 neg-entropy

- Entropy: $H(X) = -\Sigma_i p_i \log\left(p_i\right)$
  - Entropy is a measure of surprise
  - R.V. that is “more random” will have a larger entropy as more bits needed to send and vice versa
- What is the entropy of a Gaussian random variable?

- # 3 MEASURE OF GAUSSIAN

3.2 neg-entropy

- Entropy of a Gaussian: depends, but it's the largest possible value of any distribution with equal variance
- Given R.V. X which has variance σ^2, let $X_{Gauss} \sim N\left(0,\sigma^2\right)$ be a Gaussian with the same covariance matrix as X
- Denote $J(X) := H\left(X_{Gauss}\right) - H(X)$ as the negentropy of X

• 3 MEASURE OF GAUSSIAN

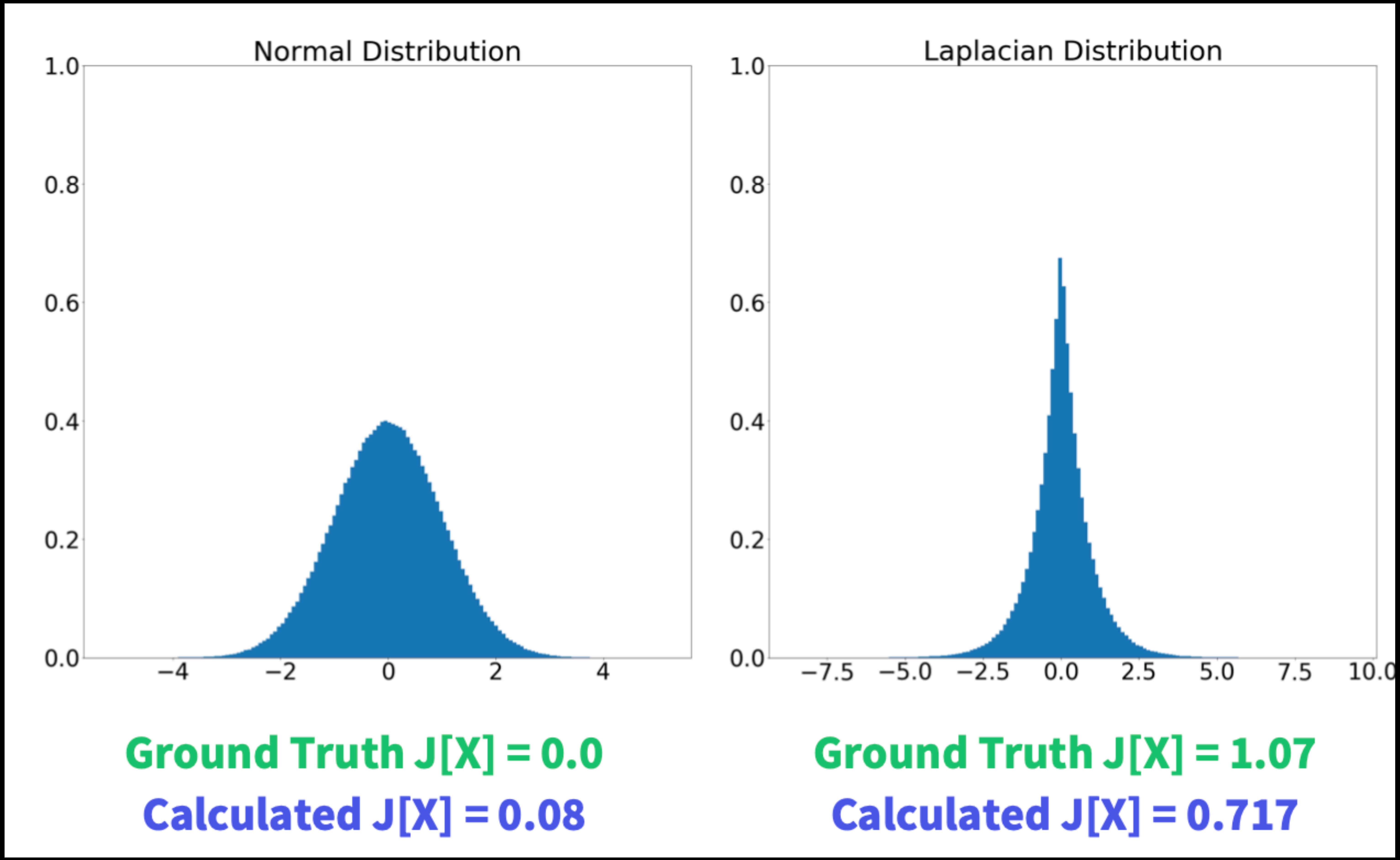
3.2 neg-entropy

- $J(X) := H\left(X_{Gauss}\right) - H(X)$
    - $J(X) \geq 0$ and the equation holds iff X is Gaussian
    - Maximize negentropy to get source
- Sounds good …

• 3 MEASURE OF GAUSSIAN

3.2 neg-entropy

- Generated with 1,000,000 examples
- Use GOOD approximation of $J(X) := H\left(X_{Gauss}\right) - H(X)$ instead of itself !!!



- # 3 MEASURE OF GAUSSIAN

3,2 neg-entropy

- Approximation of negentropy

- $$J(X) \propto \left[E\left[G(X)\right] - E\left[G(v)\right]\right]^2,$$

- where $v \sim N(0,I)$, G is a non-linear and non-quadratic functions

- Some commonly used G

pow3: $G_1(y) = \frac{y^4}{4}$

tanh: $G_2(y) = \frac{1}{a}\log\left(\cosh(ay)\right).$

skew: $G_3(y) = \frac{y^3}{3}$

- # 3 MEASURE OF GAUSSIAN

3.2 neg-entropy

- Advantages:
  - Very well justified measure of Gaussianity
- Disadvantages
  - Computationally hard
  - Must estimate the PDF of a R.V. for accuracy results but we will usually approximate negentropy and maximize over that

• Poll 2

- Which divergence below is easier to implement
  - Kurtosis divergence
  - Neg-entropy
- Which divergence below is more accurate with less deviation
  - Kurtosis divergence
  - Neg-entropy

• 4 SECOND ICA IMPLEMENTATION

4,1 Fast ICA

- Given observation X, optimize W ($S = WX$) such that it maximize $j(S)$
- $J(S) = J(WX) \propto \left[E\left[G(WX)\right] - E\left[G(v)\right]\right]^2$
- $W = argmax \left\{ E\left[G\left(W^t X\right)\right] - E\left[G(v)\right] \right\}$, condition on $\| W \|^2 = 1$
- How to solve?
- Leave the solution slides after class

- # 4 SECOND ICA IMPLEMENTATION

4,1 Fast ICA

- $$F(W) := \left\{ E\left[G\left(W^t X\right)\right] - E\left[G(v)\right] \right\}$$
- After applying the Lagrange multiplier, $F(W)$ can be rewritten in terms of the first derivatives of G and the optimal value of W, that is, G' and W_0
- $$F^*(W) = E\left[XG'\left(W^t X\right)\right] - E\left[W_0^t X G'\left(W_0^t X\right)\right] W$$
- The iteration can be reduced to the Newton method used in order to find a vector W leading to the maximal negentropy.

• 4 SECOND ICA IMPLEMENTATION

4,1 Fast ICA

- Given that we are actually dealing with the nonlinear system of equation, this has to be done using Jacobian matrix
- $Jaco\,(W) = E\left[G''\left(W^tX\right)\right]I - E\left[W_0^tXG'\left(W_0^tX\right)\right]I$
- Then, the iteration step of Fast-ICA is the following:
- $W_{n+1} = W_n - Jaco^{-1}\left(W_n\right)F*\left(W_n\right)$
- normalize W_{n+1} befrore next iteration.
- The convergence of the algorithm is verified by calculating a dot product of W_{n+1} and W_n, which ought to be zero

• 4 SECOND ICA IMPLEMENTATION

4,1 Fast ICA

- A useful toolkit: ski-learn

```
from scipy.fft import dct
from sklearn.datasets import load_digits
from sklearn.decomposition import FastICA

ica = FastICA(n_components=100)   # X.T samp
X_transformed = ica.fit_transform(X.T)
# print(X_transformed.shape) # sample 1071
w = ica.components_ #(100, 4096)
```

-

- # 4 SECOND ICA IMPLEMENTATION

4.2 Other Methods

- Joint Approximation Diagonalization of Eigen- matrices (JADE), you can find a short introduction in hidden slide
- Robust FOBI, by Cardoso. Free up third moment
- fastICA that free upon some function
- In 1995, Tony Bell and Terry Sejnowski proposed a simple infomax neural network algorithm for independent component analysis (ICA)

Another typical ICA approach in TENSORIAL DECOMPOSITIONS

JADE is a generalization of FOBI [4]. By considering covariance matrix to be a second order cumulant tensor, the kurtosis matrix (9) can be considered as a fourth order cumulant tensor of the identity matrix ($\mathbf{K}_I = \mathbf{F}(I)$). Replacing the identity matrix with a set of tuning matrices (eigenmatrices of the cumulant tensor: $\{M_1, \ldots M_p,\}$) results in a set of cumulants $\{K_{M1}, \ldots K_{Mp}\}$. The whitened de-mixing matrix **D** is estimated by jointly diagonalizing these matrices, which reduces to the maximization problem:

$$max\, J(\mathbf{D}) = max \sum \left(\left(\operatorname{diag}\left(\mathbf{D}\mathbf{K}_{M_p}\mathbf{D}\right)\right)\right)^2 , \tag{10}$$

where $\left\|diag(.)\right\|^2$ is the squared l_2 norm of the diagonal. Given that the maximization of the diagonal elements is equivalent to the minimization of the off-diagonal ones, the resulting de-mixing matrix **D** jointly diagonalize the set of cumulants. This algorithm overcomes the mentioned drawback of FOBI, but stays limited to low-dimensional problems.

• 4 SECOND ICA IMPLEMENTATION

4.2 Other Methods

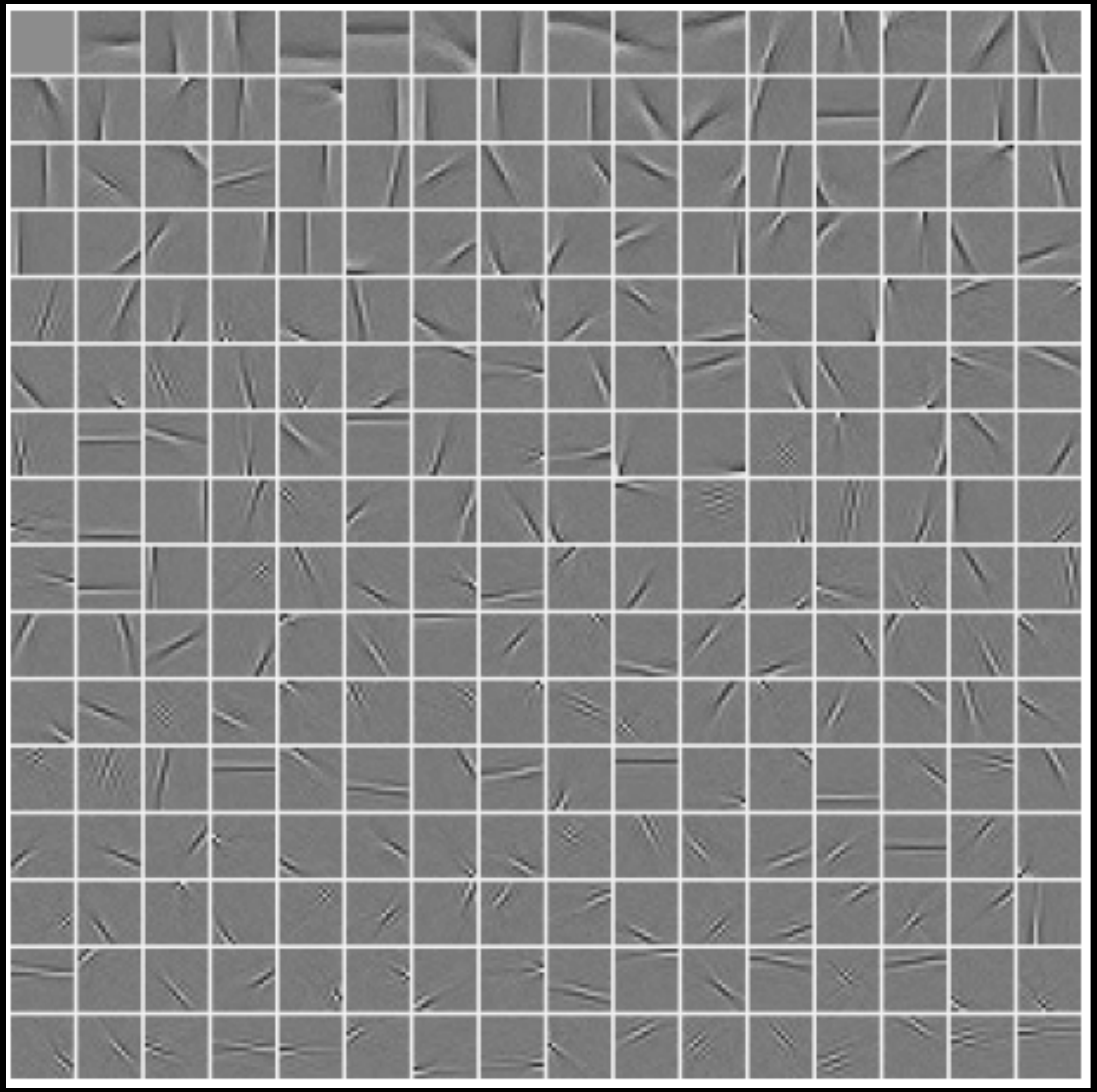
- None of them really gurrantee to give you independence.
- You can come out some other functions and put up a new method on your own, such as Try to free upon other moment beyond second moment :-)

Poll 3

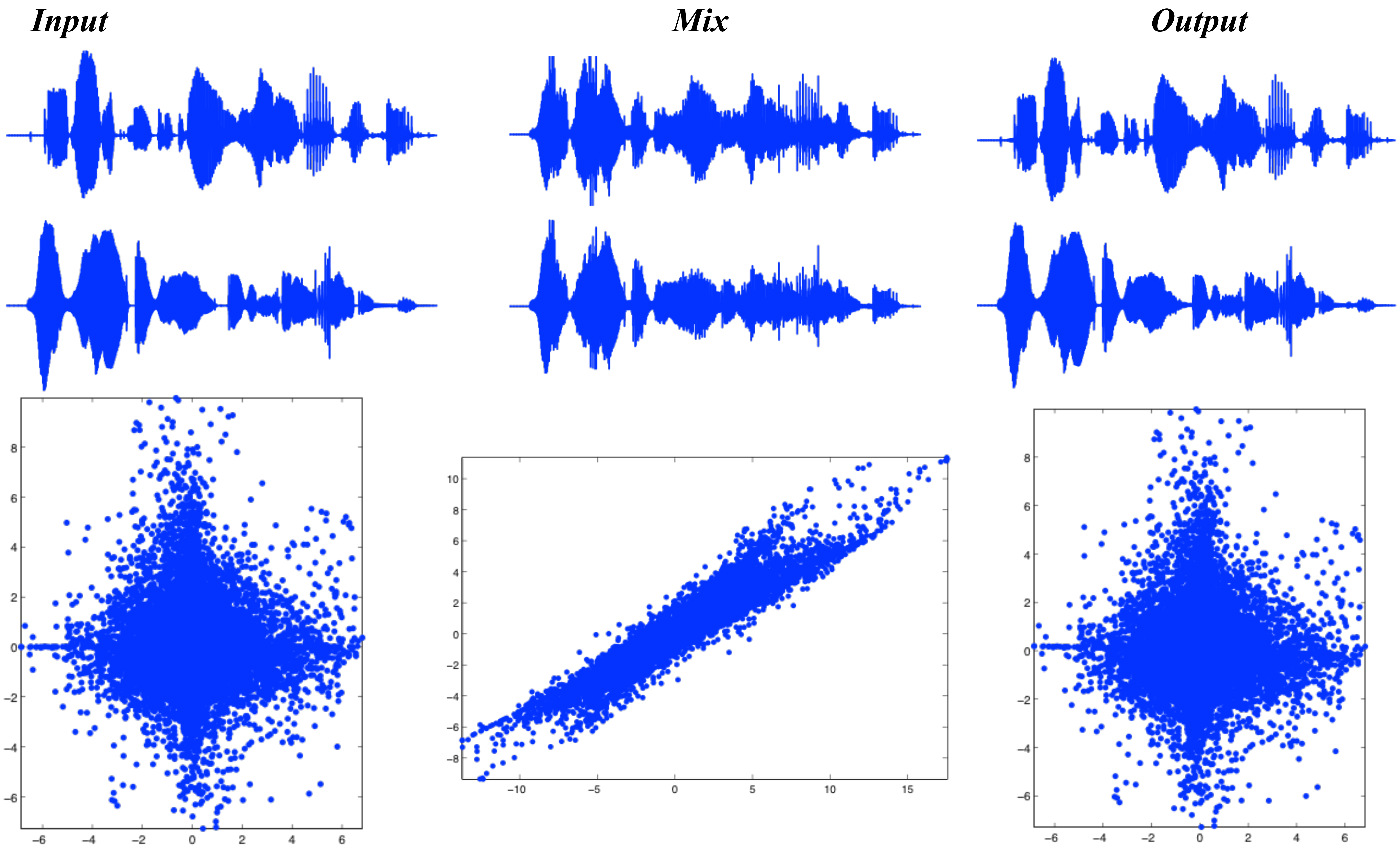
- Choose all the true statement as follows
  - FOBI focus on the uncorrelation in third moment of random variables (independent sources) to evaluate the independent component in signal
  - You can put forward your own methods to solve ICA by using another order of moment people never used before
  - Fast ICA use second orders moment to evaluate the independent component behind a signal
  - Fast ICA use a specific function instead of any orders moment to evaluate the independent component behind a signal

5 APPLICATION

- image recognition (see more details in homework3)
- example: ICA bases of a set of 16×16 pixels natural images (not only faces).

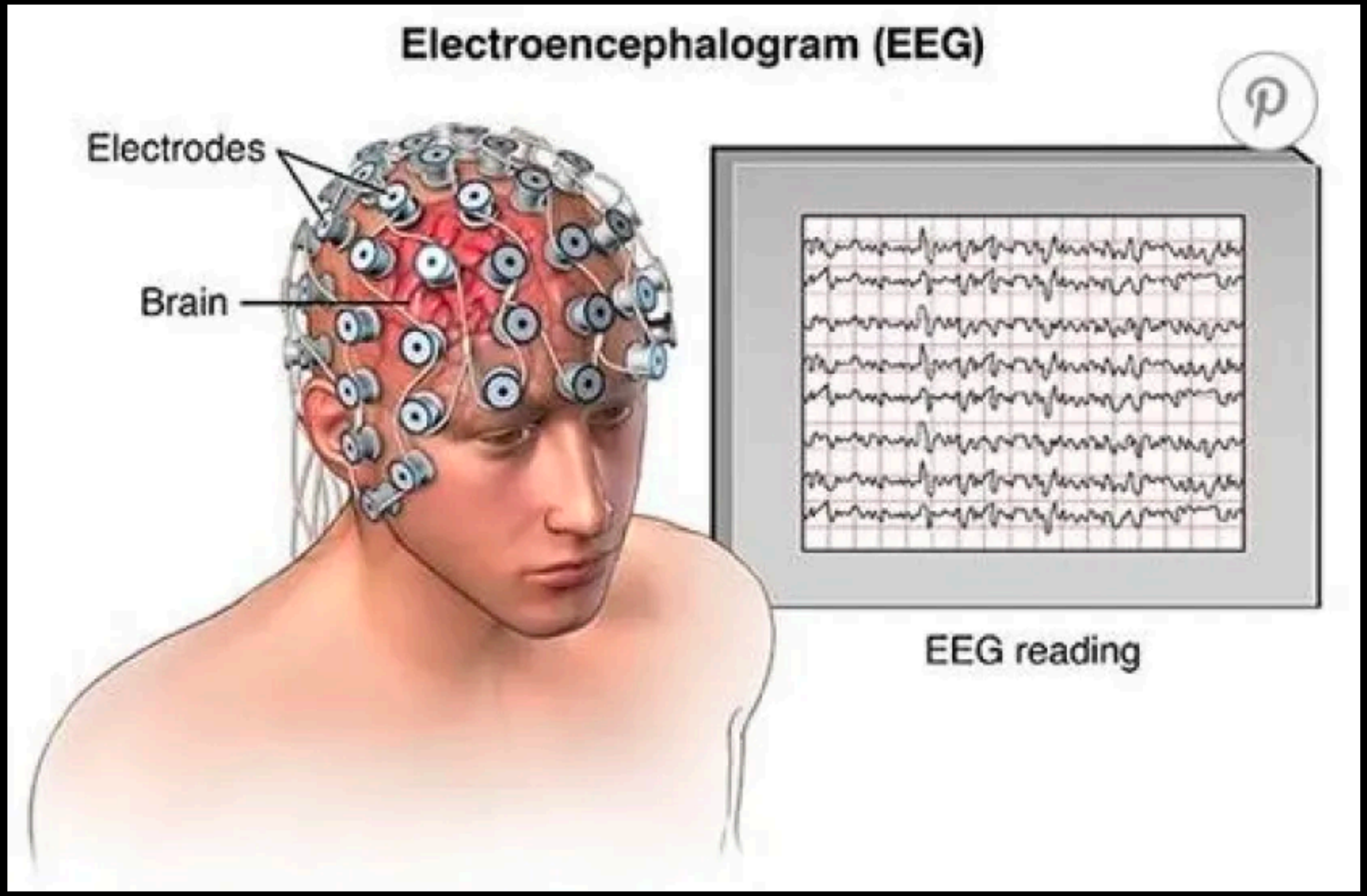


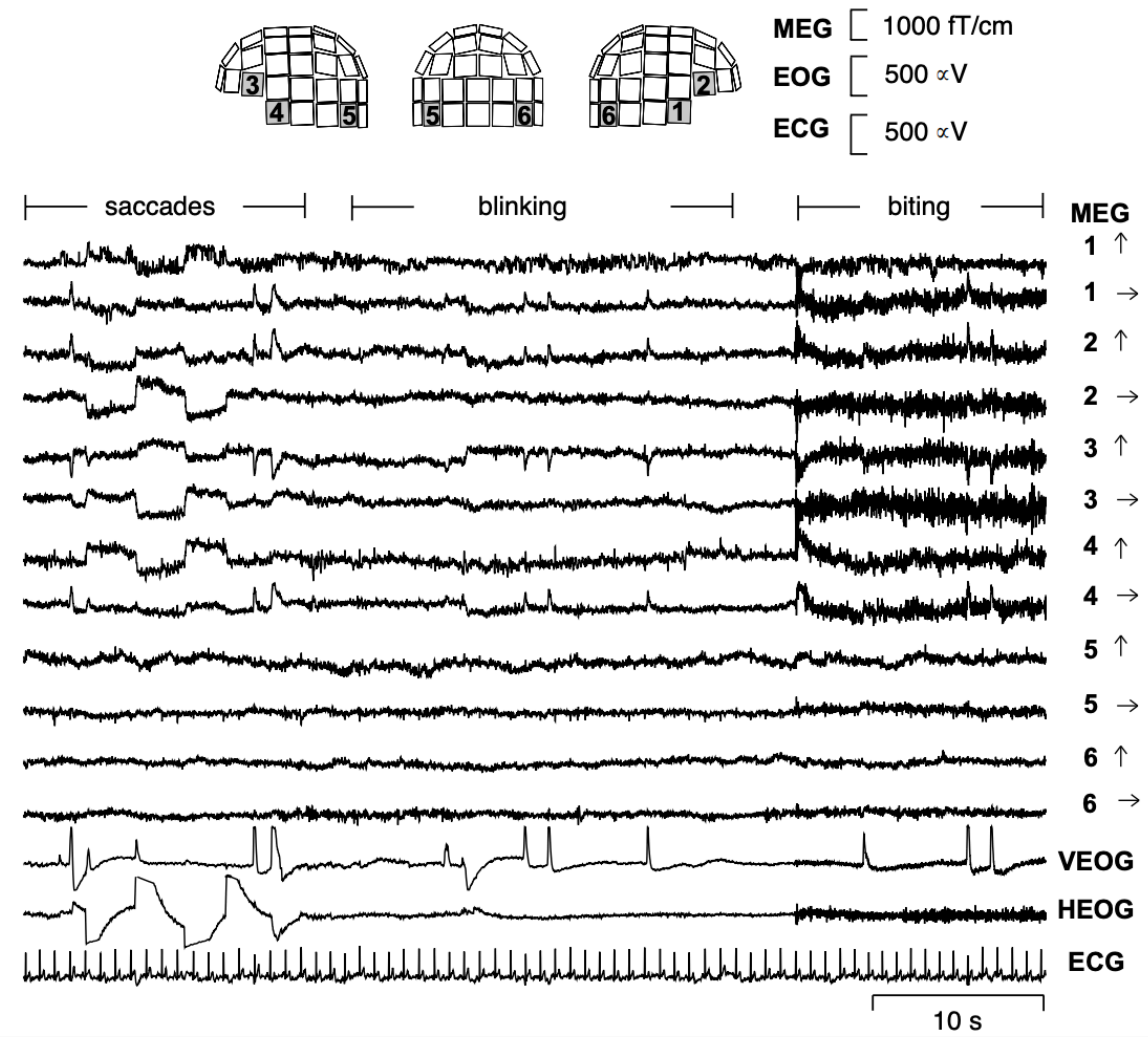
Another example!



5 APPLICATION

- Very commonly used to enhance EEG signals
- EEG signals are frequently corrupted by heartbeats and biorhythm signals and ICA can be used to separate them out




MEG 1000 fT/cm
EOG 500 αV
ECG 500 αV
saccades
blinking
biting
MEG
1 ↑
1 →
2 ↑
2 →
3 ↑
3 →
4 ↑
4 →
5 ↑
5 →
6 ↑
6 →
VEOG
HEOG
ECG
10 s

5 APPLICATION

- extracting structure from stock returns and predicting stock market prices
- Finding hidden factors in financial datas

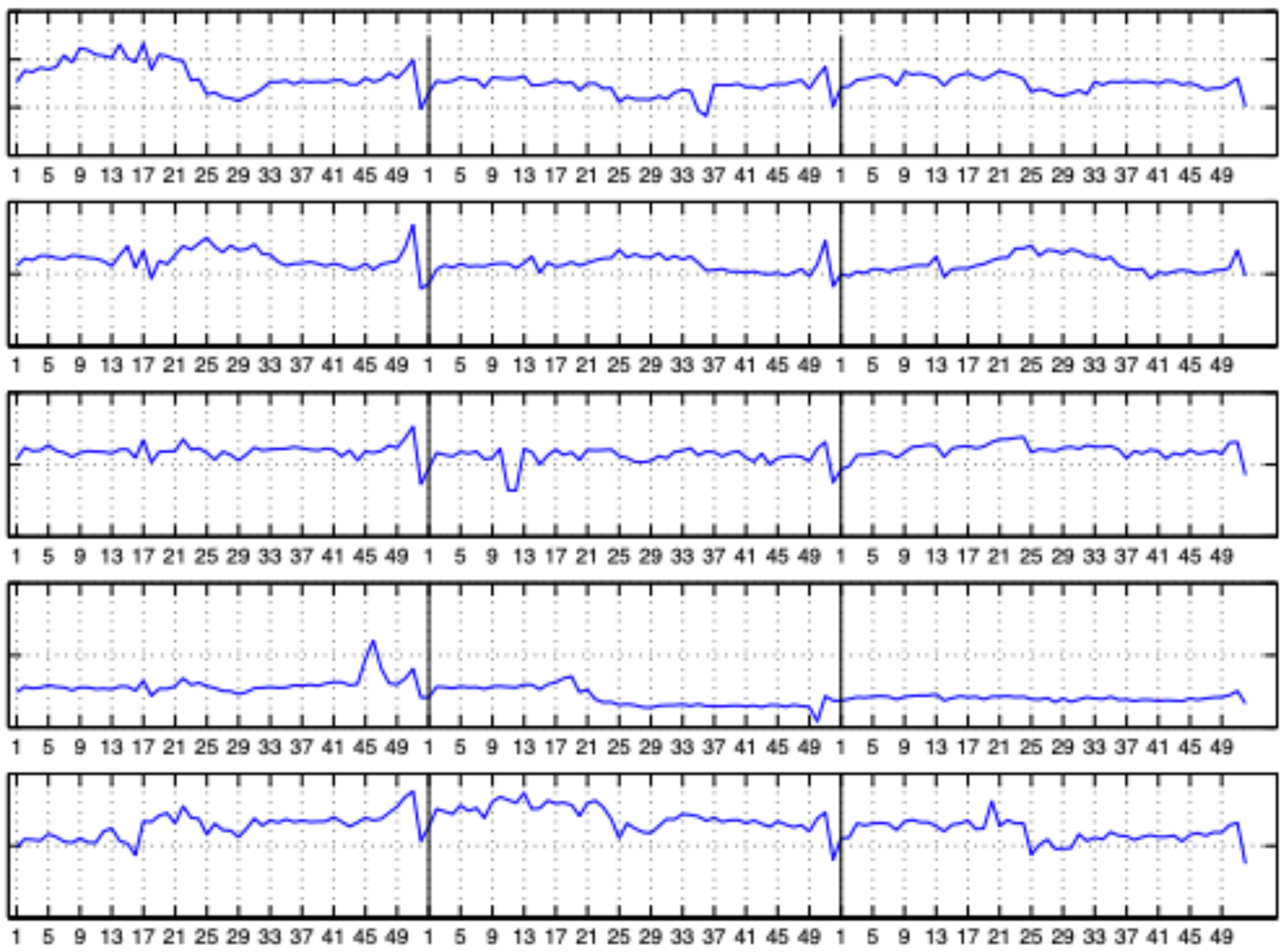


Figure 13: (from [30]). *Five samples of the original cashflow time series (mean removed, normalized to unit standard deviation). Horizontal axis: time in weeks.*

- we applied ICA on a different problem the cashflow of several stores belonging to the same retail chain trying to nd the fundamental factors common to all stores that affect the cashflow data.
- The assumption of having some underlying independent components in this specic application may not be unrealistic. For example factors like seasonal variations due to holidays and annual variations and factors having a sudden effect on the purchasing power of the customers like prize changes of various commodities can be expected to have an eect on all the retail stores and such factors can be assumed to be roughly independent of each other. Yet depending on the policy and skills of the individual manager like eg advertising efforts the effect of the factors on the cash ow of specic retail outlets are slightly dffierent. By ICA it is possible to isolate both the underlying factors and the eect weights thus also making it possible to group the stores on the basis of their managerial policies using only the cash ow time series data.

5 APPLICATION

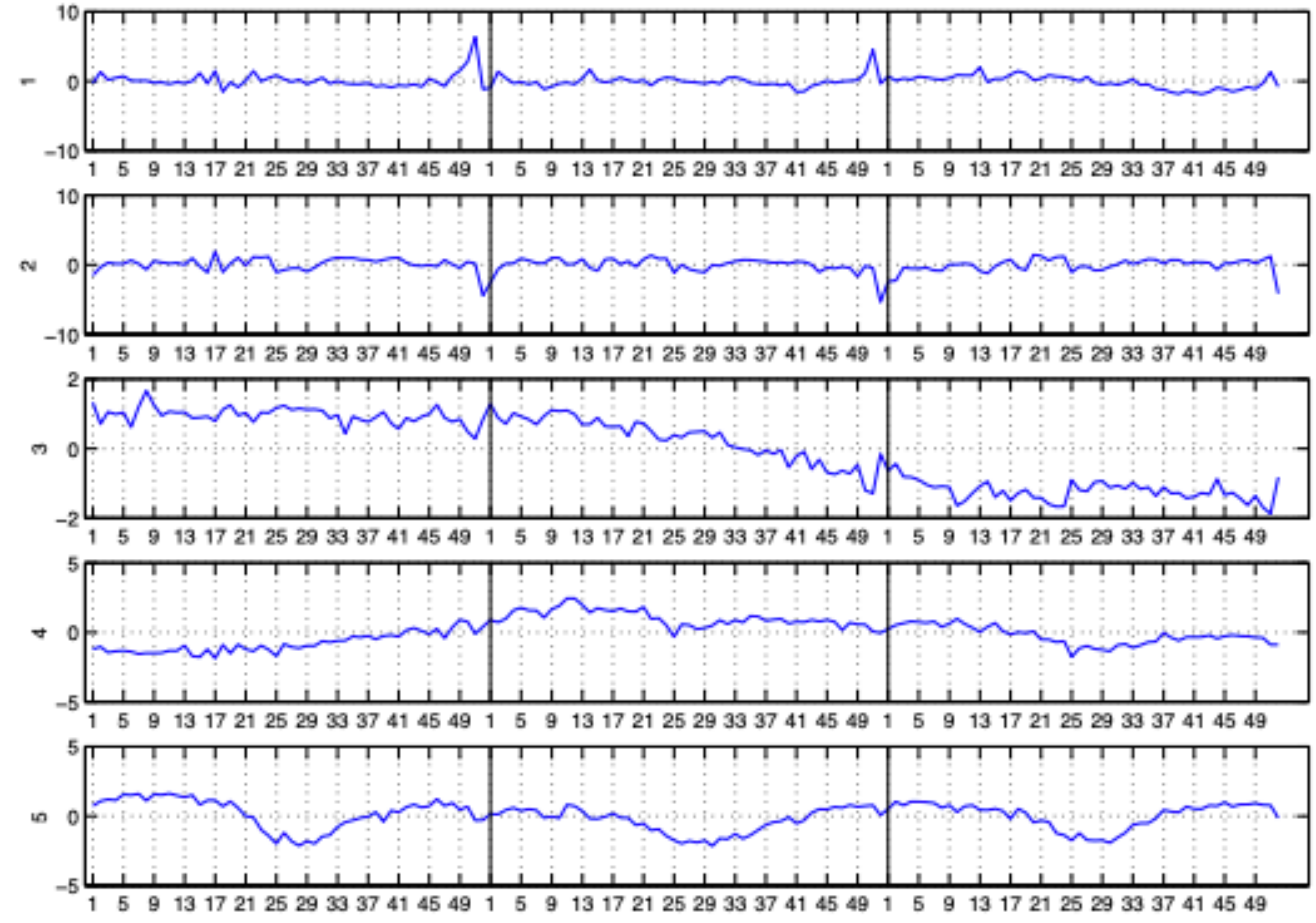
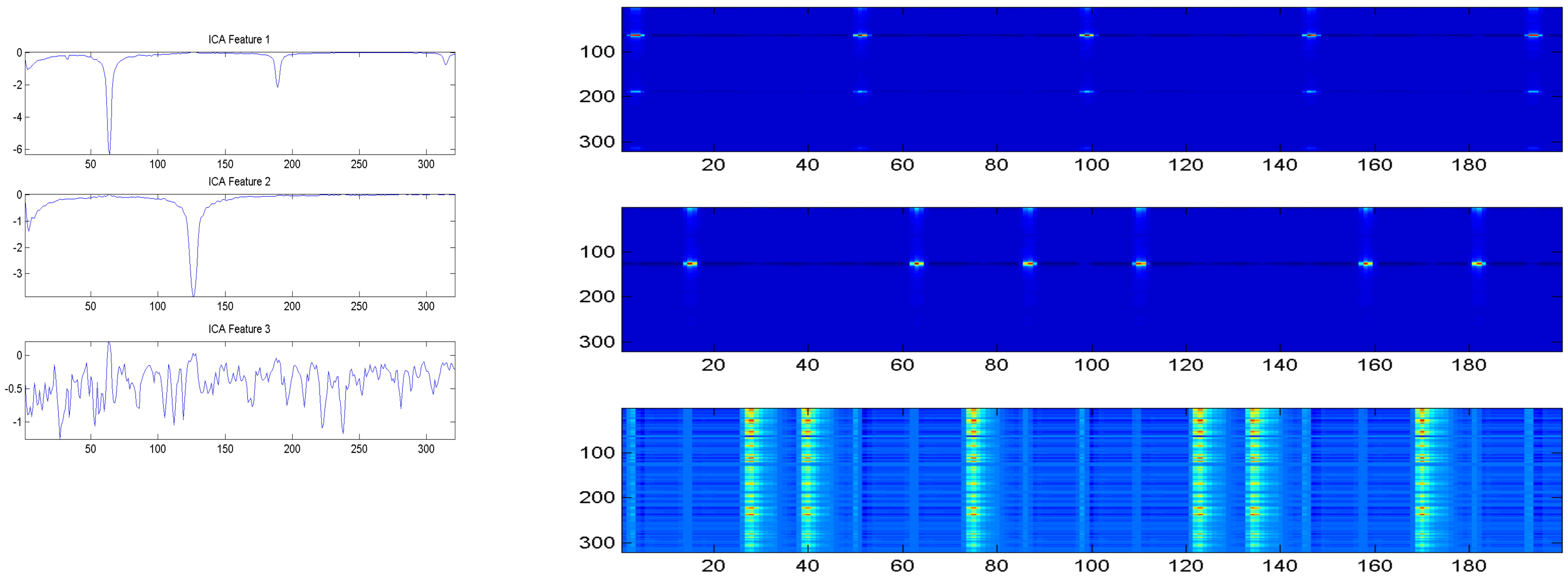


Figure 14: (from [30]). *Four independent components or fundamental factors found from the cashflow data.*

The Notes



- Three instruments..

5 APPLICATION

- analysis of changes in gene expression over time in single cell RNA-sequencing experiments
- Identify and Separate Bright Galaxy Clusters from the Low-frequency Radio Sky
- ……

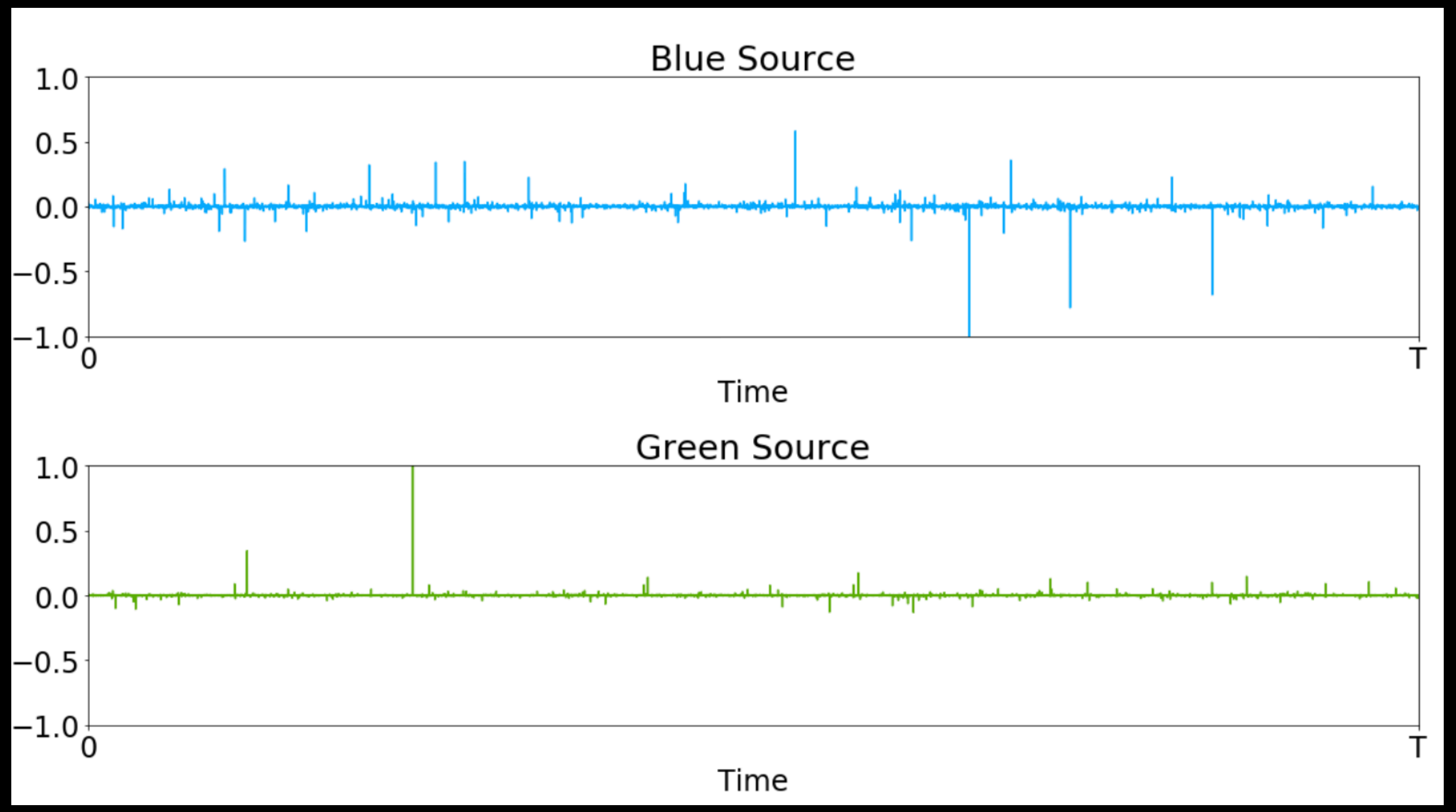
Poll 4

- Choose all the tasks you can apply ICA on
  - Speech enhancement that separate speech from a mixed sound
  - removing artifacts, such as eye blinks, from EEG data and studies of the resting state network of the brain.
  - computer vision tasks like optical Imaging of neurons or face recognition
  - extracting structure from stock returns and predicting stock market prices
  - mobile phone communications
  - analysis of changes in gene expression over time in single cell RNA-sequencing experiments
  - Identify and Separate Bright Galaxy Clusters from the Low-frequency Radio Sky

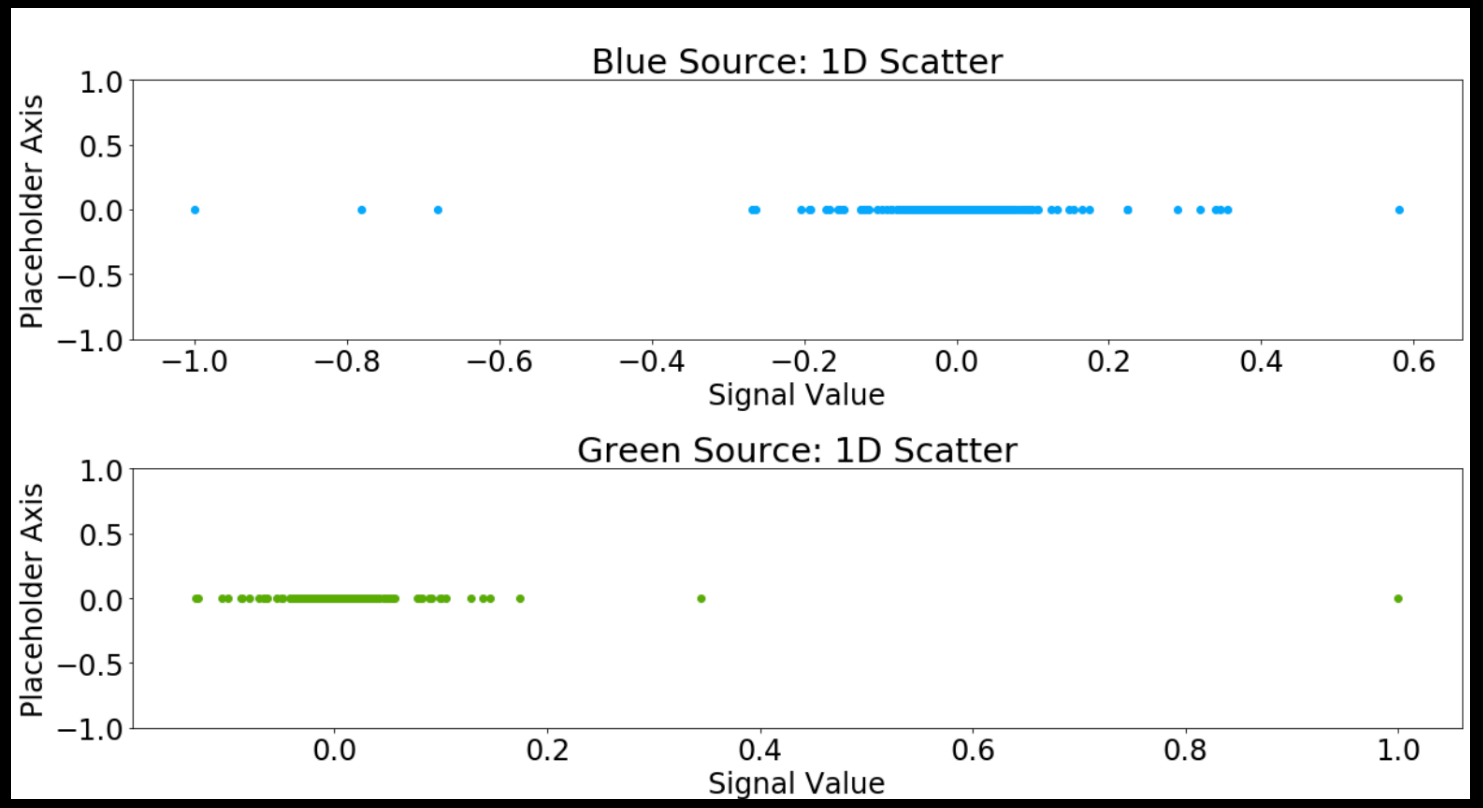
6 COMPARED WITH PCA

6.1 geometric intuition

- Observation $X(t)$



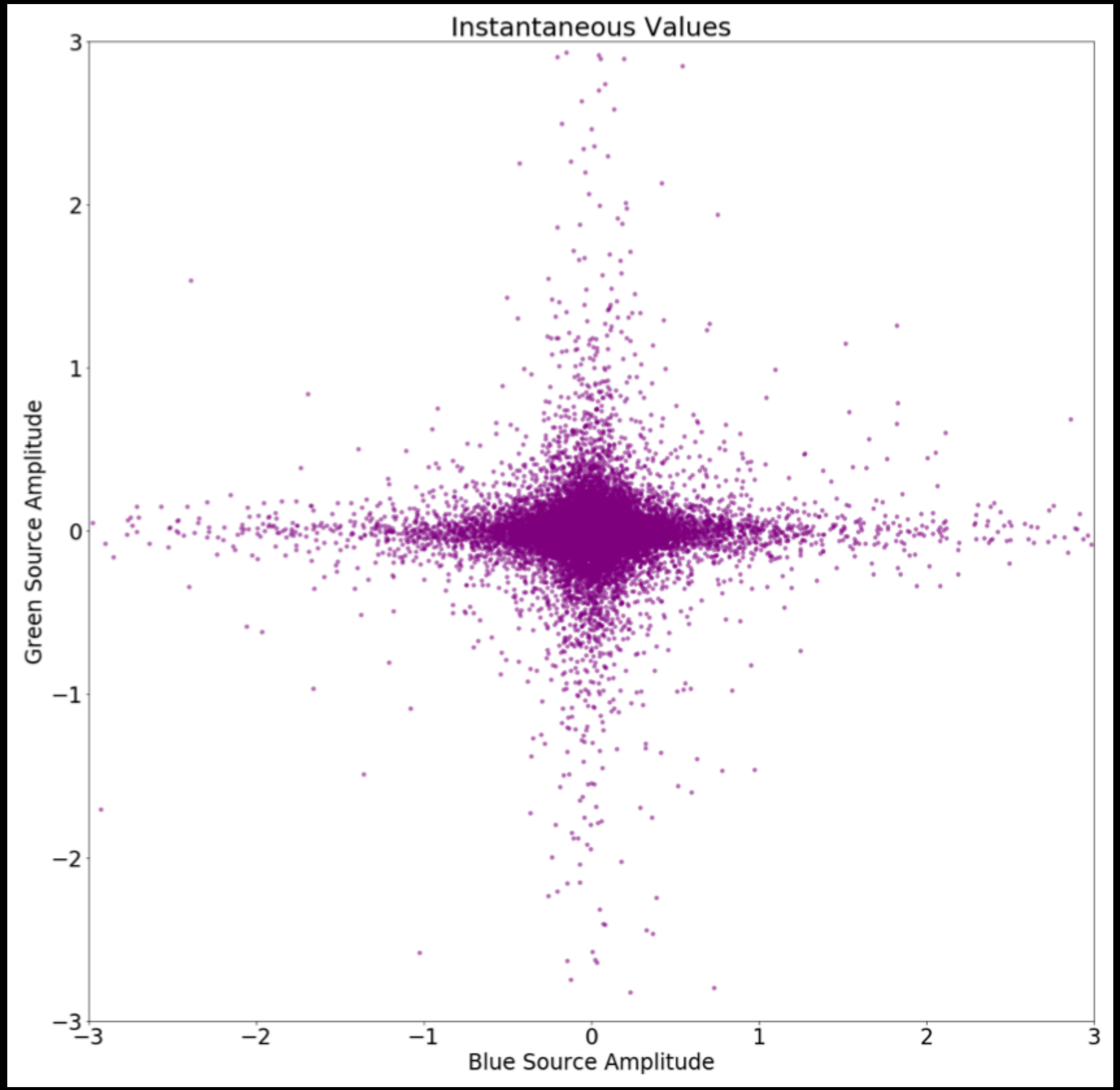
- Samples of random variable X



6 COMPARED WITH PCA

6.1 geometric intuition

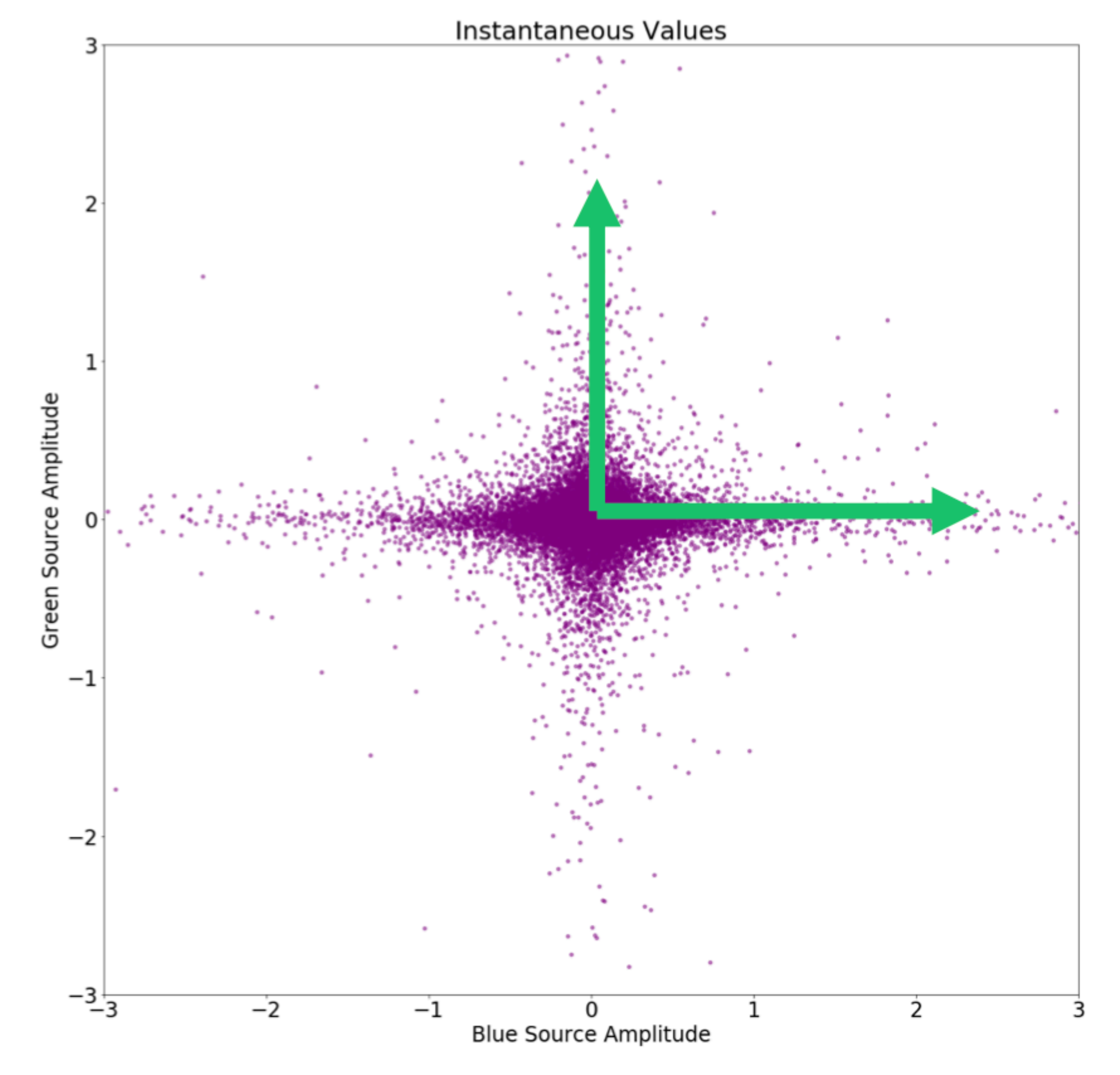
- Where are the 2 directions with maximum non-Gaussianity?



6 COMPARED WITH PCA

6.1 geometric intuition

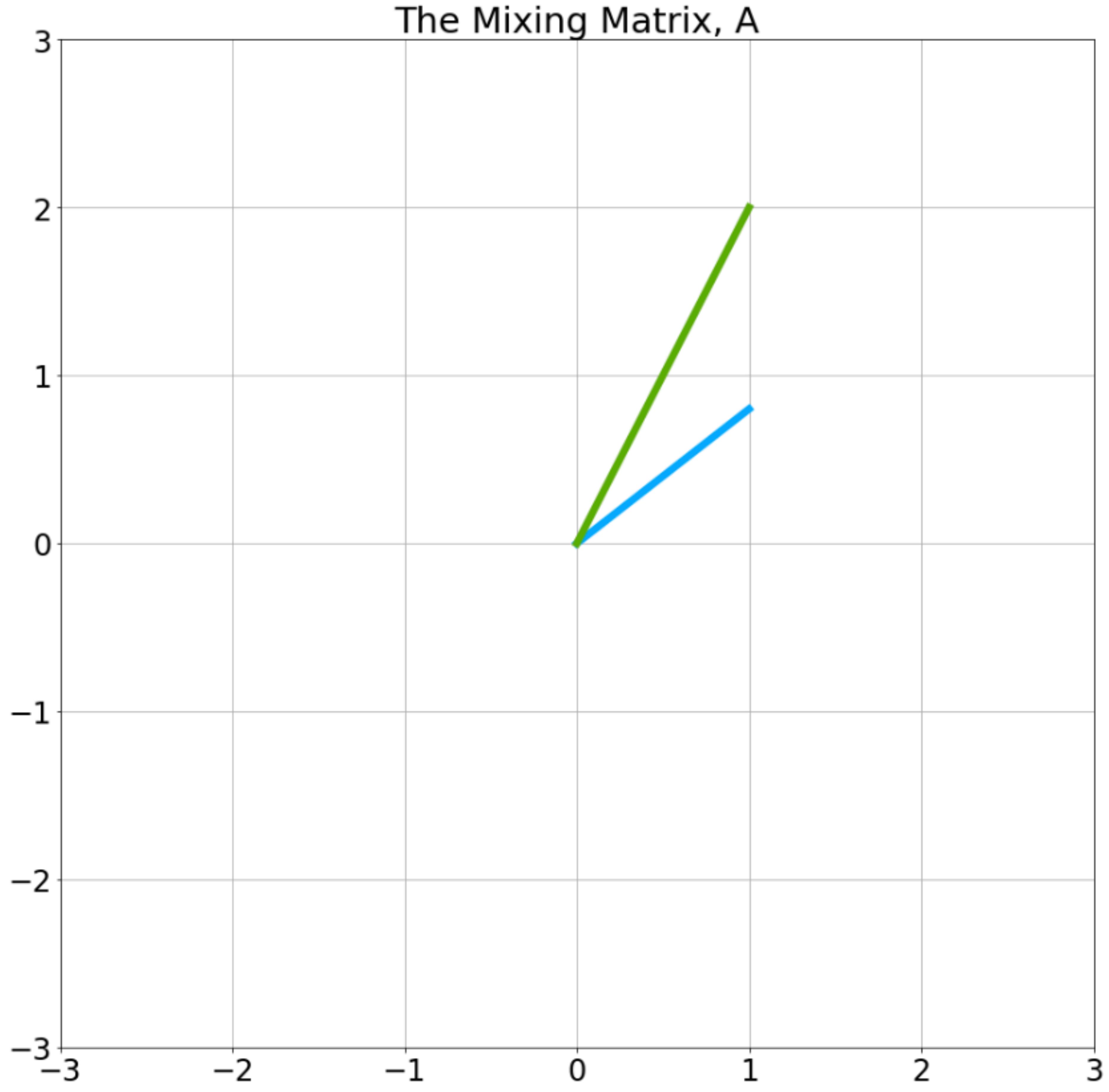
- Here are the 2 directions with maximum non-Gaussianity?



6 COMPARED WITH PCA

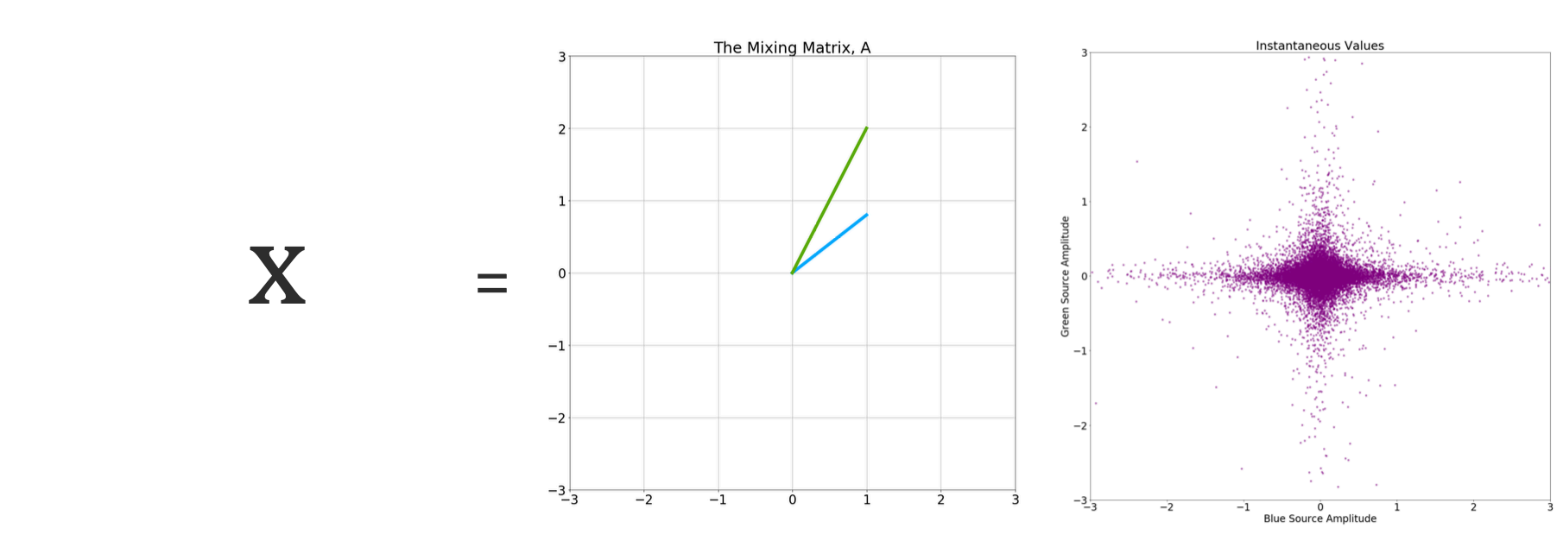
6.1 geometric intuition

$$A = \begin{bmatrix} 1.0 & 1.0 \\ 0.8 & 2.0 \end{bmatrix}$$



6 COMPARED WITH PCA
6.1 geometric intuition

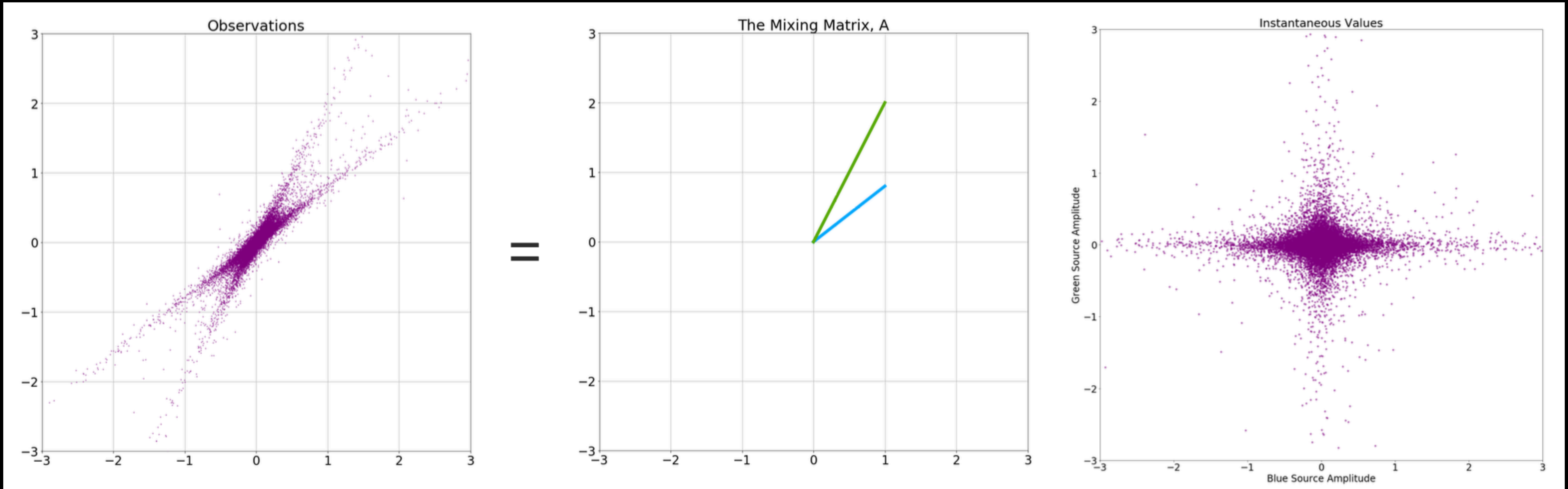
- What will X looks like?



6 COMPARED WITH PCA

6.1 geometric intuition

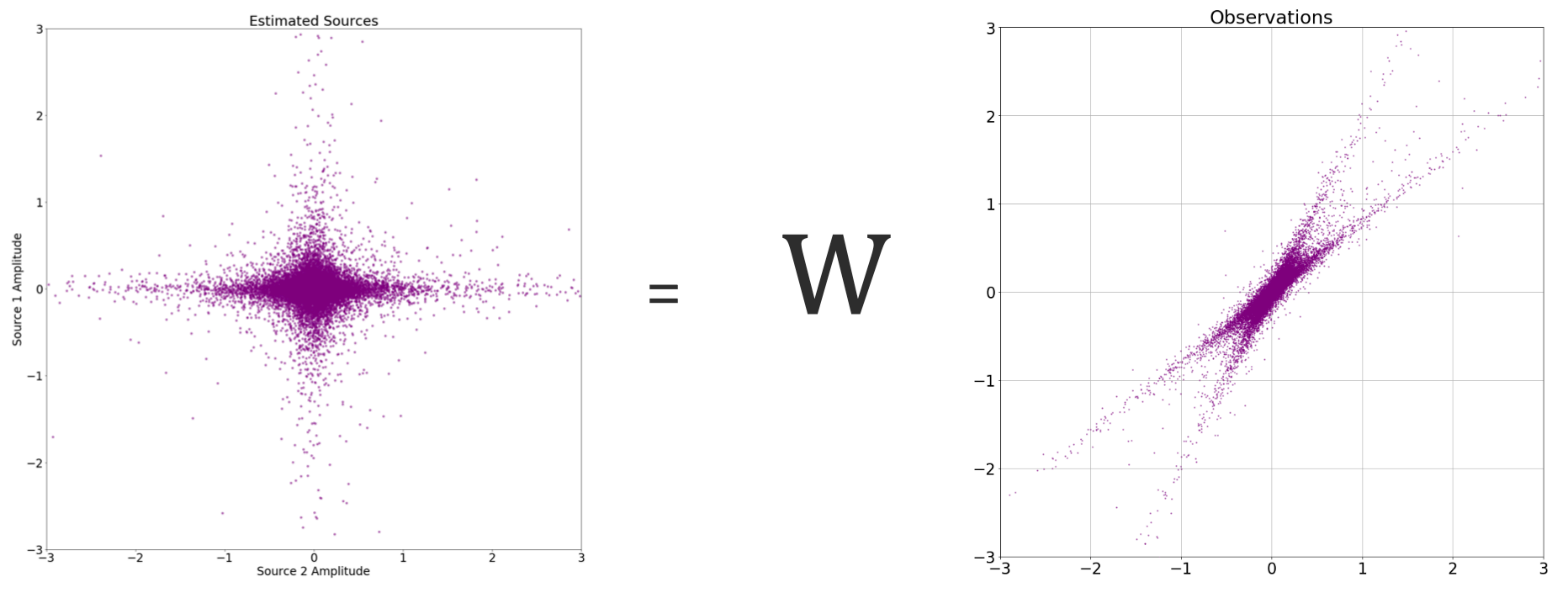
- $X = AS$



6 COMPARED WITH PCA

6.1 geometric intuition

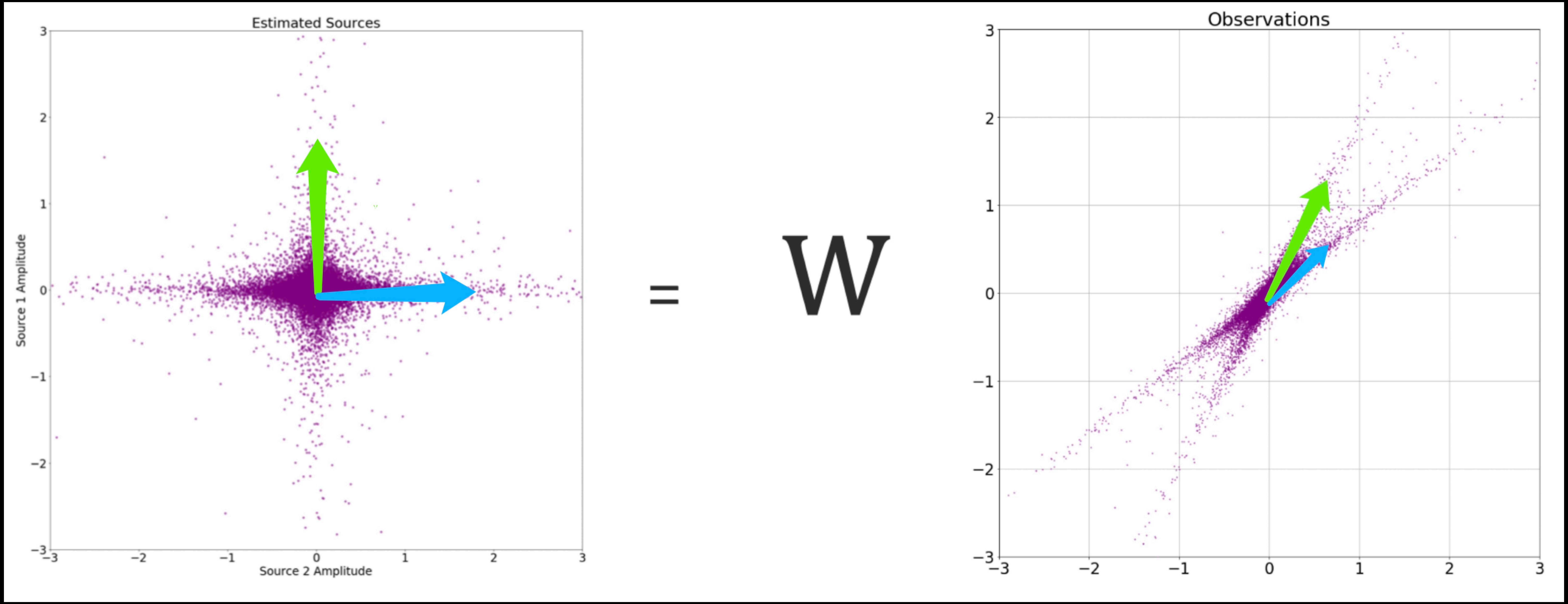
- $S = WX$



6 COMPARED WITH PCA

6.1 geometric intuition

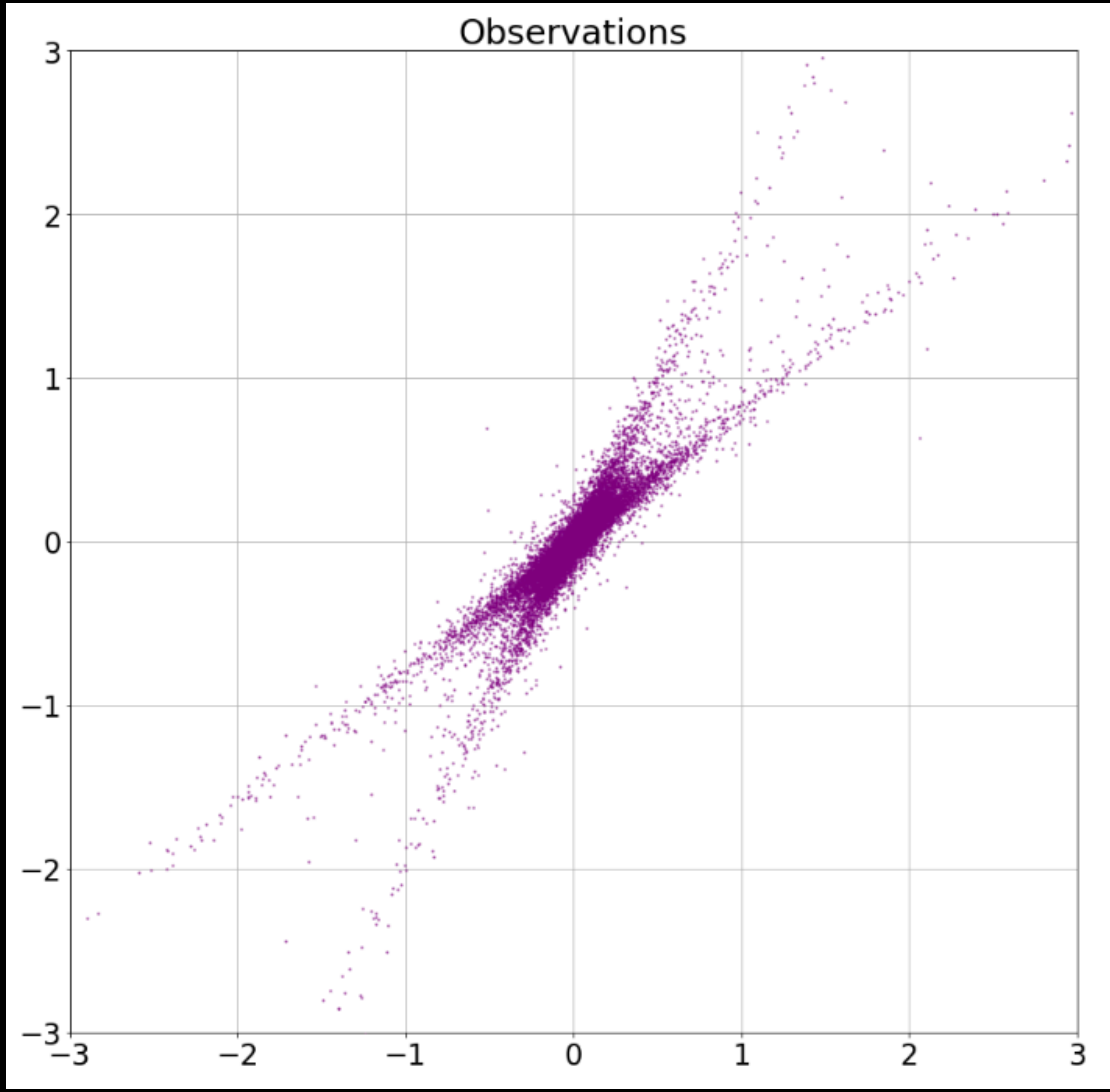
- $S = WX$



6 COMPARED WITH PCA

6.1 geometric intuition

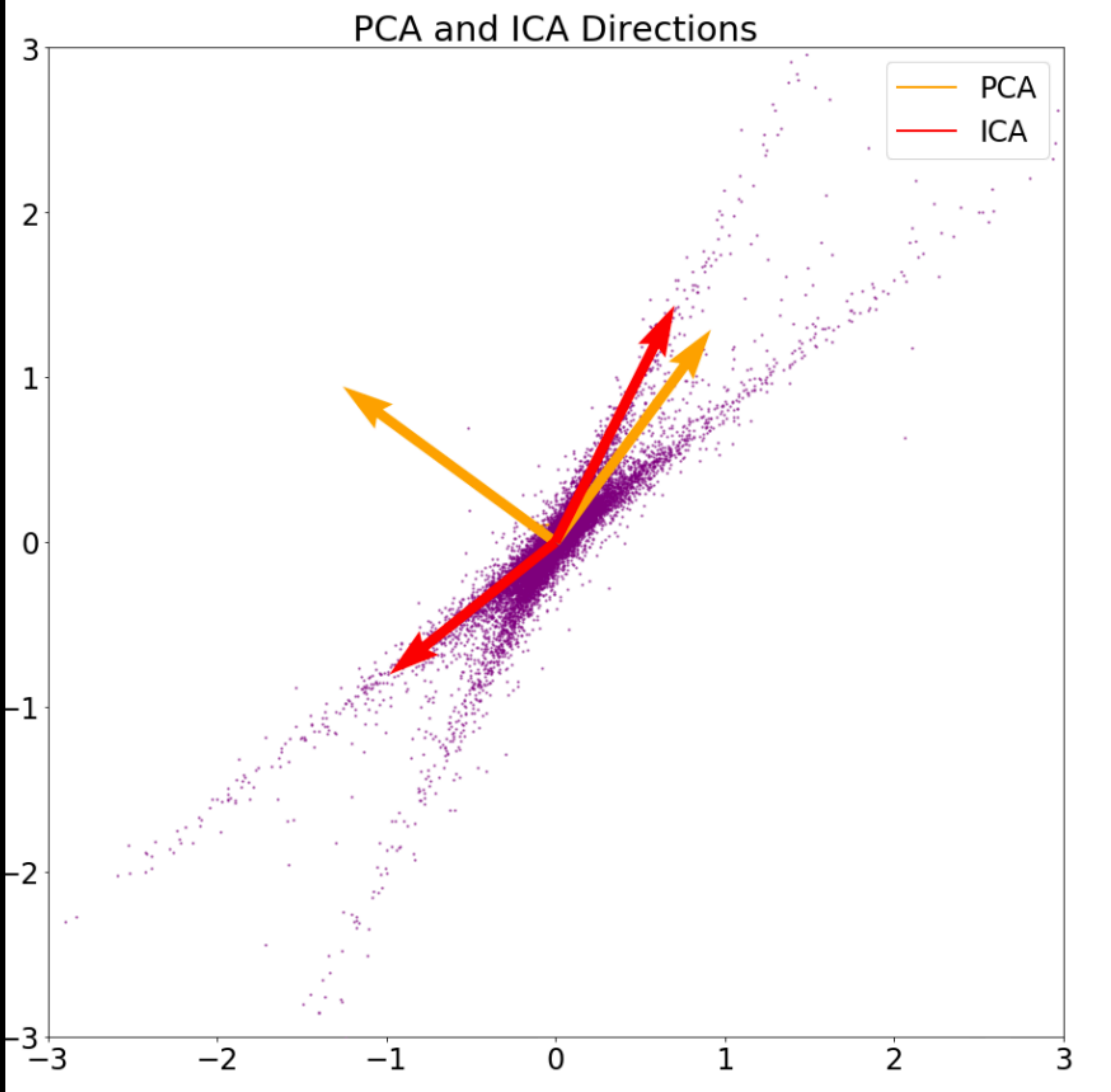
- Where is the ICA bases?
- Where is the PCA bases?



6 COMPARED WITH PCA

6.1 geometric intuition

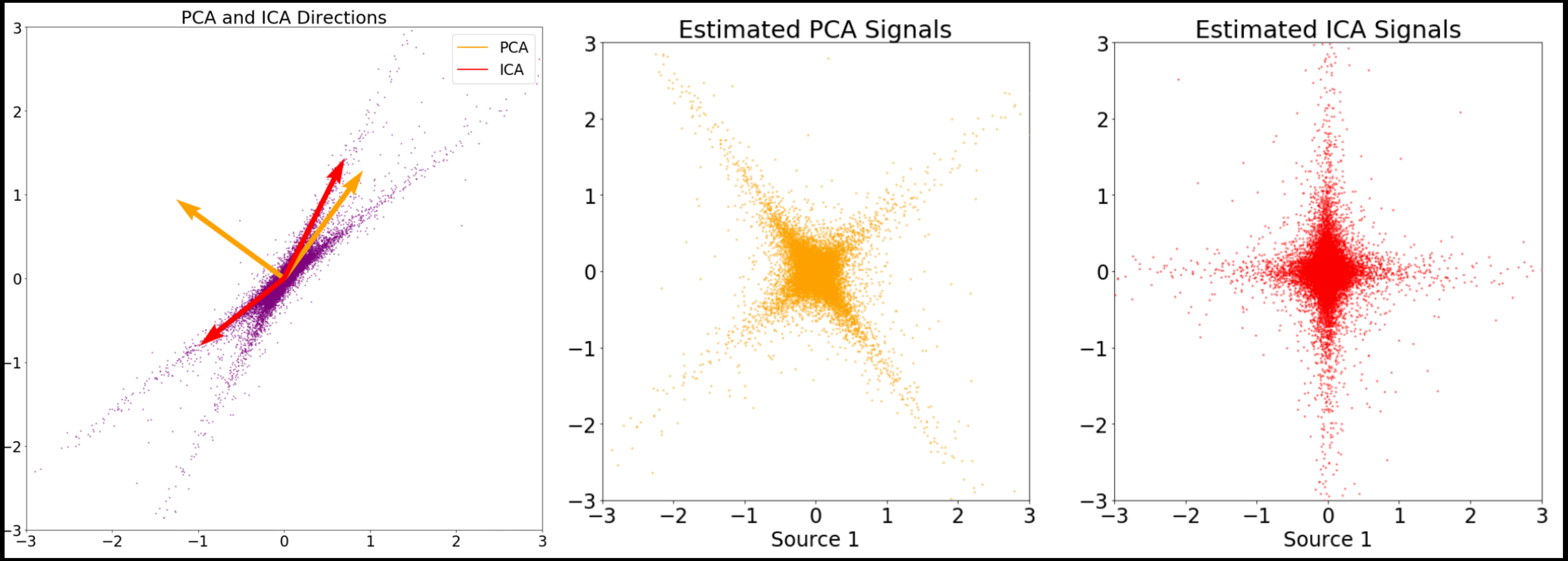
- here is the ICA bases and PCA bases



6 COMPARED WITH PCA

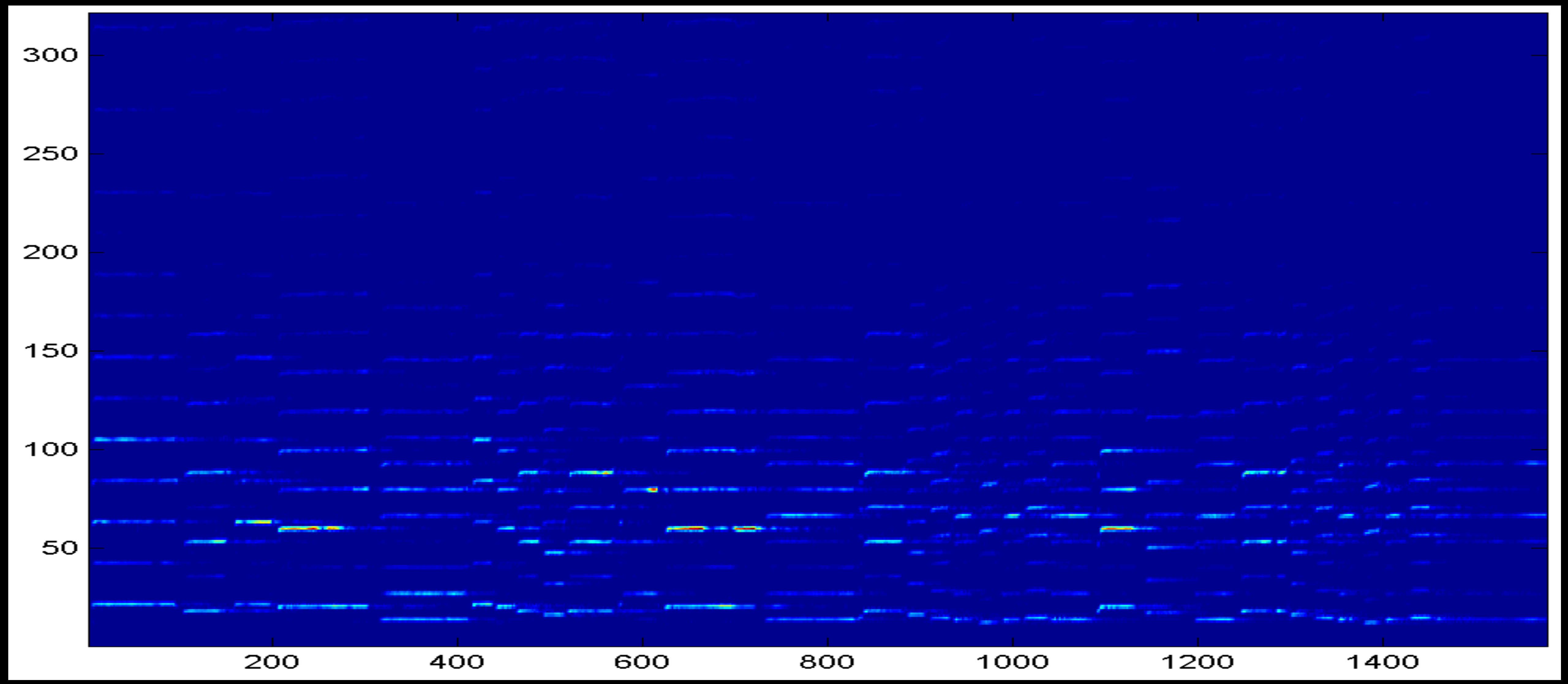
6.1 geometric intuition

- here is the ICA bases and PCA bases



6 COMPARED WITH PCA

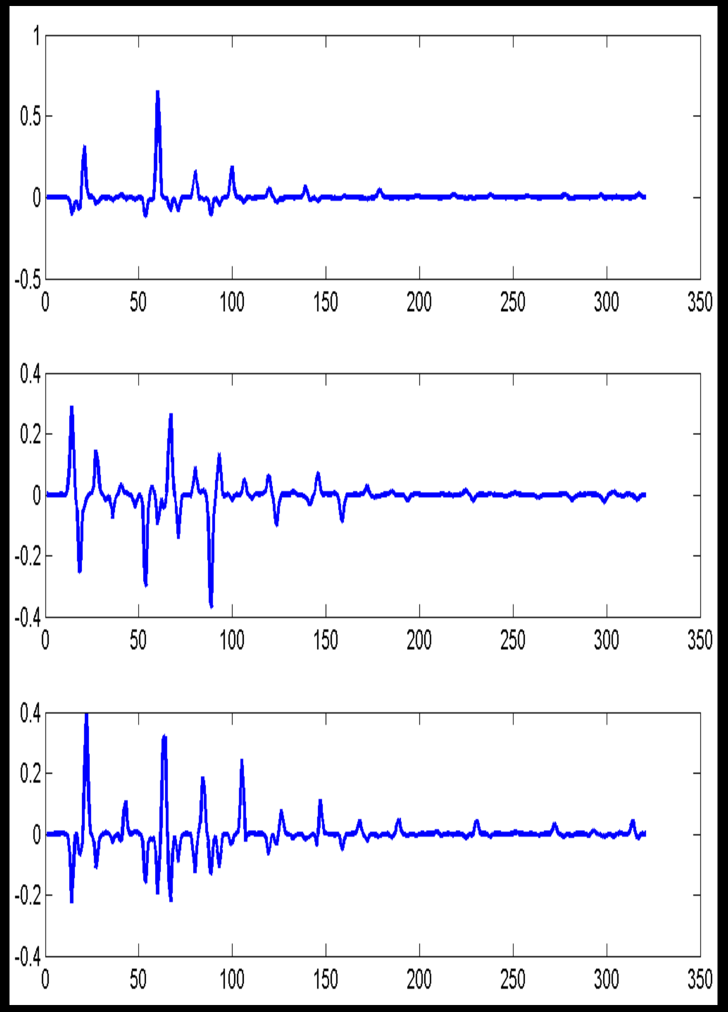
- There are 12 notes in the segment, hence we try to estimate 12 notes..

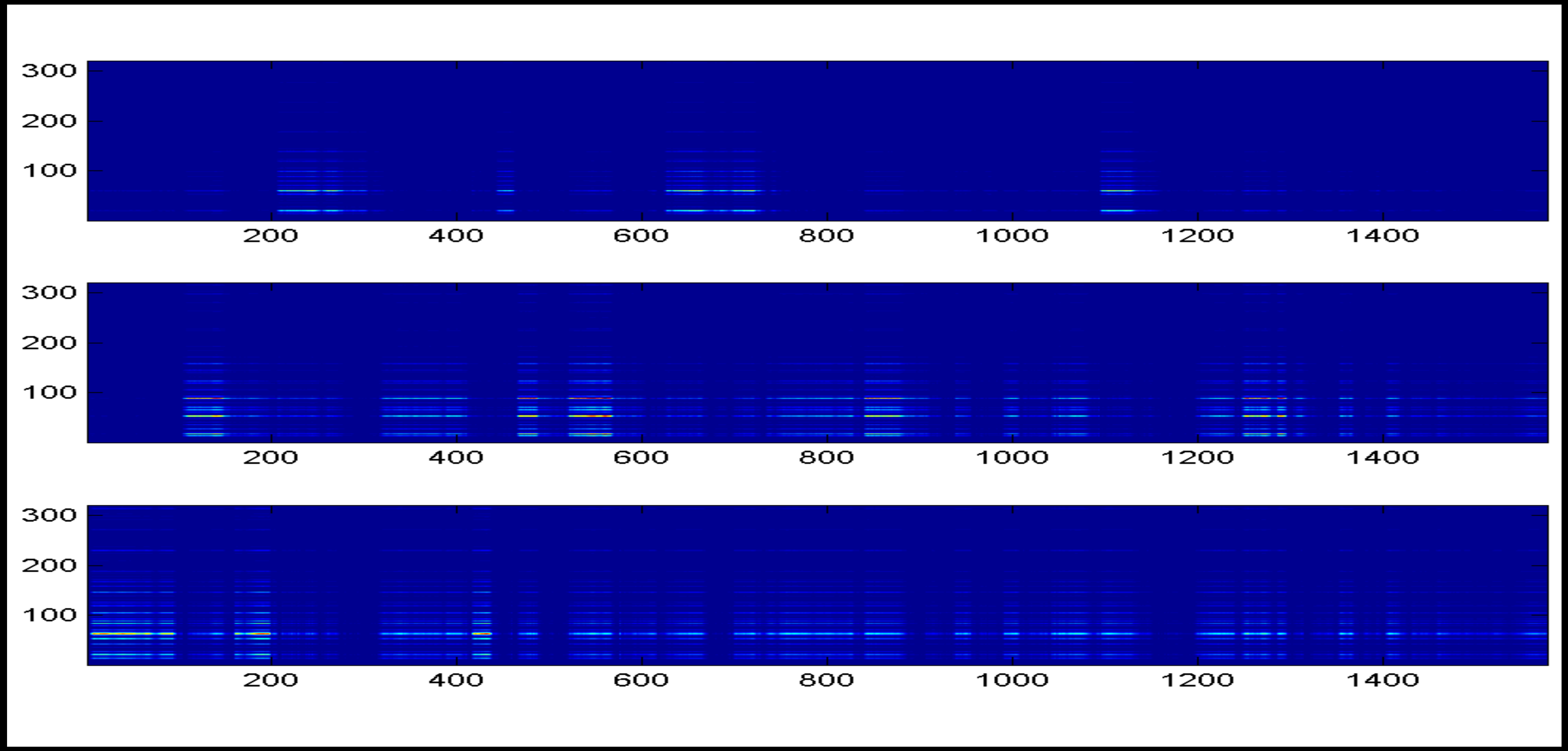


6 COMPARED WITH PCA

PCA solution

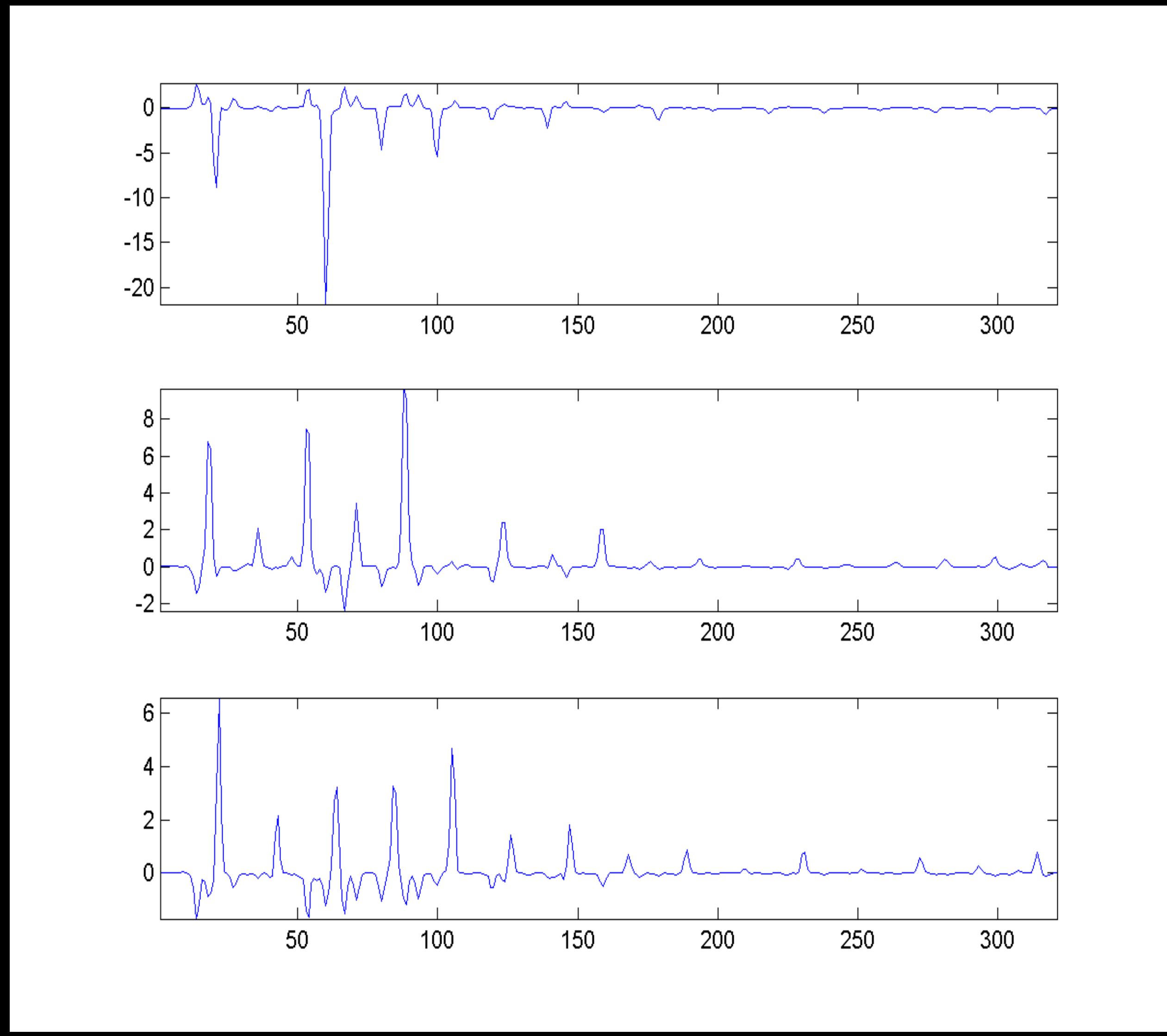
- There are 12 notes in the segment, hence we try to estimate 12 notes..

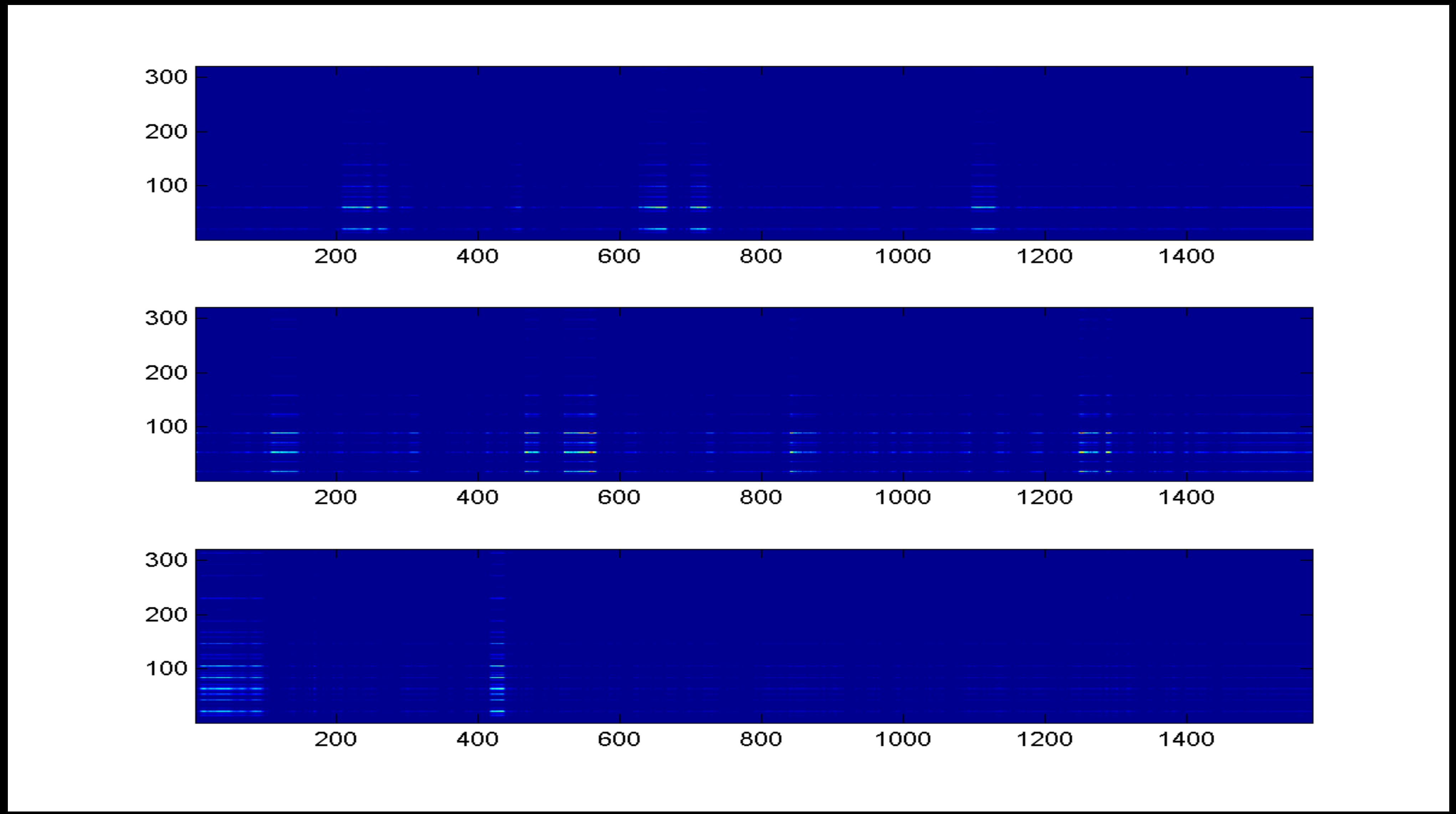




6 COMPARED WITH PCA

So how does this work: ICA solution





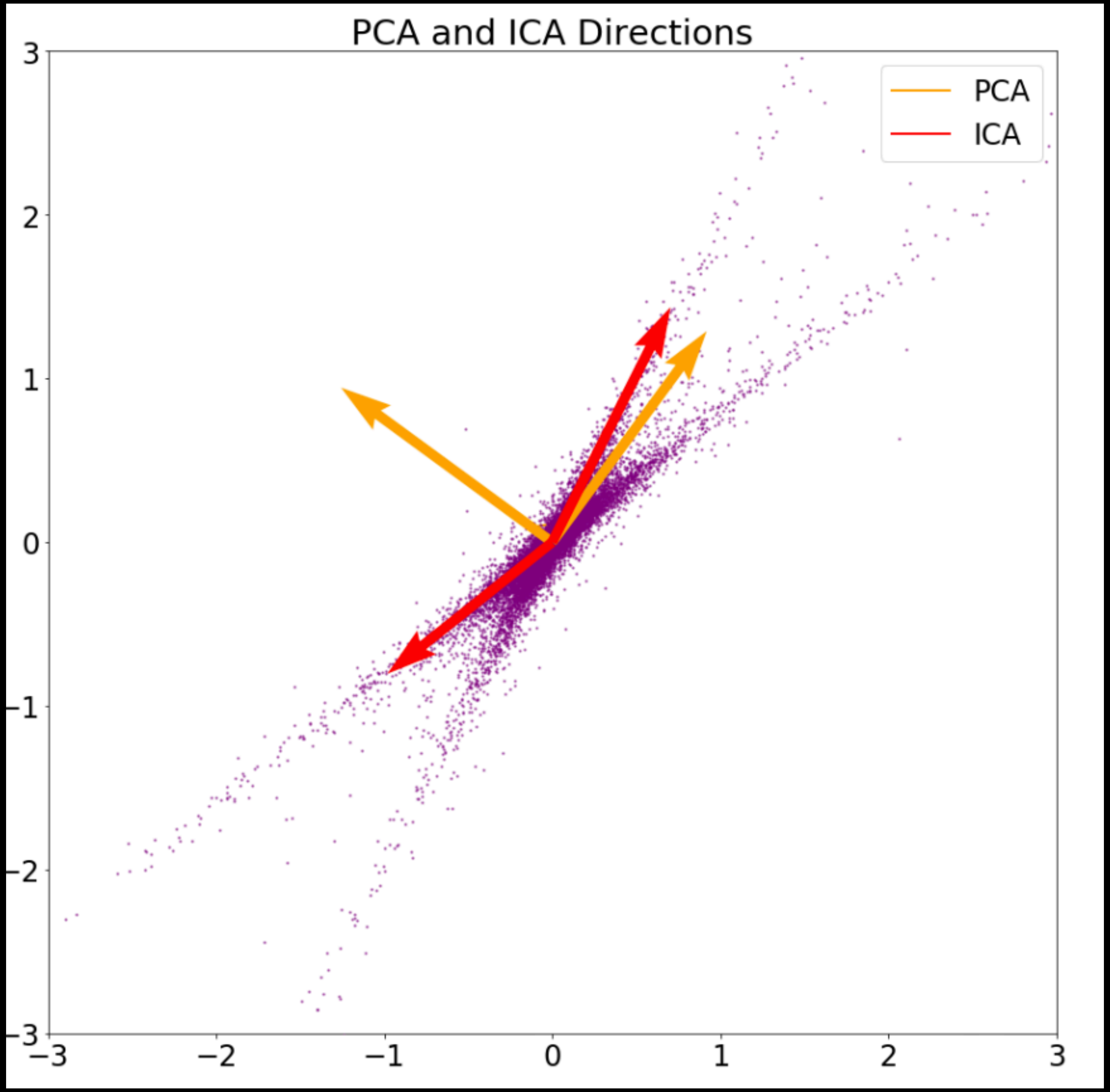
6 COMPARED WITH PCA

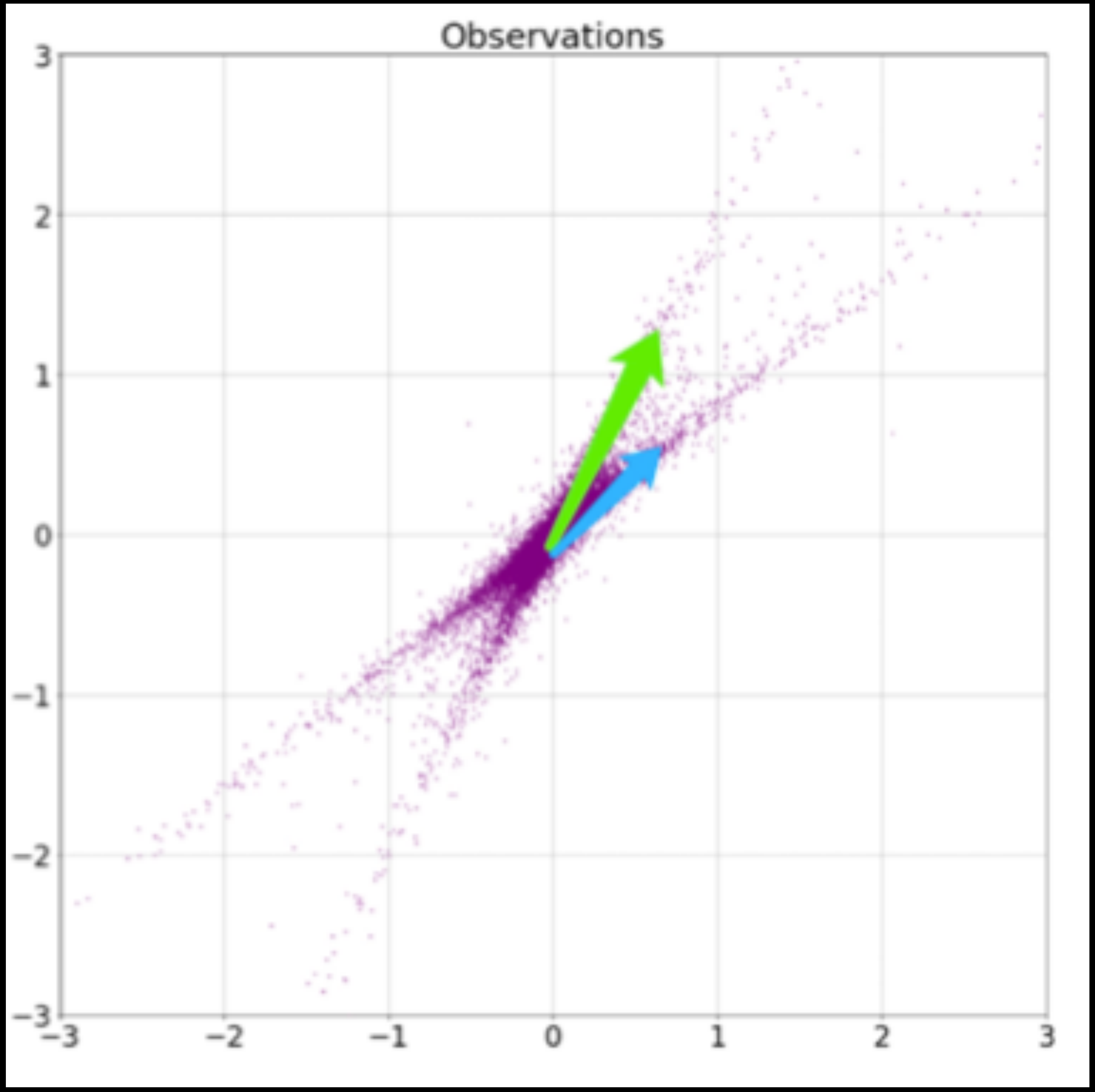
Discussion

- What’s your feeling when you hear those bases?
- Why doesn’t ICA work as well as we’d expect?

6 COMPARED WITH PCA

6.2 Unique Basis





6 COMPARED WITH PCA

6.2 Unique Basis

- permuted order: ICA basis has no sense of order
- Get K independent directions, but does not have a notion of the “best” direction
- Scale and sign: does not have sense of scaling

6 COMPARED WITH PCA

6.2 Unique Basis

- How to compute weight and reconstruction error?

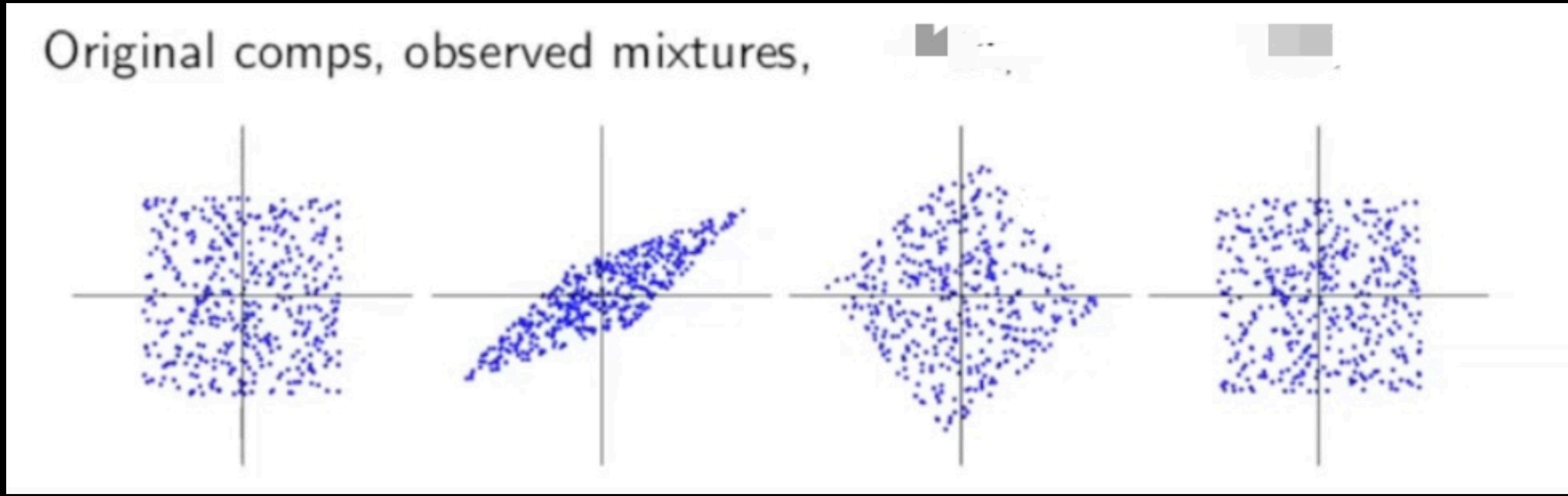
6 COMPARED WITH PCA

6.2 Unique Basis

- Do not use the projection to each ICA bases, because they could be correlated !!!
- use pseudo inverse to evaluate the projection to the whole surface the bases generated
- you can compute the weight with linear algebra :-)

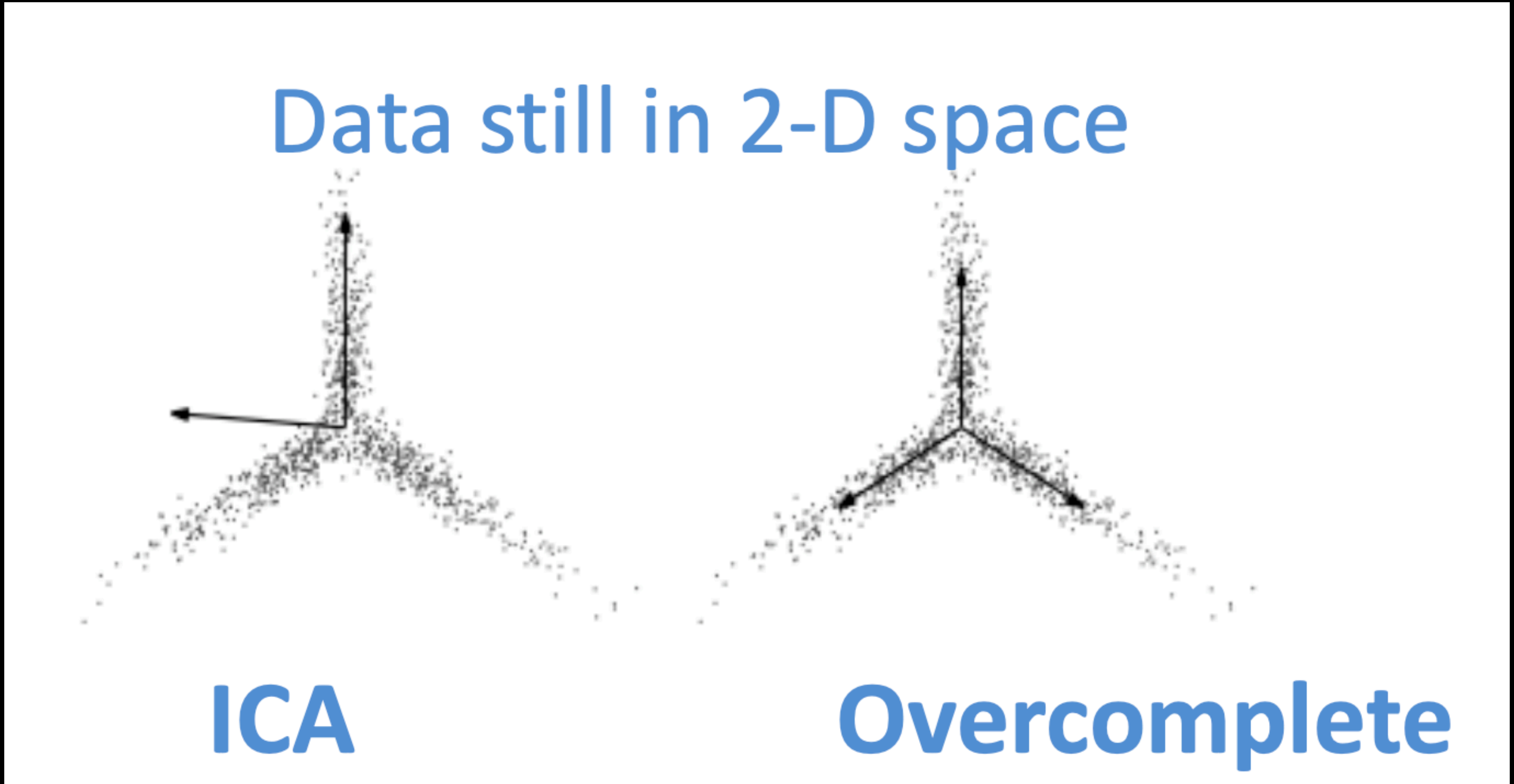
Poll 5

- Which figure in the upper image is more likely to be recovered by ICA?
    - the third one
    - the last one



• 7 DISADVANTAGE WITH REFINEMENT

7.0 What if the number of S is significantly larger than X?



• 7 DISADVANTAGE WITH REFINEMENT

7.1 Introduction of Linear Noisy ICA

- Let $Z = X + n$ be the observation with white Gaussian noise n
- n is uncorrelated with the true observation $X = AS$
- methods
  - FFT, low-pass filter, iFFT (inefficient)
  - wavelet shrinkage (not explicitly take advantage of data statistics)
  - median filter (not explicitly take advantage of data statistics)
  - Sparse Code Shrinkage (ICA related methods)

• 7 DISADVANTAGE WITH REFINEMENT

7.1 Introduction of Linear Noisy ICA

- $Z = X + n$
- $WZ = S + Wn$, where W is the best orthognal approximation of the inverse of the ICA mixing matrix
- noise term Wn is still Gaussian and white and the density of $S = Wx$ becomes highly non-Gaussian with a high positive kurtosis (with some good assumption on S)

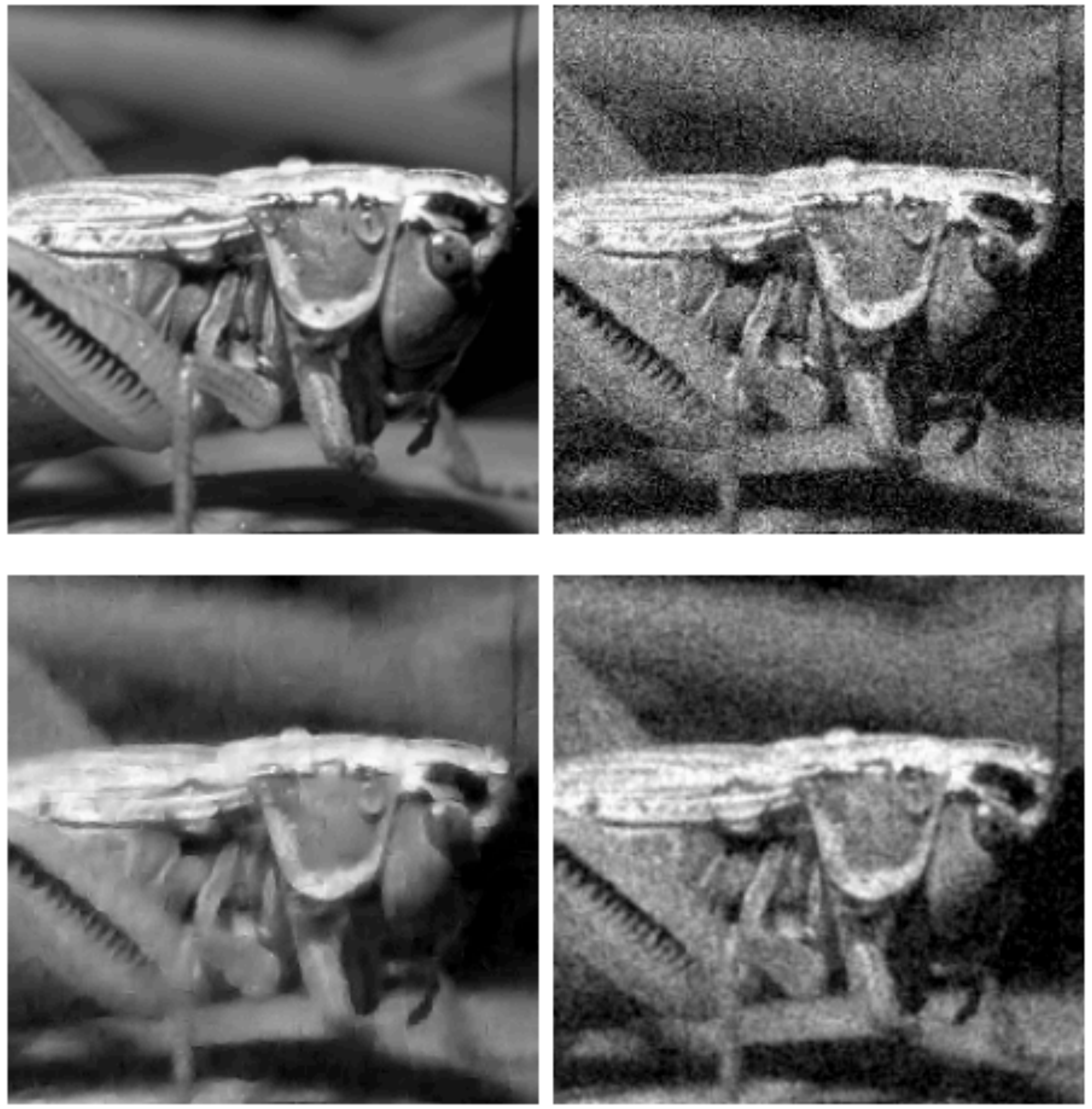
• 7 DISADVANTAGE WITH REFINEMENT

7.1 Introduction of Linear Noisy ICA

- Assuming S has a specific non-Gaussian distribution (for example, Laplician distribution), we can evaluate shirkage function $S = g\left(Wz\right)$ explicitly
- The optimal (maximum likelihood) of $\hat{X} = W^t S = W^t g\left(WZ\right)$ can be evaluate by a refined algorithm of Fast-ICA

• 7 DISADVANTAGE WITH REFINEMENT

7.1 Introduction of Linear Noisy ICA



e 15: (from [22]). *An experiment in denoising. Upper left: original image. Upper right: original corrupted with noise; the noise level is 50 %. Lower left: the recovered image after applying sparse shrinkage. Lower right: for comparison, a wiener filtered image.*

• 7 DISADVANTAGE WITH REFINEMENT

7.2 Introduction of Nonlinear ICA

- $X = f\left(S \mid \theta\right)$, where f is a non-linear function and θ is the parameter of the function
- in general, non-linear ICA do not have unique solution
  - Suppose s_1 and s_2 are independent sources, and $X = \left(x_1, x_2\right)$ is the observation
  - Define $g\left(a, b\right) := P\left(s_2 \leq b \mid s_1 = a\right)$ for all a and b
  - Random variable $y = g(s_1, s_2)$ is independent of s_1
  - It is absurd to regard s_1 and y to be the independent component of X with another non-linear function f

• 7 DISADVANTAGE WITH REFINEMENT

7.2 Introduction of Nonlinear ICA

- The construction is dened recursively as follows. Assume that we have already m independent random variables y1; : : : ; ym which follow a joint uniform distribution in [0; 1]m . (It is not a restriction to assume that the distributions of the yi are uniform: this follows directly from the recursion, as will be seen below.) Denote by x any random variable, and by a1; : : : ; am; b some non-random scalars. Dene

- $$\begin{aligned} g(a_1, \ldots, a_m, b; p_{y,x}) &= P(x \leq b | y_1 = a_1, \ldots, y_m = a_m) \qquad (2) \\ &= \frac{\int_{-\infty}^{b} p_{y,x}(a_1, \ldots, a_m, \xi) d\xi}{p_y(a_1, \ldots, a_m)} \end{aligned}$$

- where py() and py;x() are the (marginal) probability densities of (y1; : : : ; ym) and (y1; : : : ; ym; x), respectively (it is assumed here implicitly that such densities exist), and P (j) denotes the conditional probability.

- # 7 DISADVANTAGE WITH REFINEMENT

7.2 Introduction of Nonlinear ICA

Theorem 1 *Assume that* $y_1, \ldots, y_m$ *are independent scalar random variables which follow a joint uniform distribution in the unit cube* $[0,1]^m$*. Let* x *be any scalar random variable (such that the joint distribution of* $y_1, \ldots, y_m, x$ *has a probability density with respect to the Lebesgue measure of* $\mathbb{R}^{m+1}$*). Define* g *as in (2), and set*

$$y_{m+1} = g(y_1, \ldots, y_m, x; p_{y,x}). \tag{3}$$

Then y_{m+1} *is independent from the* $y_1, \ldots, y_m$*. In particular, the variables* $y_1, \ldots, y_{m+1}$ *are jointly uniformly distributed in the unit cube* $[0,1]^{m+1}$*.*

• 7 DISADVANTAGE WITH REFINEMENT

7.2 Introduction of Nonlinear ICA

where $c_1, c_2, \ldots$ are some irrelevant quantities, and

$$K = \frac{p_{y,x}(v_1, \ldots, v_m, \xi)}{p_y(v_1, \ldots, v_m)}. \tag{6}$$

The determinant of JF equals K. Thus, one obtains the density p_{y+} of the vector $(y_1, \ldots, y_m, y_{m+1})$ as

$$\begin{aligned} p_{y+}(v_1, \ldots, v_{m+1}) &= p_{y,x}(v_1, \ldots, v_m, \xi) \left[\frac{p_{y,x}(v_1, \ldots, v_m, \xi)}{p_y(v_1, \ldots, v_m)} \right]^{-1} \\ &= p_y(v_1, \ldots, v_m) \end{aligned} \tag{7}$$

From (2) it follows that $y_{m+1} \in [0, 1]$. Thus (7) implies that p_{y+} is a uniform density in $[0, 1]^{m+1}$, which implies that the $y_1, \ldots, y_{m+1}$ are mutually independent (Pajunen et al., 1996). □

• 7 DISADVANTAGE WITH REFINEMENT

7.2 Introduction of Nonlinear ICA

Proof. Denote by

$$F(v_1, \ldots, v_m, \xi) = (v_1, \ldots, v_m, g(v_1, \ldots, v_m, \xi; p_{y,x})) \tag{4}$$

the transformation made on the vector $(y_1, \ldots, y_m, x)$ to obtain $(y_1, \ldots, y_{m+1})$. The Jacobian of this transformation equals

$$JF(v_1, \ldots, v_m, \xi) = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ c_1 & c_2 & \ldots & K \end{bmatrix} \tag{5}$$

• 7 DISADVANTAGE WITH REFINEMENT

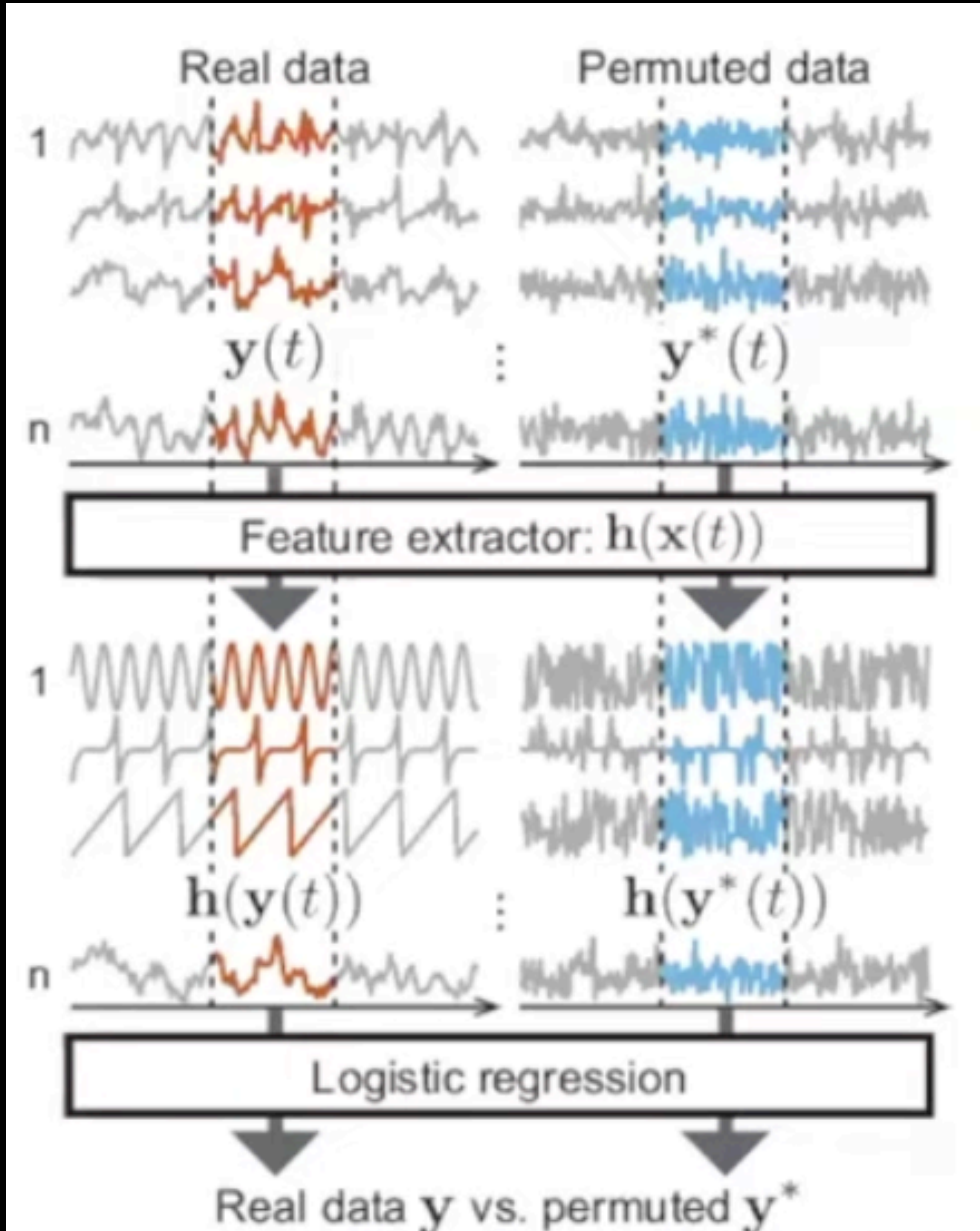
7.2 Introduction of Nonlinear ICA

- You need some extra assumption for X and S to make the solution unique
  - Use temporal structures in the time series (non-stationary and autocorrelation for stationary)
  - Use an auxiliary variable such as multimodal for audio and video
  - You can combined it with some useful estimation methods like self-supervised learning or Variational autoencoder (VAE)

• 7 DISADVANTAGE WITH REFINEMENT

7.2 Introduction of Nonlinear ICA

- Permutation-constructive learning (OCL)
- Take short time windows for autocorrelation for the stationary data$y(t) = \langle x(t), x(t-1) \rangle$
- Permute $x(t)$ and evaluate $y^*(t) = \langle x(t), x(t^*) \rangle$
- Use MLP with hidden layer $h(x)$ with dimension n to predict
- MLP turn nonlinear ICA to linear ICA



• 7 DISADVANTAGE WITH REFINEMENT

7.2 Introduction of Nonlinear ICA

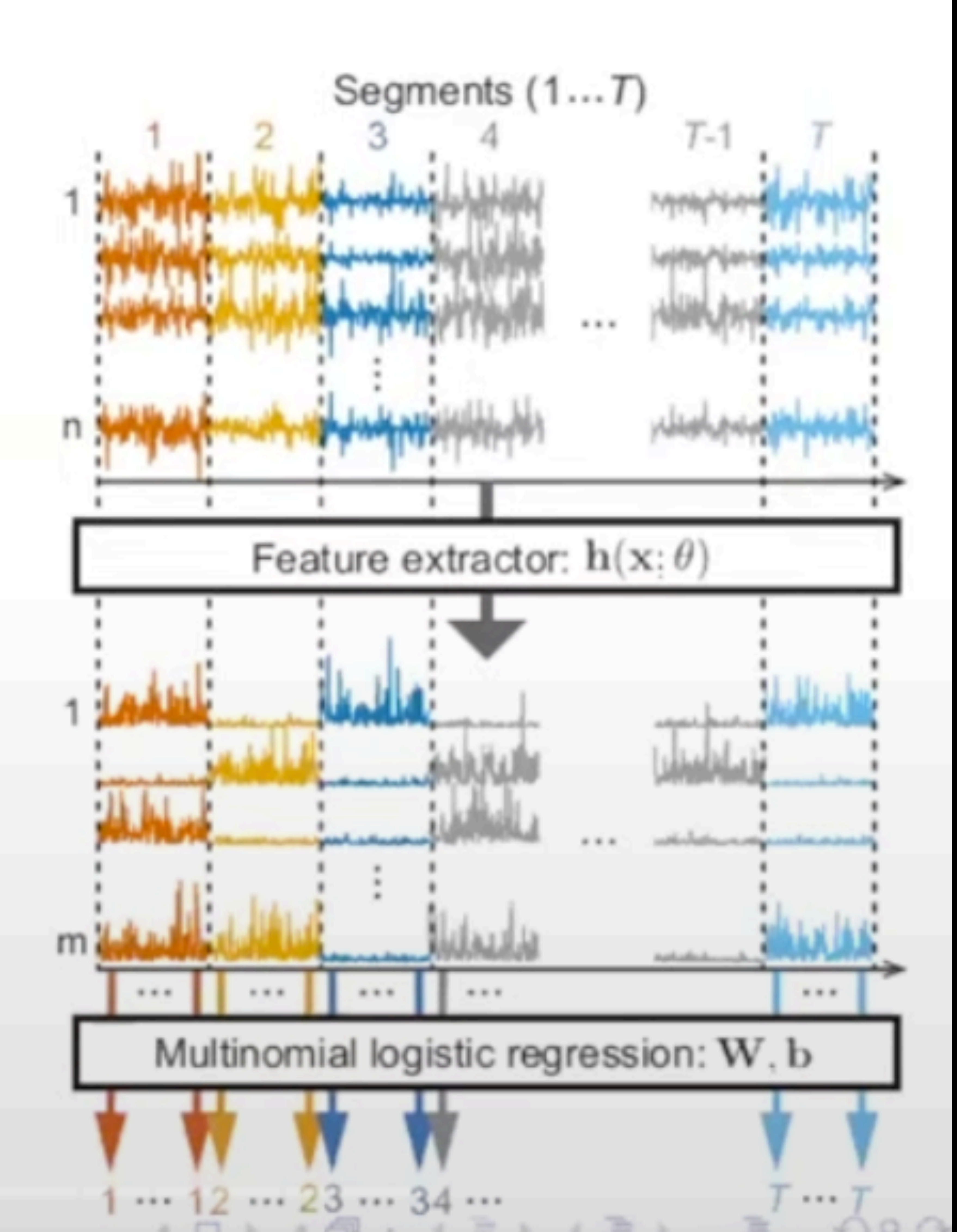
- Time-contrastive learning (TCL)

- Non-stationary time series $s(t)$ are mixed to the observation data follows nonlinear ICA $\mathbb{R}^n$ method $x(t) = f(s(t))$ where $f(\cdot)$ is smooth and invertible map from $\mathbb{R}^n$ to $\mathbb{R}^n$

- Chop them to different segmentation

- Use MLP with hidden layer $h(x)$ with dimension n to predict

- MLP turn nonlinear ICA to linear ICA



• 7 DISADVANTAGE WITH REFINEMENT

7.2 Introduction of Nonlinear ICA

- classical reference without deep learning: Nonlinear Independent Component Analysis: Existence and Uniqueness Results http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=EB9962CA3821E70E34A14B61A23082DF?doi=10.1.1.54.6547&rep=rep1&type=pdf

- non-linear ICA in 21 century

  - https://www.youtube.com/watch?v=_cBLSNRWt8c&t=761s

  - Variational Autoencoders and Nonlinear ICA: A Unifying Framework https://arxiv.org/abs/1907.04809

- **7 DISADVANTAGE WITH REFINEMENT**

7.3 introduction to Quantum ICA

8 REFERENCE

- Independent Component Analysis A Tutorial https://www.cs.jhu.edu/~ayuille/courses/Stat161-261-Spring14/HyvO00-icatut.pdf
- COMPARATIVE ANALYSIS OF THE ICA ALGORITHMS APPLIED ON A 2D SIGNAL http://oaji.net/articles/2017/4249-1487183273.pdf
- ICA ppt from prof. Bhiksha Raj
- ICA ppt from Patrick, TA in previous year