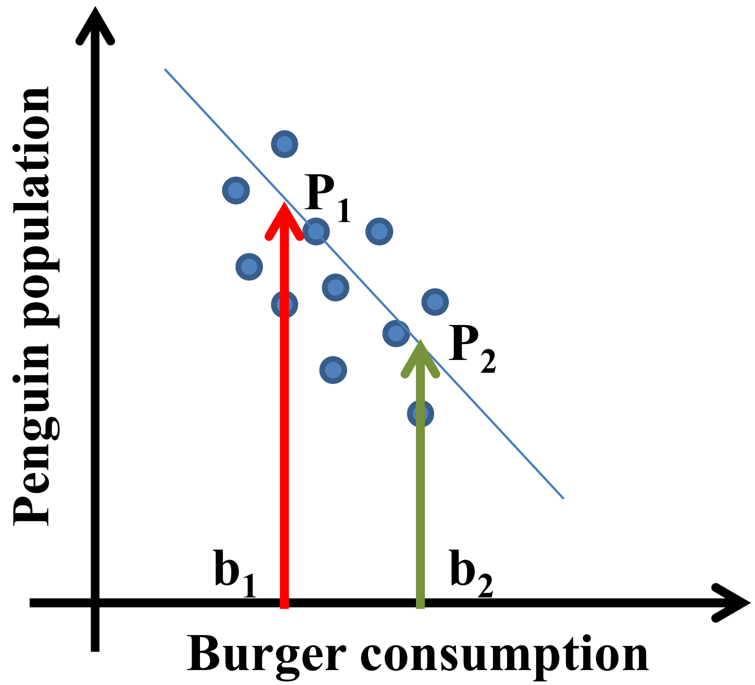
Machine Learning for Signal Processing

Independent Component Analysis

Instructor: Bhiksha Raj

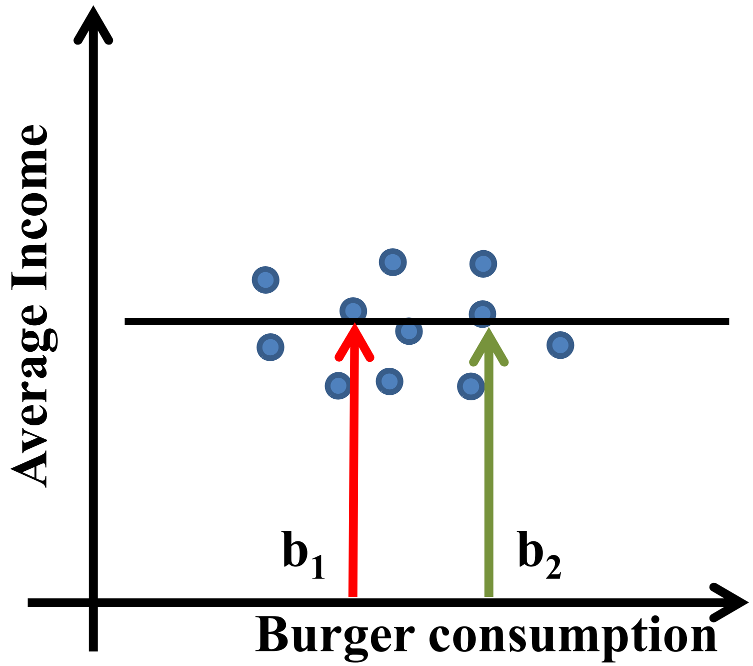
Slides (the good ones) are by Patrick Conrey

Recap: Correlated Variables



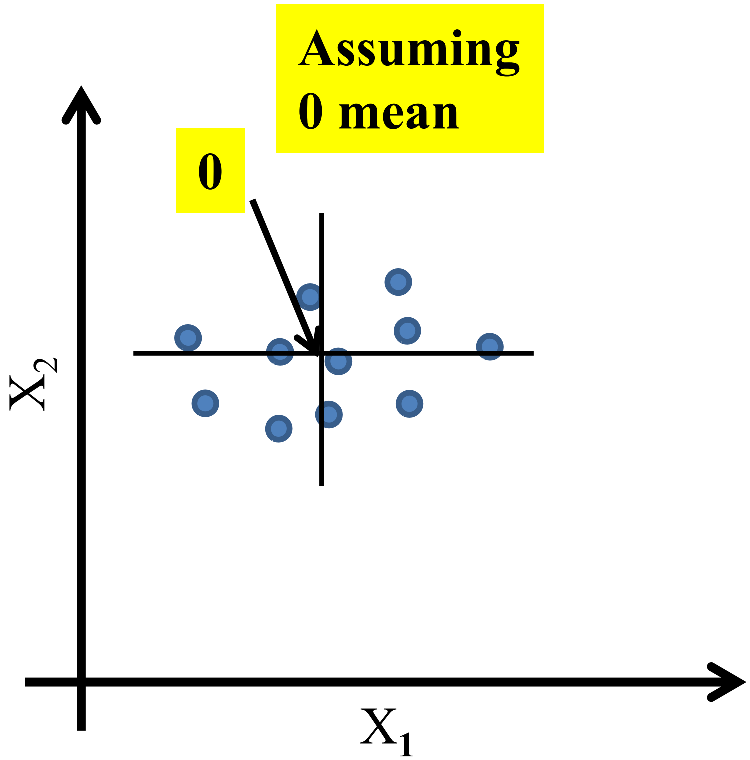
- Expected value of Y given X varies with X
  - And vice versa

Uncorrelatedness



- Knowing X does not tell you what the *average* value of Y is
  – And vice versa

Recap: Uncorrelatedness



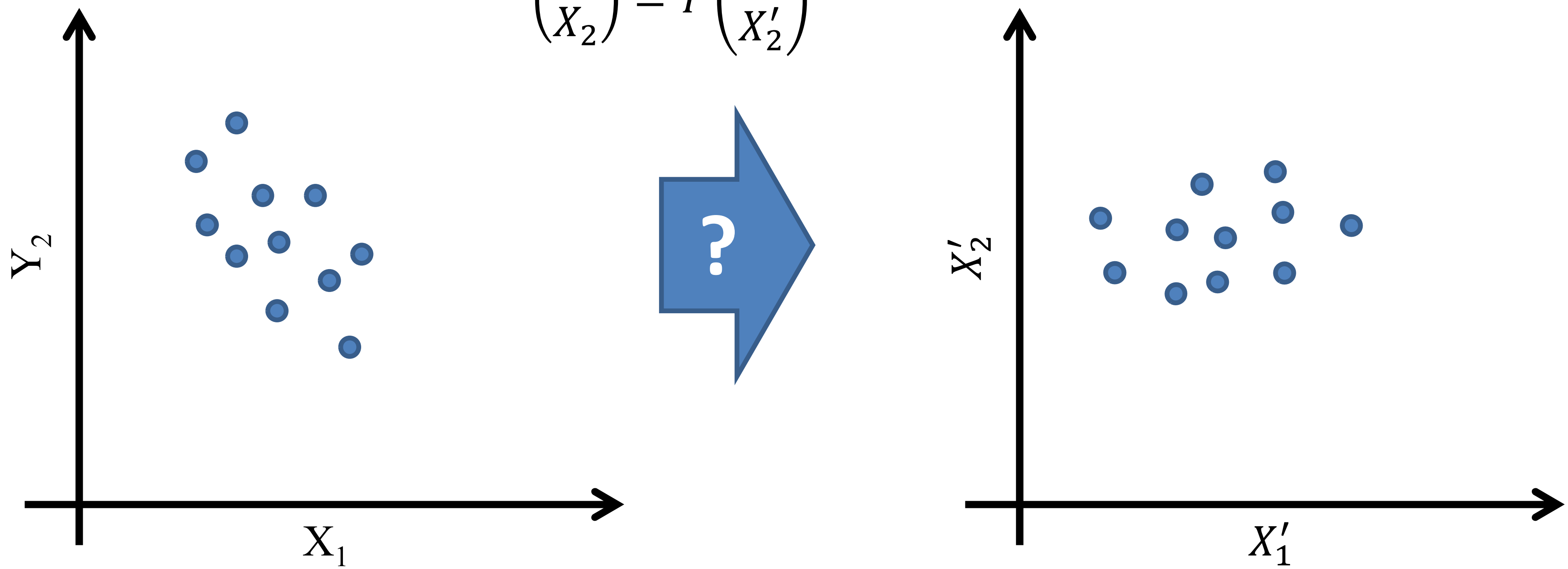
- $E[X_1] = constant$
- $E[X_2] = constant$
- $E[X_2|X_1] = constant$
- $E[X_1 X_2] = E[X_1]E[X_2]$
- All will be 0 for centered data

$$E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} (X_1 \quad X_2)\right] = E\begin{pmatrix} X_1^2 & X_2 X_1 \\ X_1 X_2 & X_2^2 \end{pmatrix} = \begin{pmatrix} E[X_1^2] & 0 \\ 0 & E[X_2^2] \end{pmatrix} = diagonal\ matrix$$

- If $\mathbf{X}$ is a matrix of vectors, $\mathbf{XX}^T$ = diagonal

Recap: Decorrelation

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = T \begin{pmatrix} X_1' \\ X_2' \end{pmatrix}$$



- So how does one transform the correlated variables (X_1, X_2) to the uncorrelated (X_1', X_2')

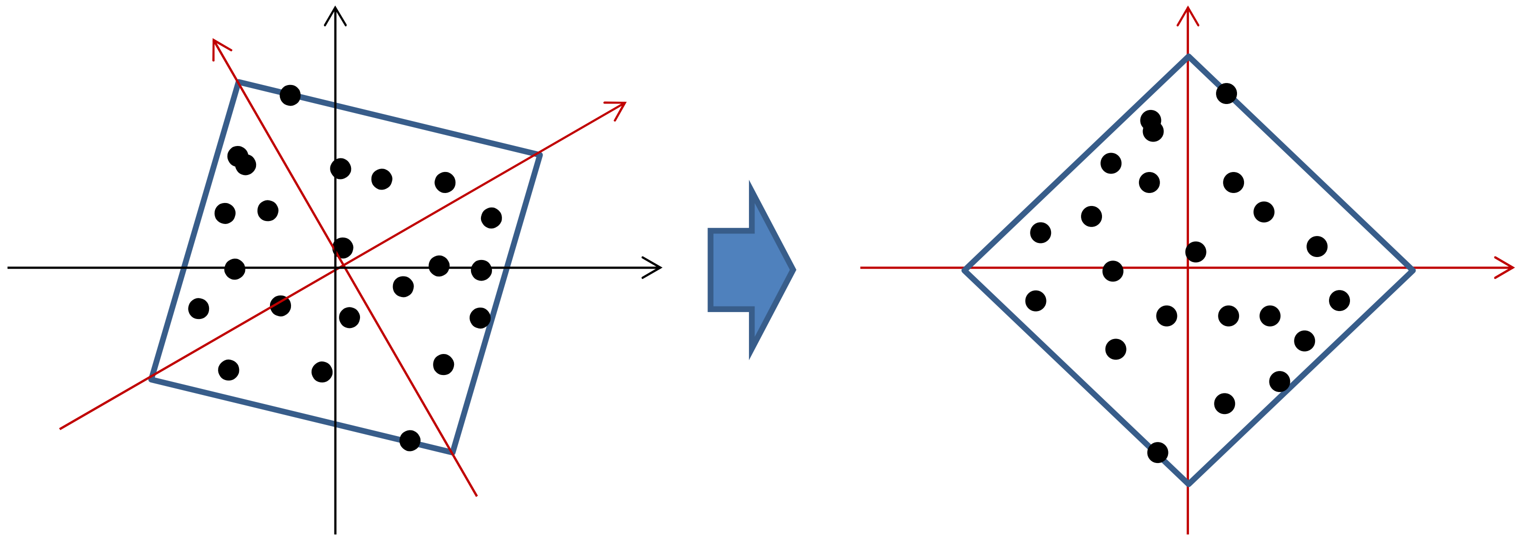
Recap: PCA

- Let $\mathbf{X}$ be the matrix of correlated data vectors
  - Each component of $\mathbf{X}$ informs us of the mean trend of other components

- Need a transform $\mathbf{T}$ such that if $\mathbf{Y} = \mathbf{TX}$, the covariance of $\mathbf{Y}$ is diagonal
  - $\mathbf{YY}^{\mathrm{T}}$ is diagonal

- **PCA:** $\mathbf{T}$ is the (transposed) matrix of Eigenvectors of the covariance matrix $\mathbf{XX}^{\mathrm{T}}$

Recap: Decorrelating by PCA



- PCA finds the principal axes of the scatter of the data
  - The Eigen vectors of the covariance matrix
- The PCA transformation transforms the principal axes of the data scatter to the main axes of the space
- This also has the *side effect* of decorrelating the data

PCA decorrelates data

- For centered (zero-mean) data $\mathbf{X}$
- The Eigenvectors of the covariance matrix are identical to the left singular vectors

$$\text{SVD: } \mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

- We can write $\mathbf{Y} = \mathbf{S}\mathbf{V}^{\mathbf{T}}$ and

$$\mathbf{X} = \mathbf{U}\mathbf{Y} \quad (\text{and } \mathbf{Y} = \mathbf{U}^T\mathbf{X})$$

  - i.e. we're setting the transform $\mathbf{T} = \mathbf{U}^T$ and $\mathbf{Y} = \mathbf{T}\mathbf{X}$
- $\mathbf{Y}$ is the representation of $\mathbf{X}$ in terms of the columns of $\mathbf{U}$
- But

$$\mathbf{Y}\mathbf{Y}^T = (\mathbf{S}\mathbf{V}^{\mathbf{T}}\mathbf{V}\mathbf{S}^{\mathbf{T}}) = \mathbf{S}\mathbf{S}^{\mathbf{T}} = \text{Diagonal}$$

- I.e. the new representations $\mathbf{Y}$ are uncorrelated

Recap: The statistical concept of *Independence*

- Two variables X and Y are *dependent* if If knowing X gives you *any information about* Y

- X and Y are *independent* if knowing X tells you nothing at all of Y

Recap: Independence

- ***Independence:*** Two random variables X and Y are independent iff:
  - Their joint probability equals the product of their individual probabilities
- $P(X,Y) = P(X)P(Y)$
- Independence implies uncorrelatedness
  - The average value of X is the same regardless of the value of Y
    - $E[X|Y] = E[X]$
  - But uncorrelatedness does not imply independence

Recap: Independence

- *Independence:* Two random variables X and Y are independent iff:
- The average value of ***any function*** *of* X is the same regardless of the value of Y
  - Or any function of Y
- **E[f(X)g(Y)] = E[f(X)] E[g(Y)] for all f(), g()**

Poll 1

- The objective of PCA is to decorrelate the data
  - True
  - False

- If two random values x and y are independent, then which of the following is true of $E[x^2y^2]$?
  - $E[x^2y^2] = E[x]^2E[y]^2$
  - $E[x^2y^2] = E[x^2]E[y^2]$

Poll 1

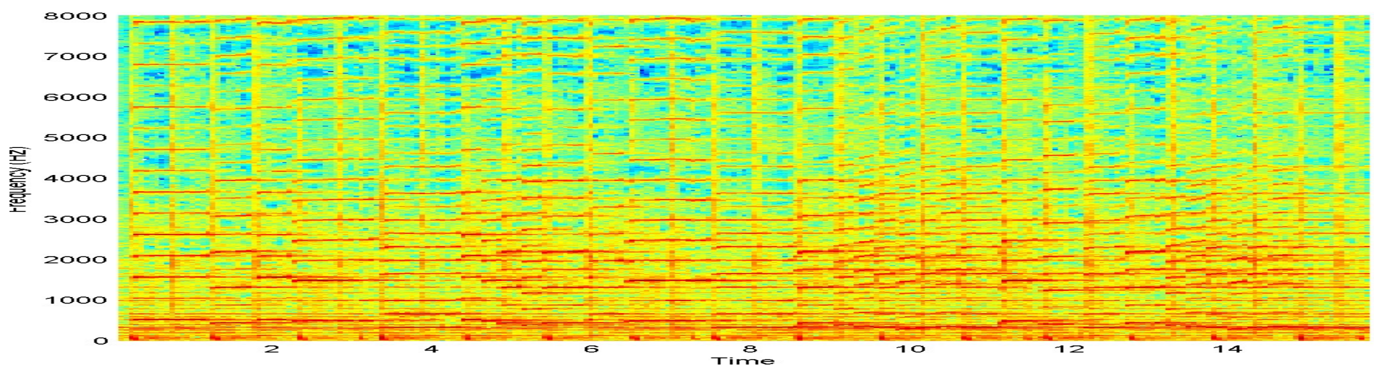
- The objective of PCA is to decorrelate the data
  - True
  - **False**

- If two random values x and y are independent, then which of the following is true of $E[x^2y^2]$?
  - $E[x^2y^2] = E[x]^2E[y]^2$
  - **$E[x^2y^2] = E[x^2]E[y^2]$**

Moving on: Finding bases…

Recap: Finding bases, aka building blocks..

M =



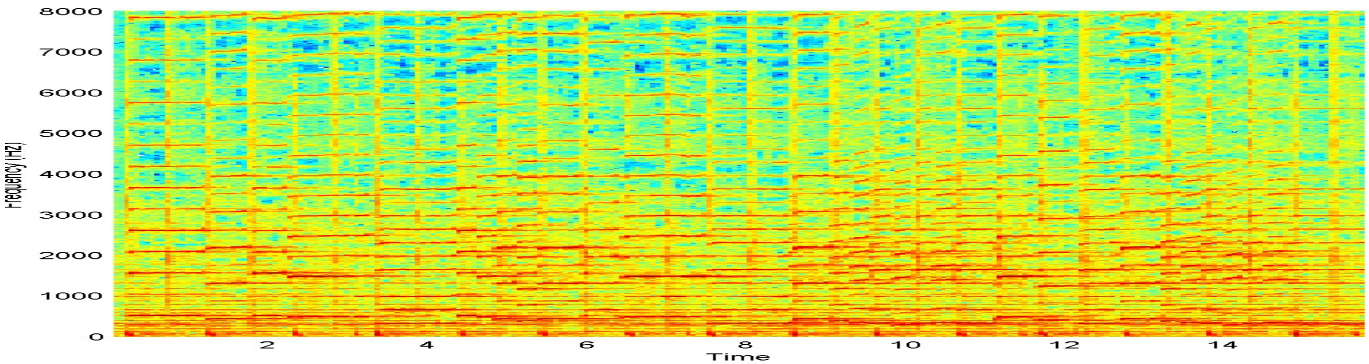
H = ?

W = ?

U = ?

- Find the bases **W** that best explain the data *in a meaningful way*

Recap: Finding bases, aka building blocks..

M = 

H = **?**

W = **?**

U = **?**

- Meaningful – try1: The bases are ***orthogonal***

A least squares solution

$$\mathbf{W}, \mathbf{H} = \arg\min_{\overline{\mathbf{W}}, \overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{W}}\overline{\mathbf{H}} \|_F^2 + \Lambda(\overline{\mathbf{W}}^T \overline{\mathbf{W}} - \mathbf{I})$$

- Constraint: $\mathbf{W}$ is orthogonal
  - $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$
- The solution:
  - $\mathbf{W}$ are the Eigen vectors of $\mathbf{M}\mathbf{M}^{\mathbf{T}}$
  - PCA!!

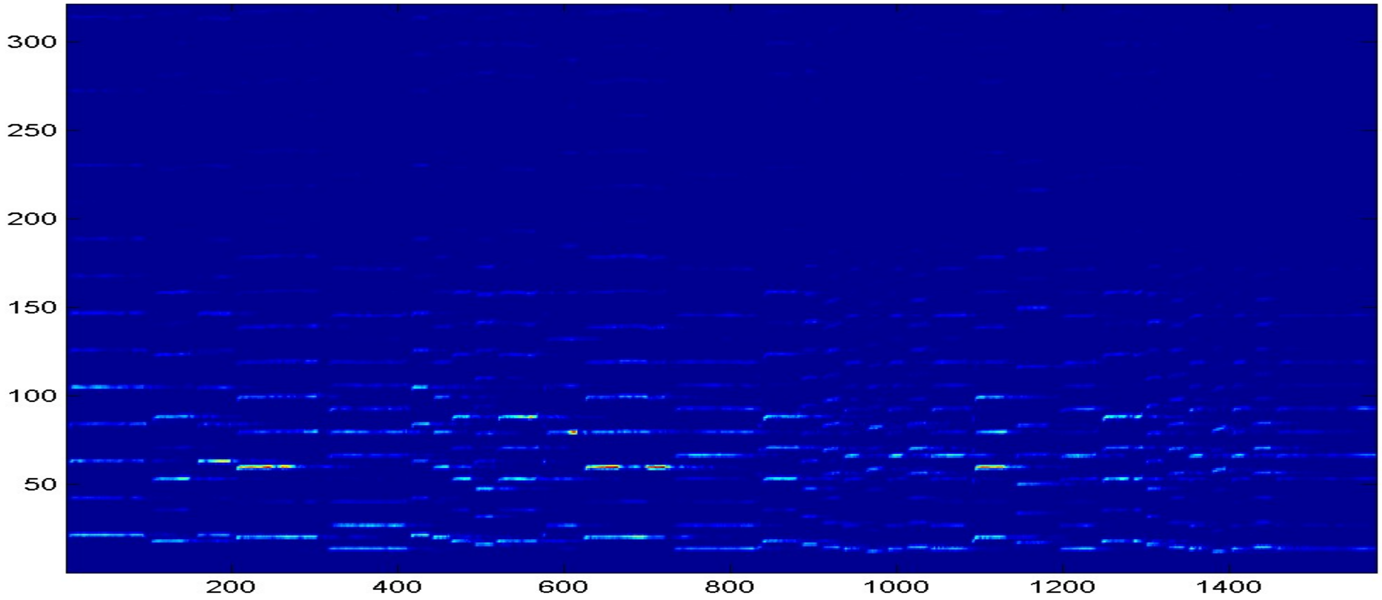
- $\mathbf{M} \sim \mathbf{W}\mathbf{H}$ is an approximation
- Also, the rows of $\mathbf{H}$ are *decorrelated*

PCA

$$\mathbf{M} = \mathbf{WH}$$

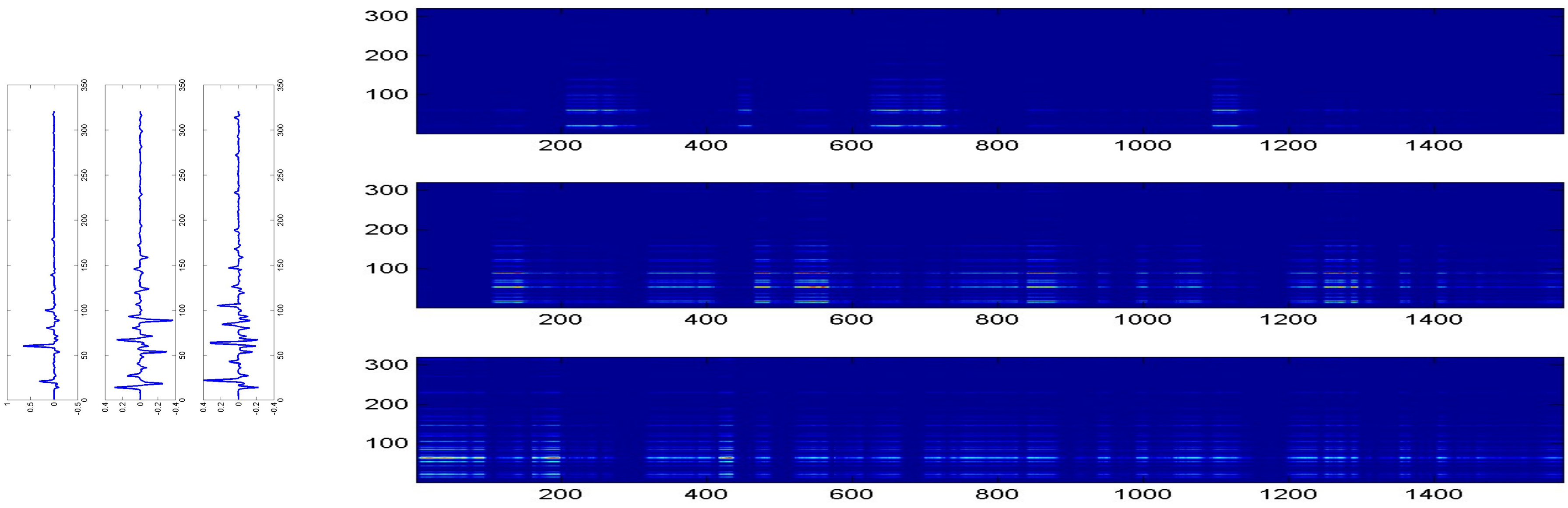
- The orthogonal columns of $\mathbf{W}$ are the bases we have learned
  - The linear “building blocks” that compose the music
- They represent “learned” notes
  - $\mathbf{w}_i \mathbf{h}_i$ is the contribution of the ith note to the music
    - $\mathbf{w}_i$ is the ith column of $\mathbf{W}$
    - $\mathbf{h}_i$ is the ith row of $\mathbf{H}$

So how does that work?



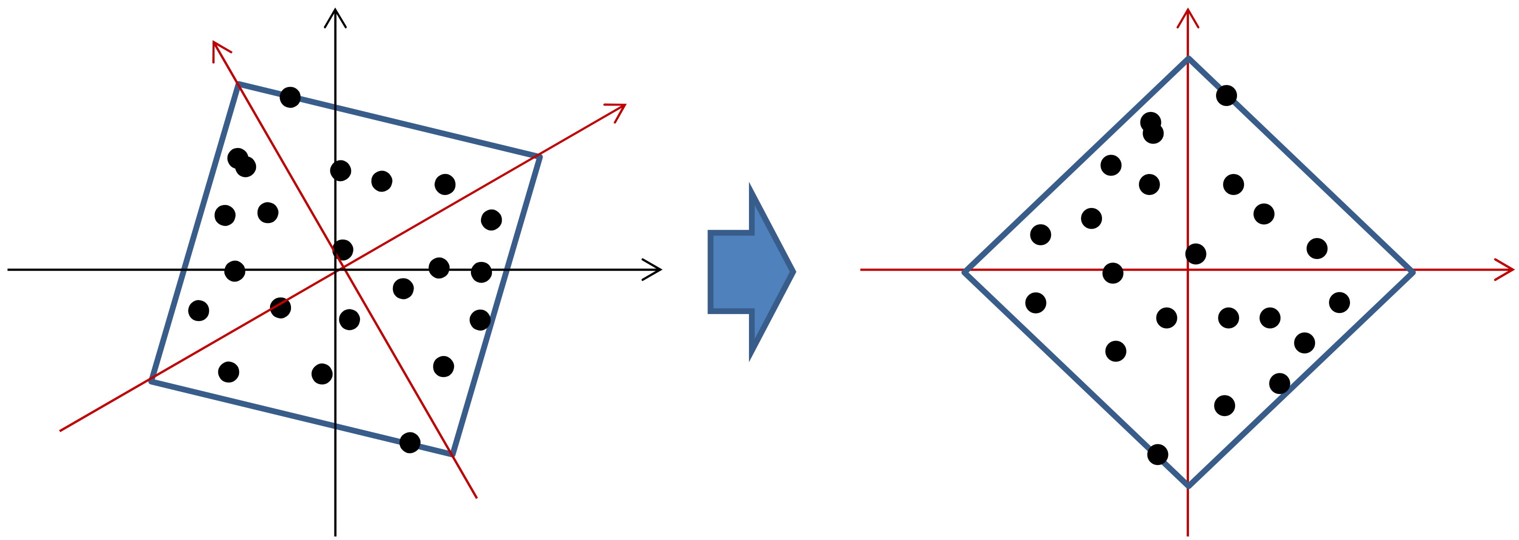
- There are 12 notes in the segment, hence we try to estimate 12 notes..

So how does that work?



- There are 12 notes in the segment, hence we try to estimate 12 notes..
- Results are not good

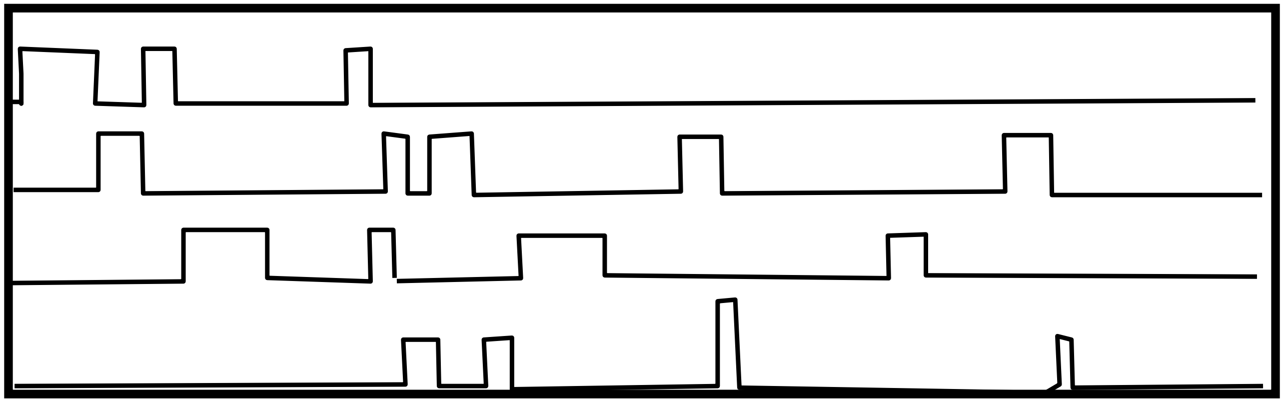
Recap: Decorrelating by PCA



- PCA decorrelates the data *incidentally*
- The focus is on the orthogonality of the axes, decorrelated representations is a side effect
- What if we focus, instead, on *decorrelating* the data directly?

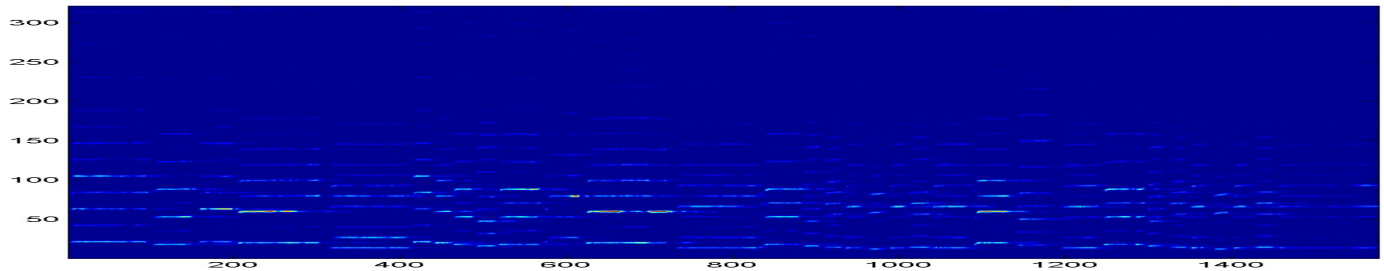
PCA through decorrelation of notes

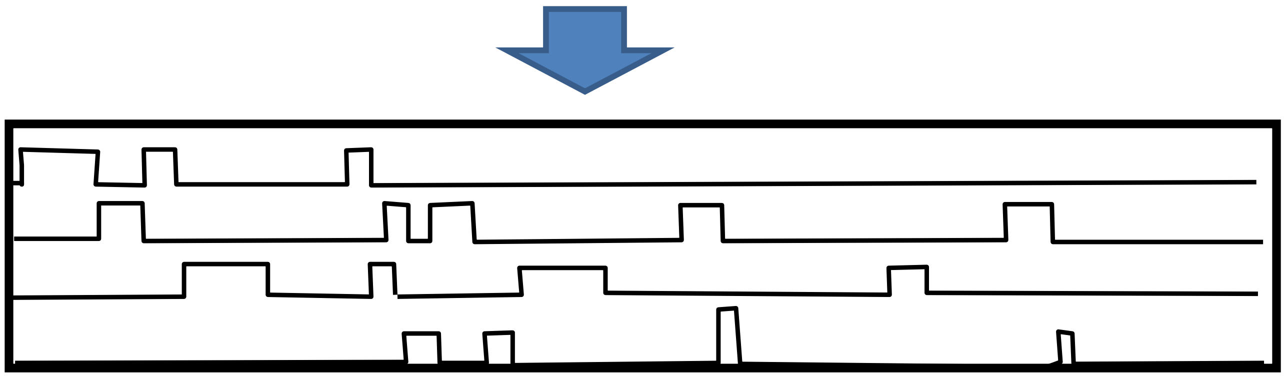
$$\mathbf{W},\mathbf{H} = \arg\min_{\overline{\mathbf{W}},\overline{\mathbf{H}}} \|\mathbf{M} - \overline{\mathbf{H}}\|_F^2 + \Lambda(\overline{\mathbf{H}}\overline{\mathbf{H}}^T - \mathbf{D})$$



- Different constraint: Constraint $\mathbf{H}$ to be decorrelated
  - $\mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{D}$

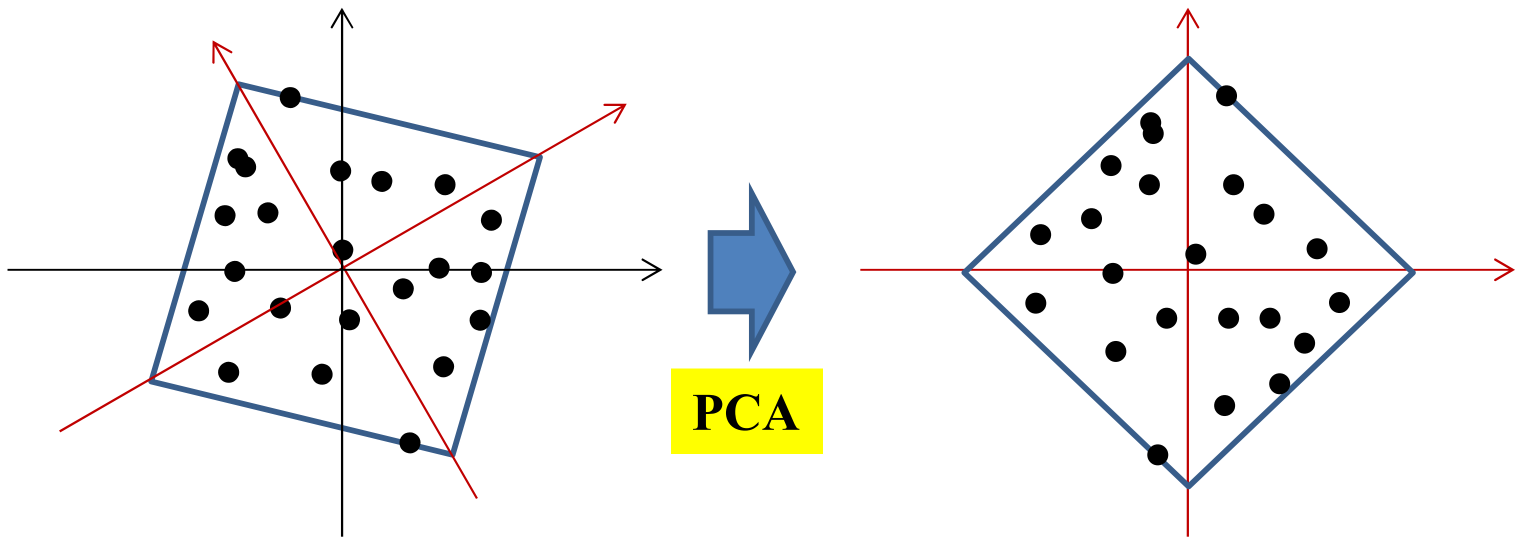
Decorrelation





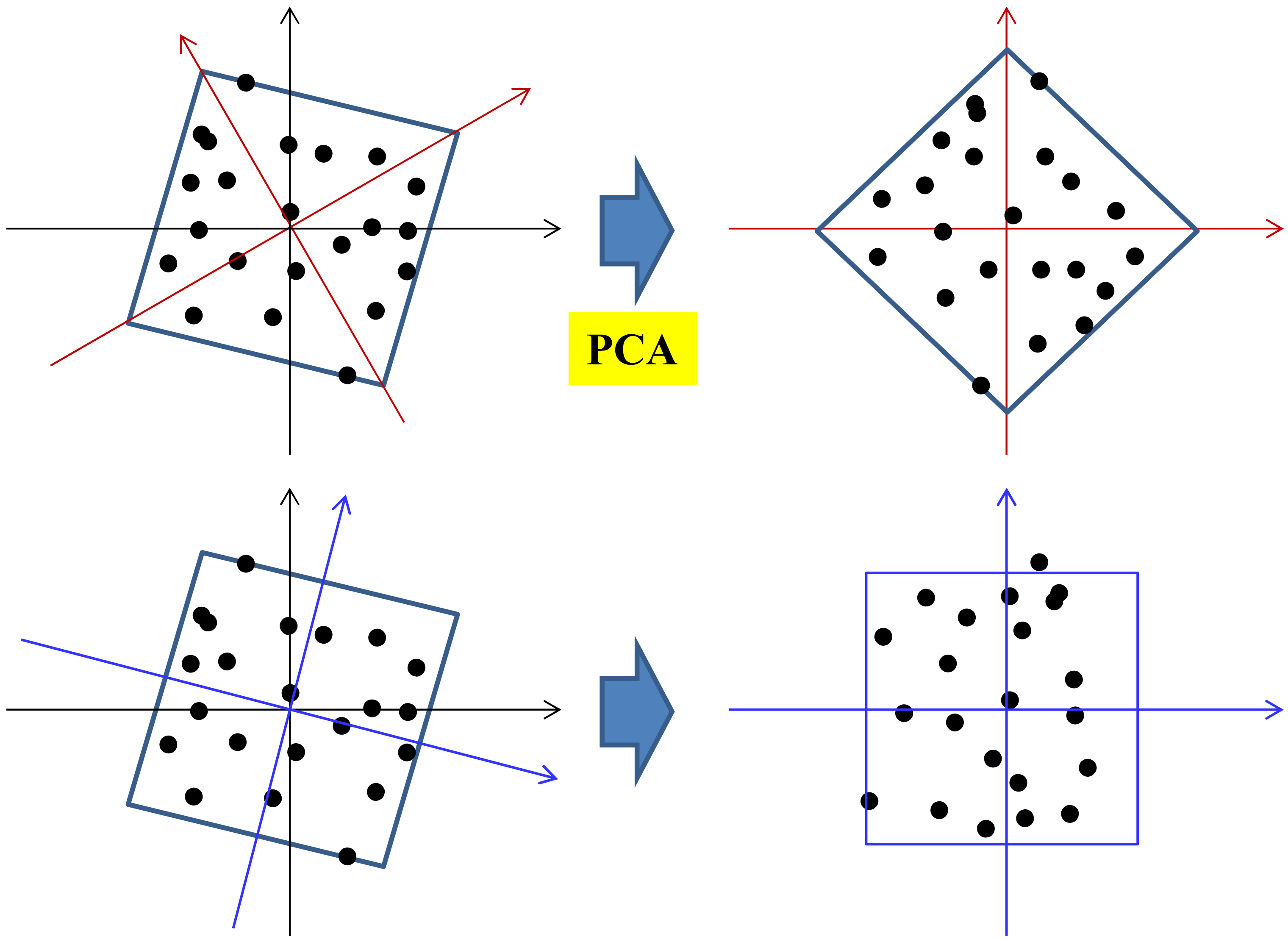
- Alternate view: Find a matrix $\mathbf{B}$ such that the rows of $\mathbf{H}=\mathbf{B}\mathbf{M}$ are uncorrelated
- PCA is one solution already
- **Are there others?**

Decorrelating the data



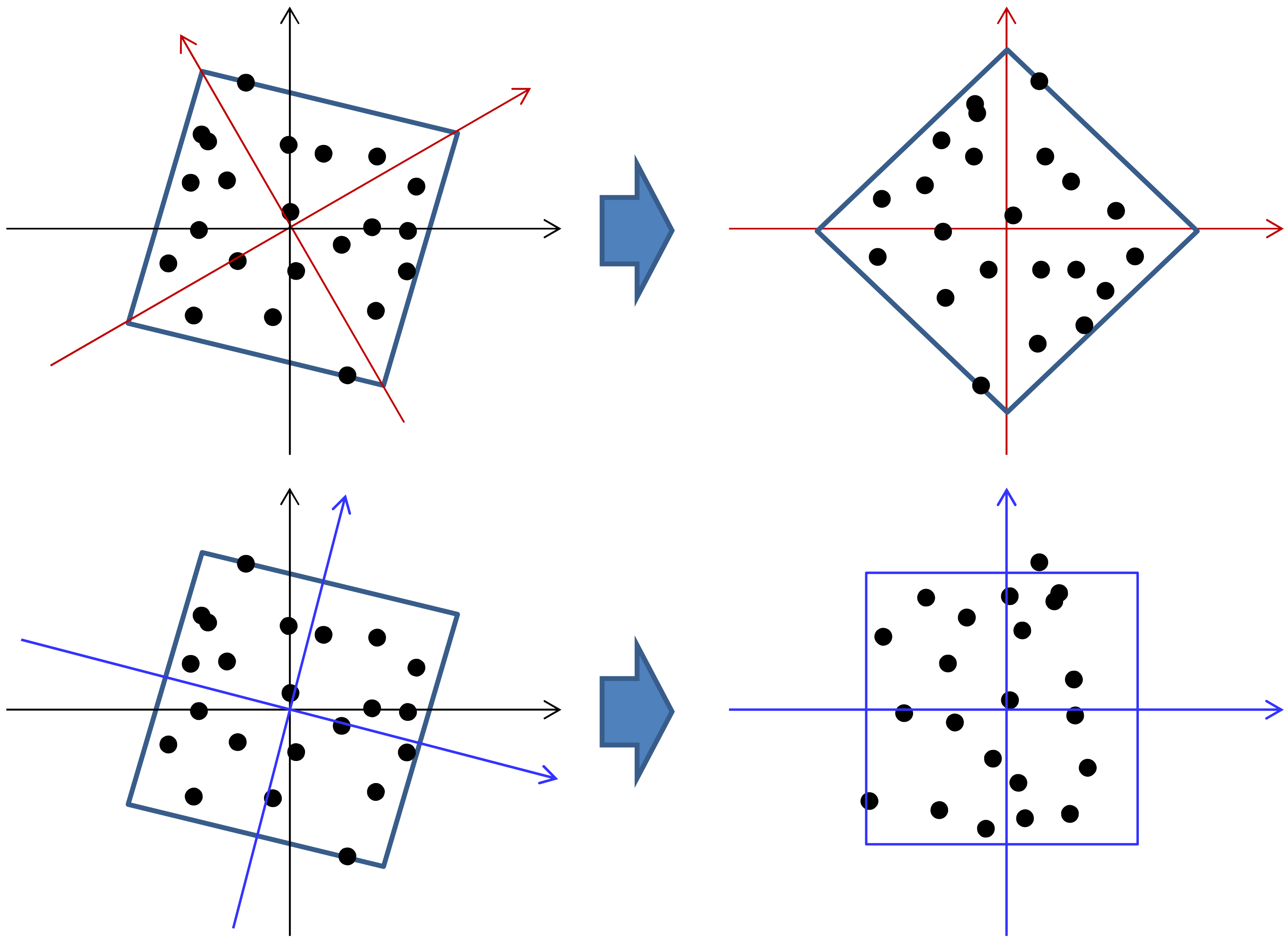
- Are there other decorrelating axes?

Decorrelating the data



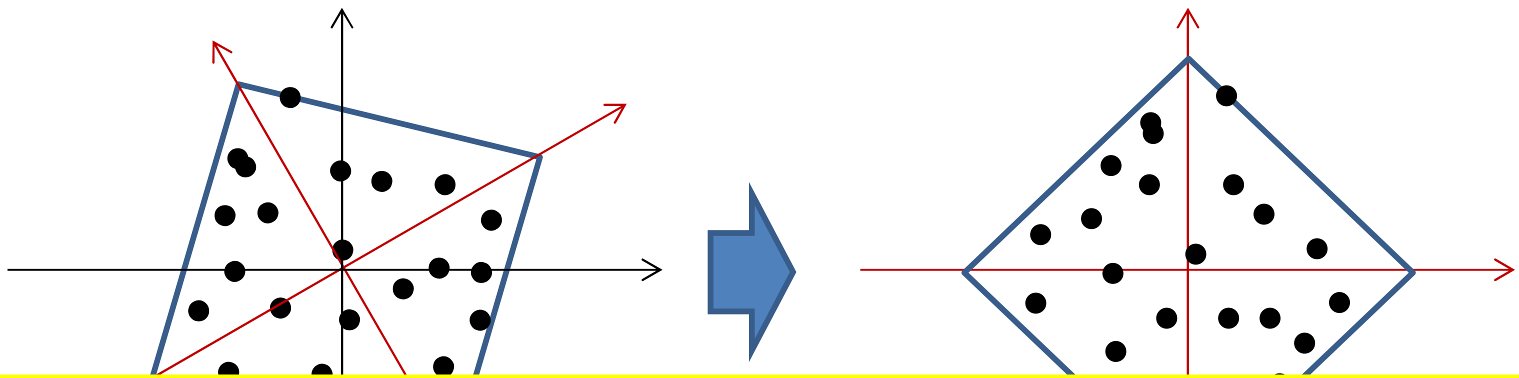
- But PCA will find only one of them, why?

Decorrelating the data

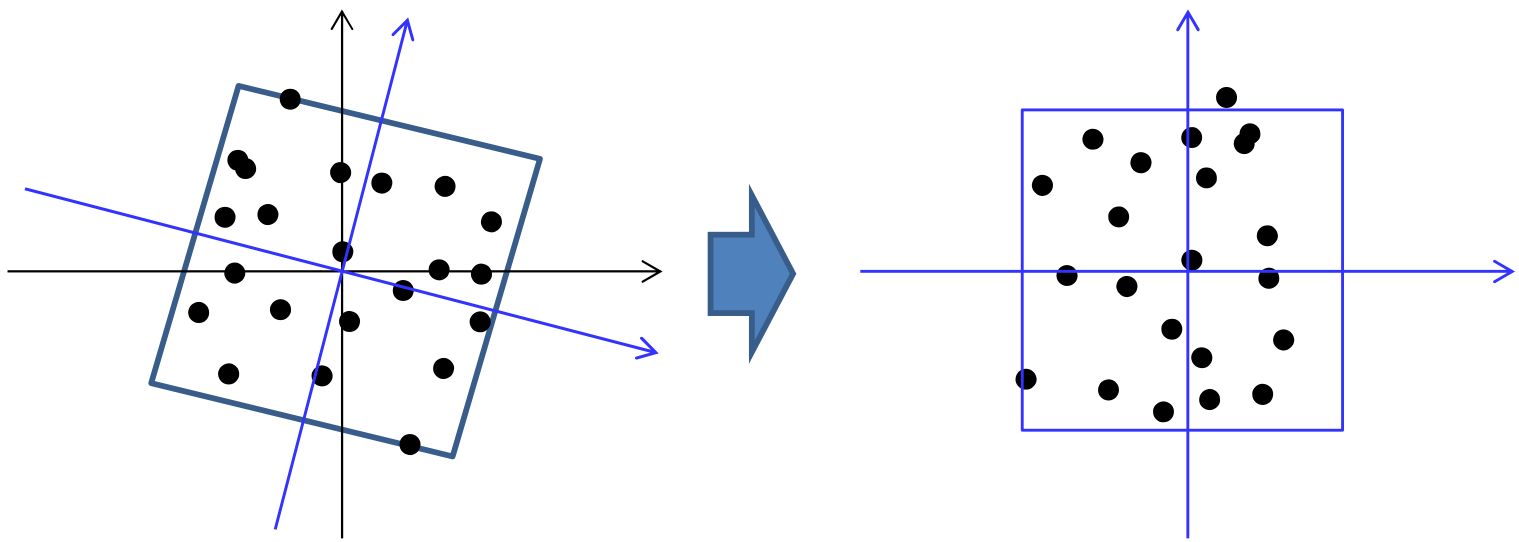


A decorrelation-based decomposition can find either of them.
The solution is non-unique

Decorrelating the data

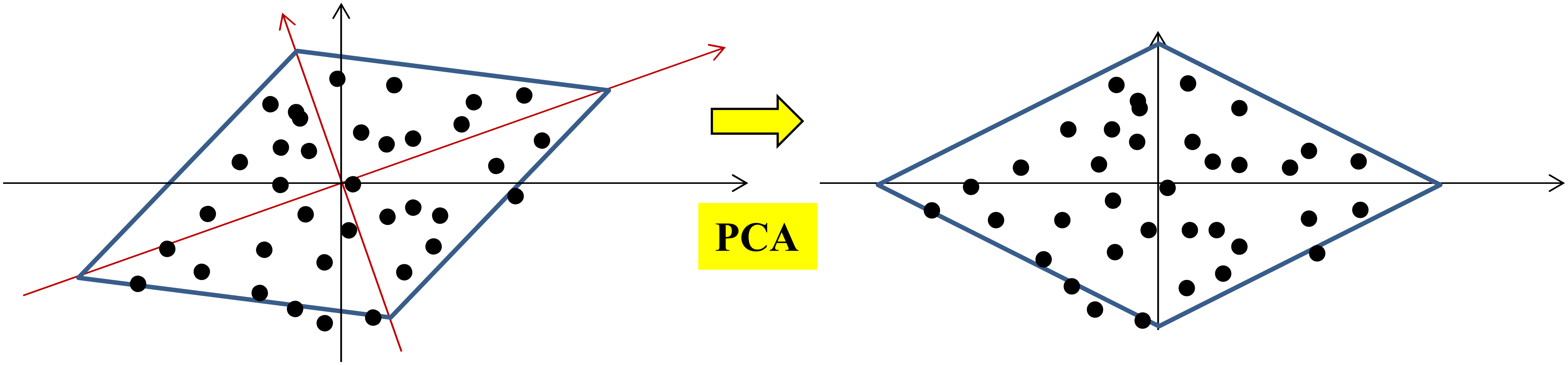


What is special about the blue axes,
and how can we modify our decomposition to find them instead



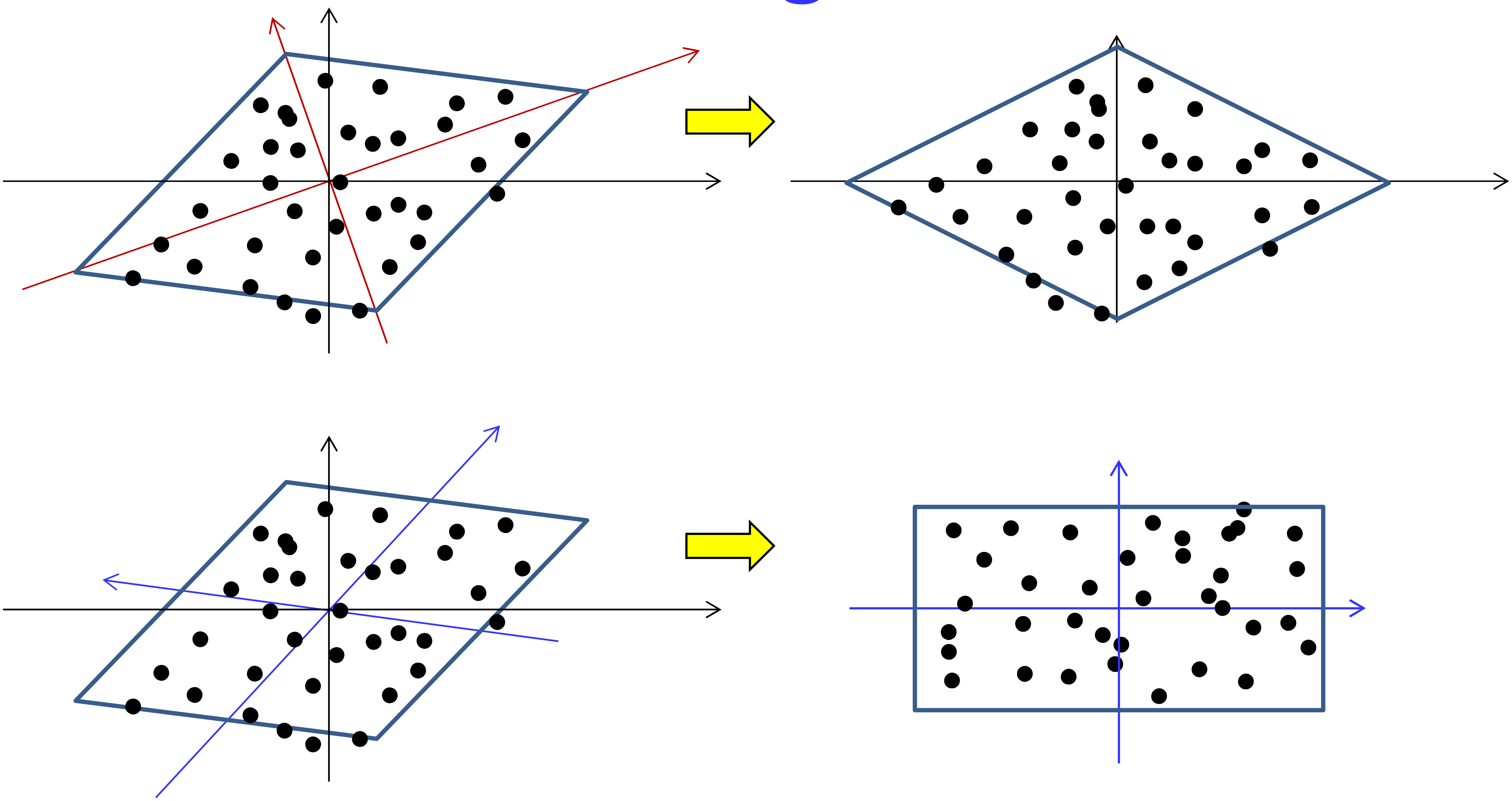
A decorrelation-based decomposition can find either of them.
The solution is non-unique

Decorrelating the data



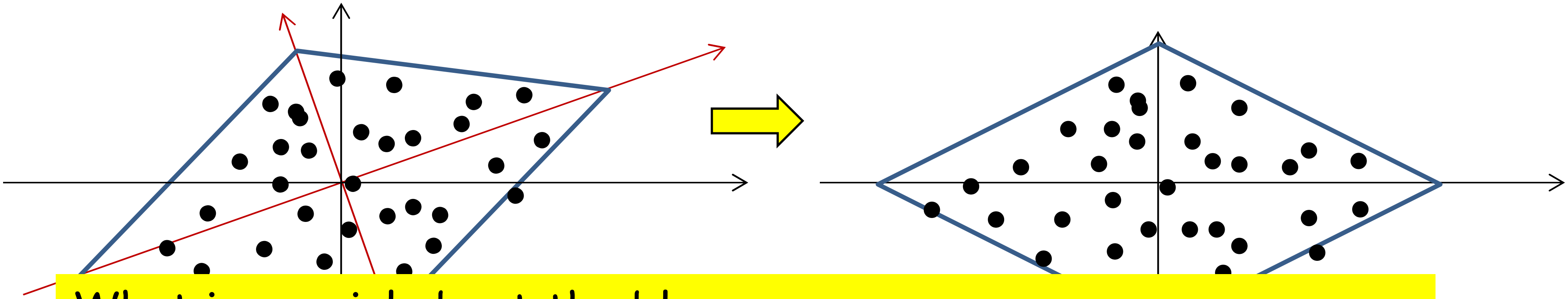
- Are there other decorrelating axes?

Decorrelating the data

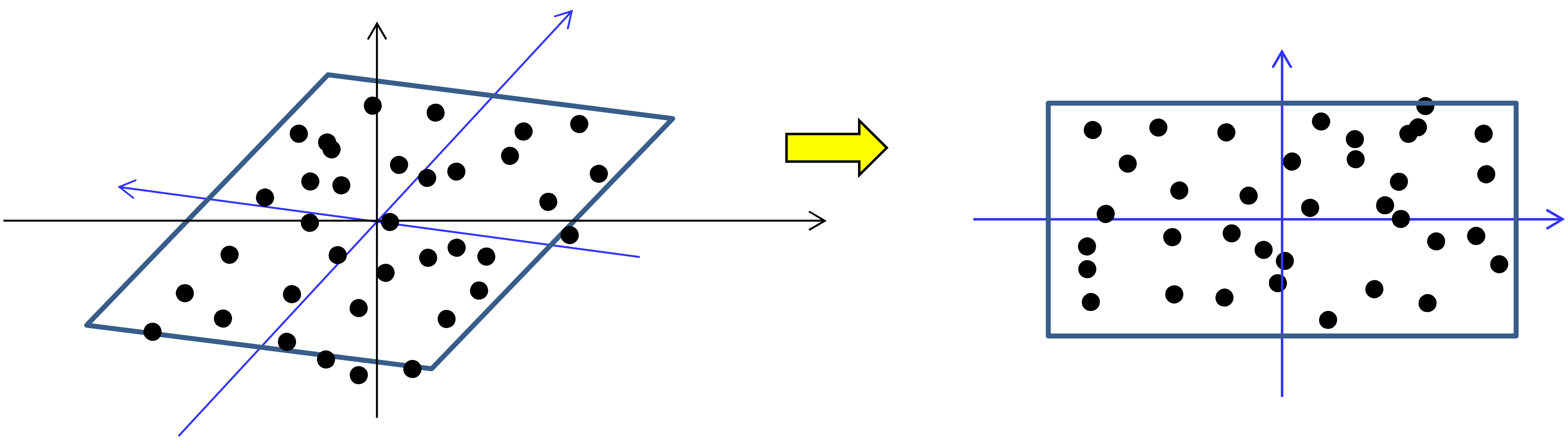


- The decorrelation-based decomposition has multiple solutions, but PCA will find only one of them

Decorrelating the data

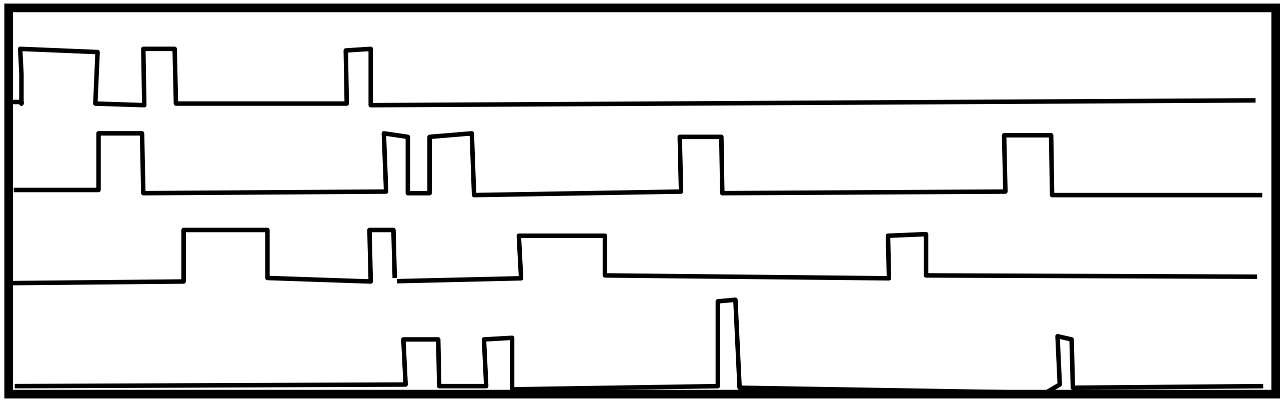


What is special about the blue axes,
and how can we modify our decomposition to find them instead



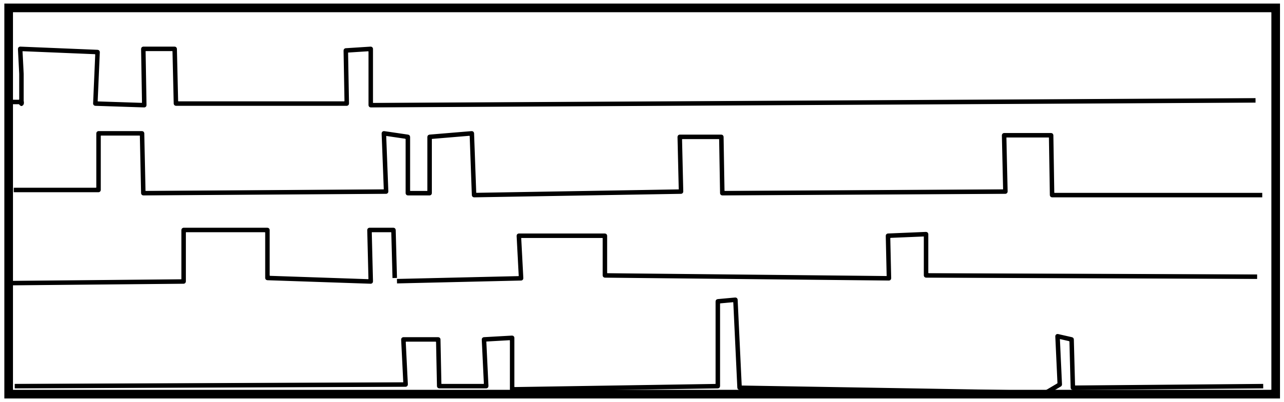
- The decorrelation-based decomposition has multiple solutions, but PCA will find only one of them

What *else* can we look for?



- Assume: The "transcription" of one note does not depend on what else is playing
  - Or, in a multi-instrument piece, instruments are playing independently of one another
- Not strictly true, but still..

What *else* can we look for?



- Assume: The “transcription” of one note does not depend on what else is playing
  - Or, in a multi-instrument piece, instruments are playing independently of one another

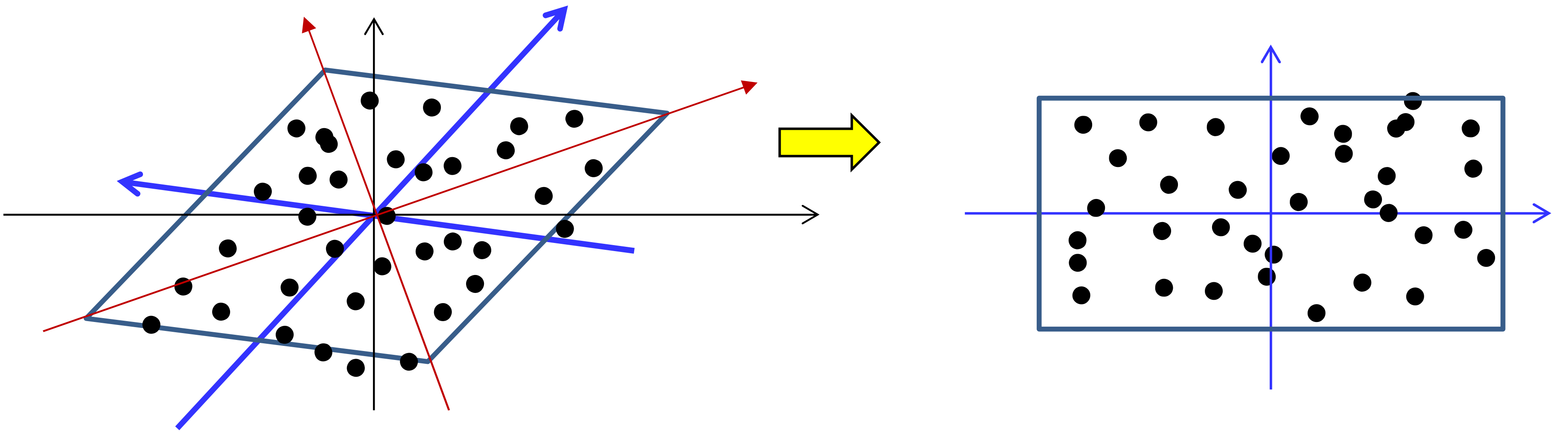
- **Attempting to find statistically independent components of the mixed signal**
  - ***Independent Component Analysis***

Formulating it with Independence

$$\mathbf{W}, \mathbf{H} = \arg\min_{\overline{\mathbf{W}}, \overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{W}}\overline{\mathbf{H}} \|_F^2 + \Lambda(\textit{rows of } \mathbf{H} \textit{ are independent})$$

- Impose statistical independence constraints on decomposition

Independent Component Analysis



- **Independent Component Analysis** searches through all possible combinations of bases to find the set that makes the representations in terms of these bases maximally independent

Poll 2

- If there are multiple decorrelating axes, the solution to PCA will always be indeterminate
  - True
  - False

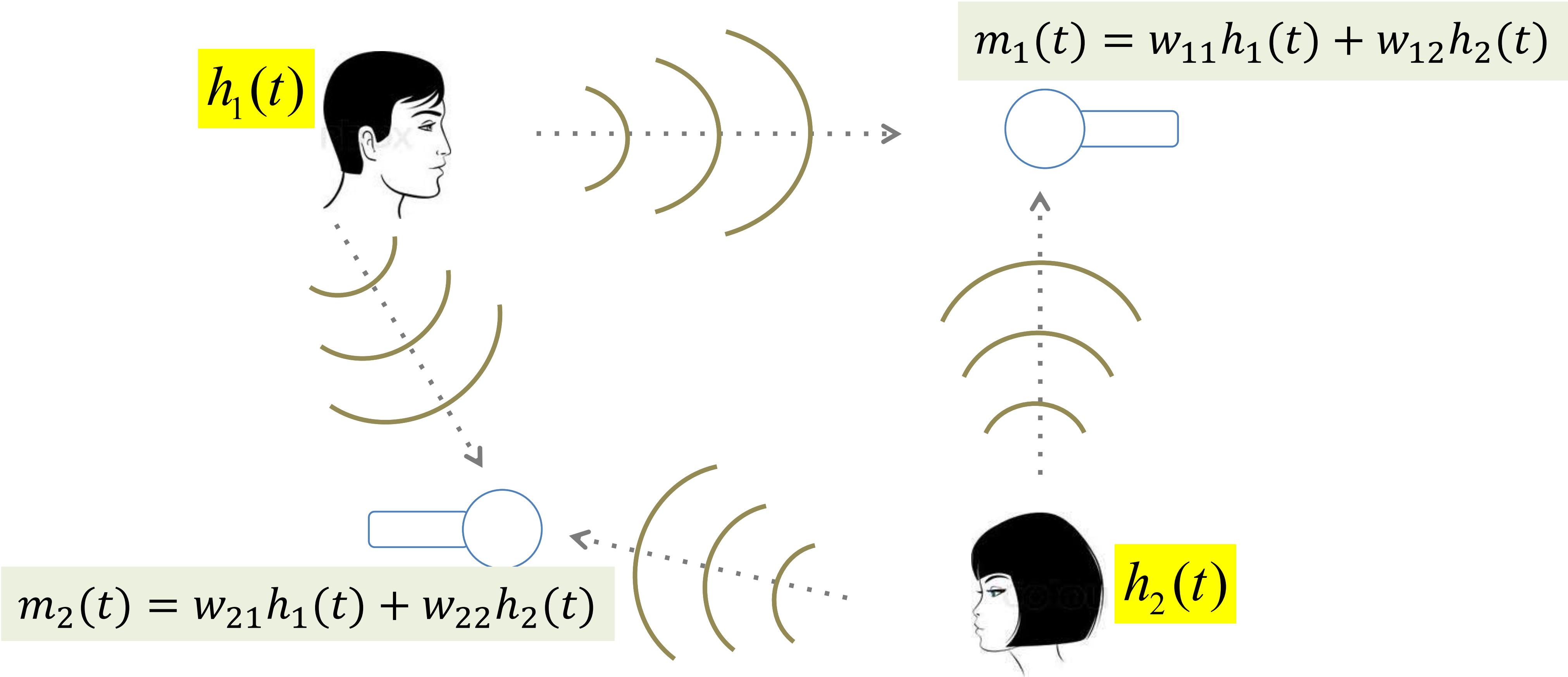
- Independent Component Analysis attempts to decompose a data matrix into the product of a bases matrix and a weights matrix, such that the components of the weights vectors are statistically independent
  - True
  - False

Poll 2

- If there are multiple decorrelating axes, the solution to PCA will always be indeterminate
  - True
  - **False**

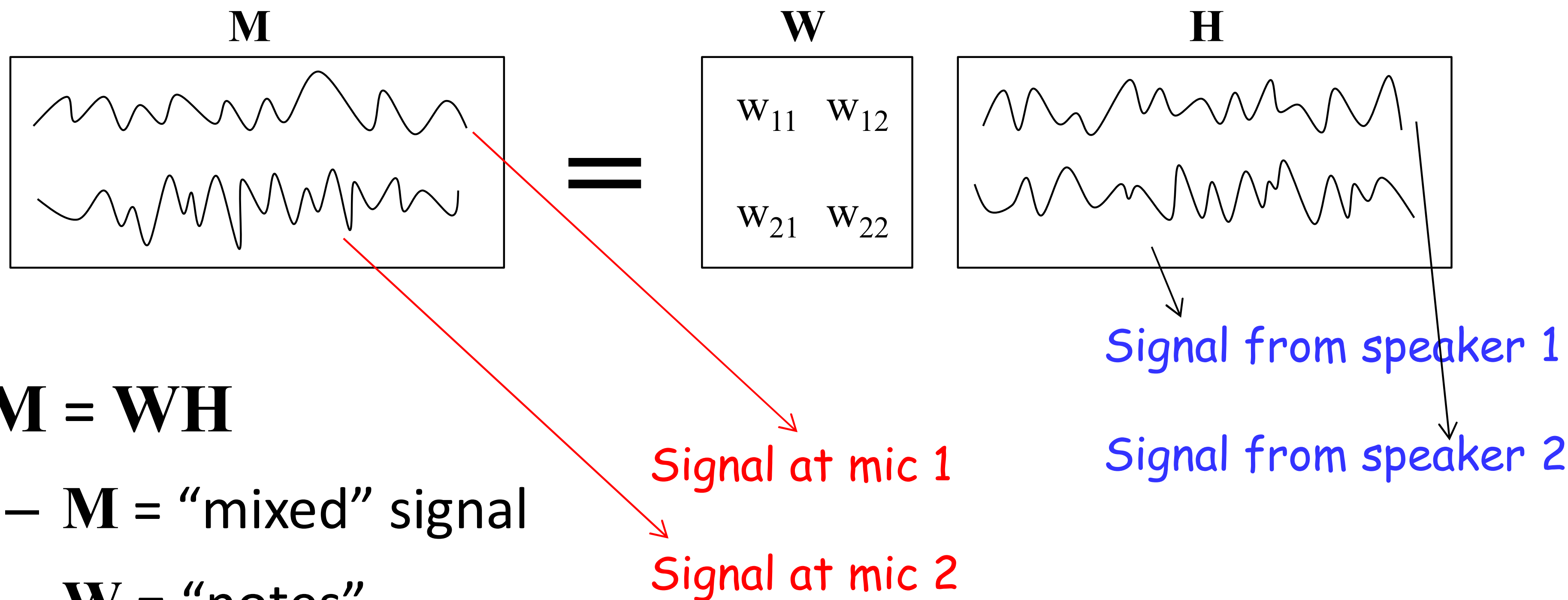
- Independent Component Analysis attempts to decompose a data matrix into the product of a bases matrix and a weights matrix, such that the components of the weights vectors are statistically independent
  - **True**
  - False

Changing problems for a bit



- Two people speak simultaneously
- Recorded by two microphones
- Each recorded signal is a mixture of both signals

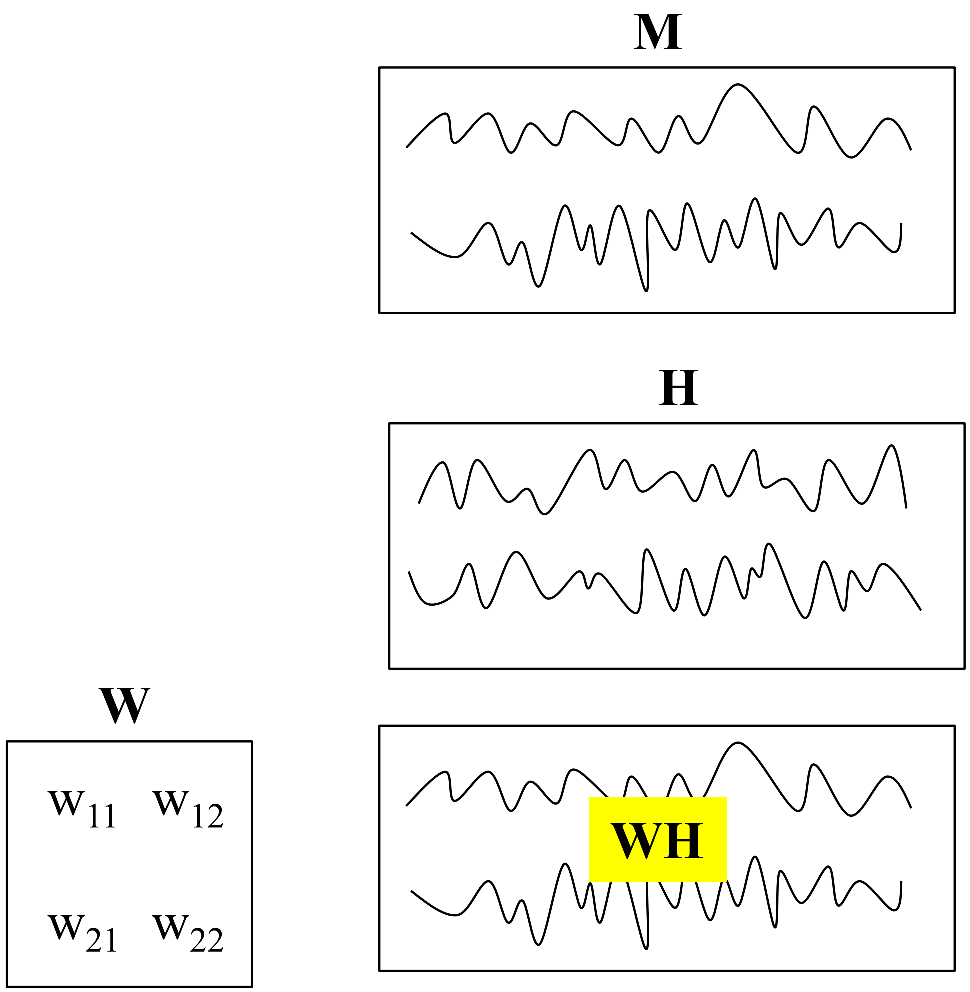
A Separation Problem



- **M** = **WH**
  - **M** = "mixed" signal
  - **W** = "notes"
  - **H** = "transcription"

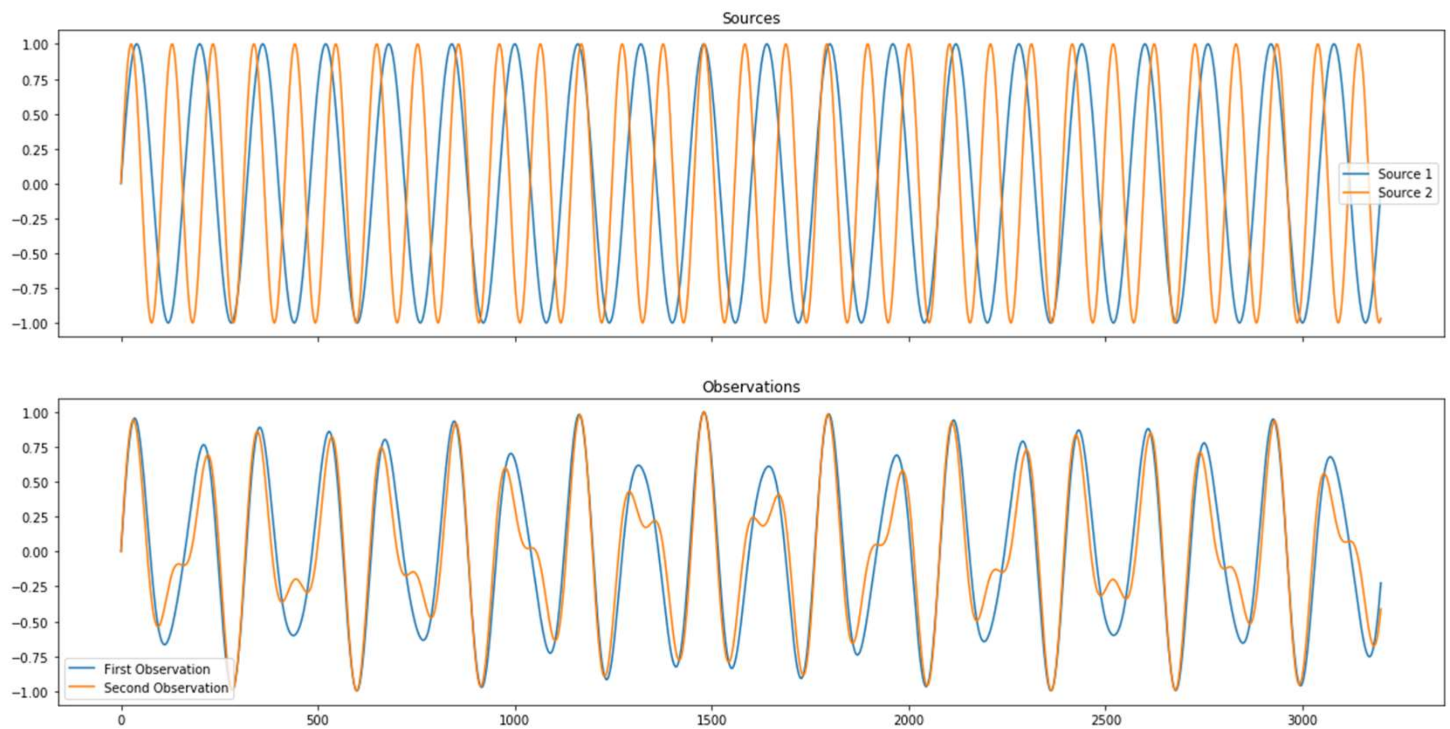
- Separation challenge: Given only **M** estimate **H**
- Identical to the problem of "finding scores (and notes)"

A Separation Problem



- Separation challenge: Given only $\mathbf{M}$ estimate $\mathbf{H}$
- **Identical to the problem of "finding scores"**

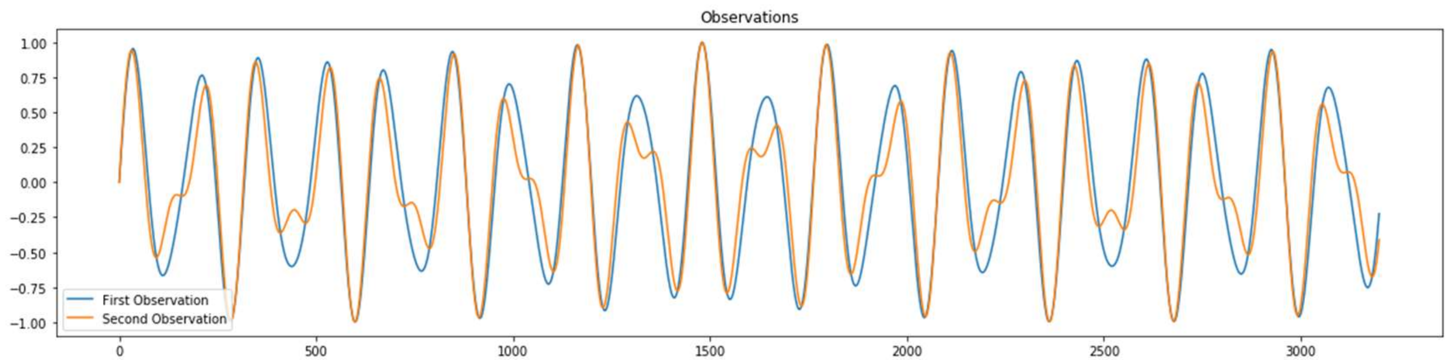
Example: Sources & Mixing



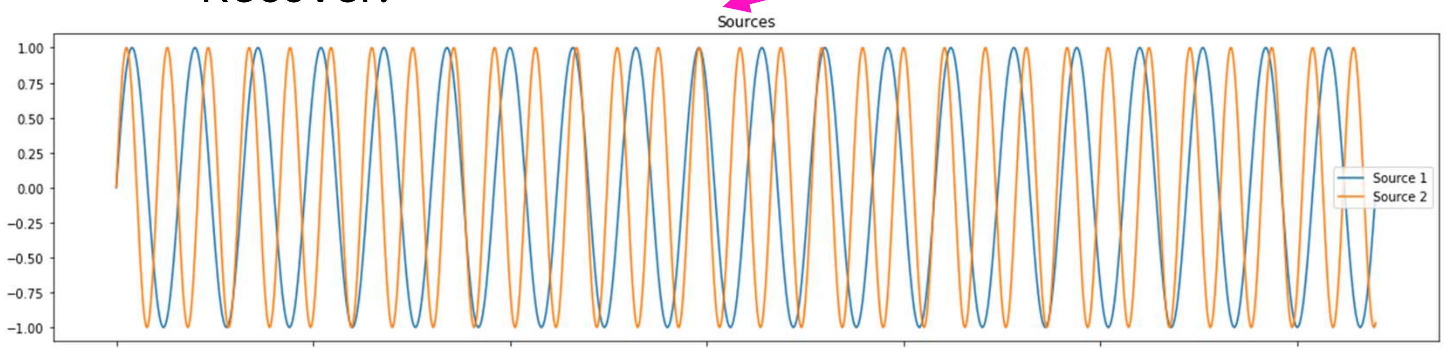
$$W = \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix}$$

Problem Statement

Given:

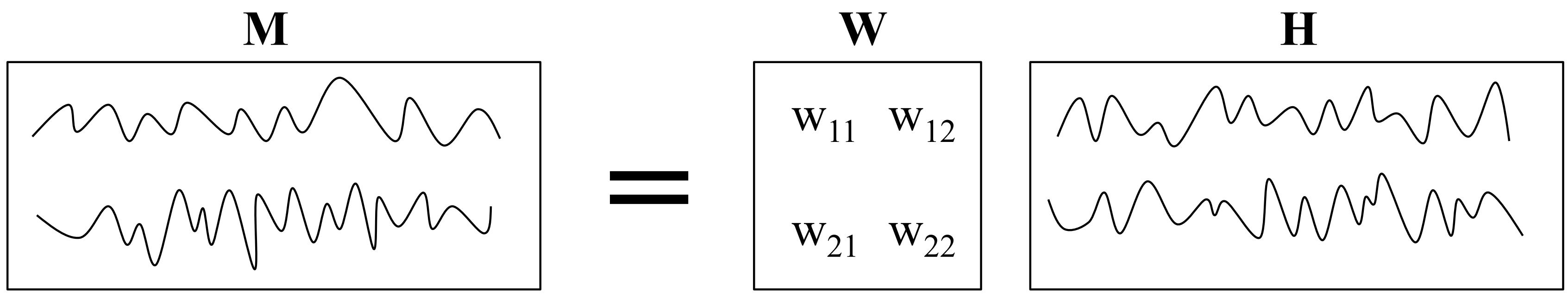


Recover:



$$W = \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix}$$

Imposing Statistical Constraints



- $\mathbf{M} = \mathbf{WH}$
- Given only $\mathbf{M}$ estimate $\mathbf{H}$
- $\mathbf{H} = \mathbf{W}^{-1}\mathbf{M} = \mathbf{AM}$
- Only known constraint: The rows of $\mathbf{H}$ are independent
- Estimate $\mathbf{A}$ such that the components of $\mathbf{AM}$ are statistically independent
  - $\mathbf{A}$ is the *unmixing* matrix

Statistical Independence

- $\mathbf{M = WH}$ $\quad \boxed{\mathbf{H = AM}}$

Remember this form

In order to recover the original unmixed signals H from the mixed signal M

An ugly algebraic solution

$$\mathbf{M} = \mathbf{W}\mathbf{H} \xrightarrow{\textbf{decorrelate}} \mathbf{H} = \mathbf{A}\mathbf{M}$$

- **Solution 1**: "Recover" $\mathbf{H}$ by decorrelating $\mathbf{M}$
  - We know uncorrelated signals have diagonal correlation matrix
- Find a transform $\mathbf{A}$ such that the rows of $\mathbf{H}=\mathbf{A}\mathbf{M}$ are decorrelated
  - i.e. $\mathbf{H}\mathbf{H}^T$ = Diagonal (assuming 0 mean signals)
  - $\mathbf{A}$ was obtained by eigen decomposition of the correlation matrix of $\mathbf{M}$
    - I.e. by Eigen decomposition of $\mathbf{M}\mathbf{M}^T$

- We know this does not work, however
- Can we do the same for independence
  - Is there a linear transform that will enforce independence?

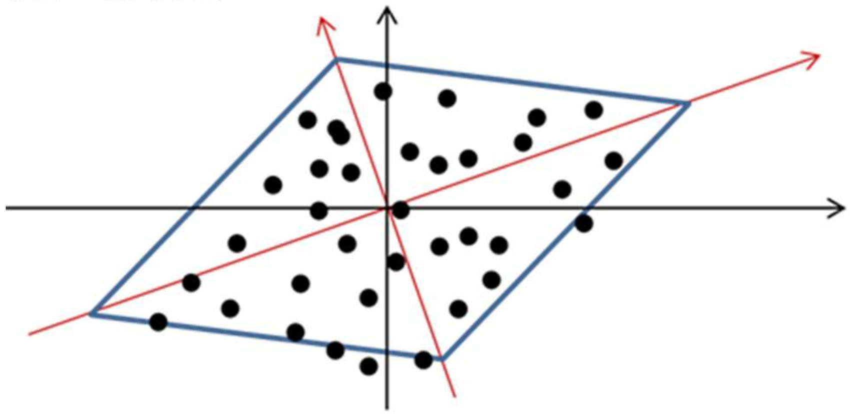
An ugly algebraic solution

- We *decorrelated* signals by diagonalizing the covariance matrix through Eigen decomposition

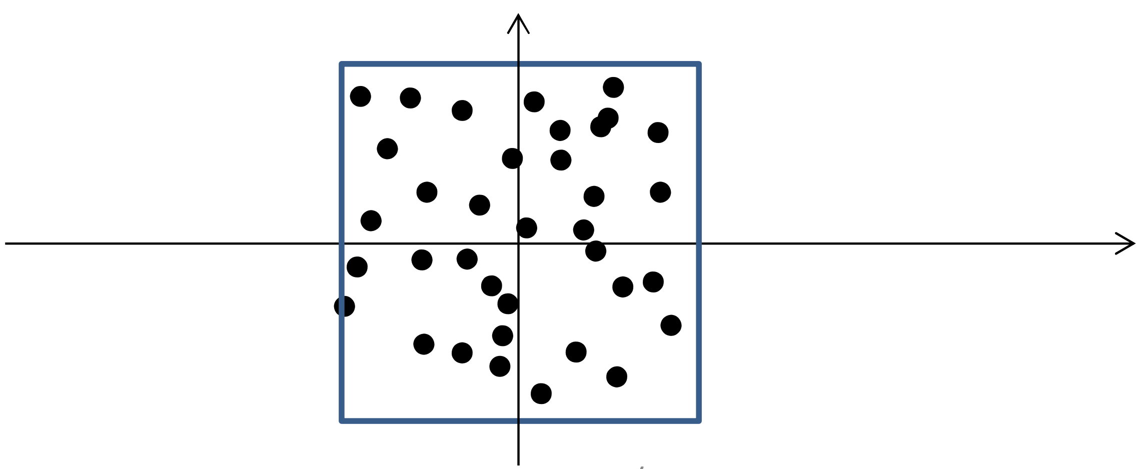
- *Is there a simple matrix we could just similarly diagonalize to make them independent?*
  - Some matrix whose Eigenvector matrix gives us the transform $\mathbf{A}$ such that the rows of $\mathbf{AM}$ are independent

Actual question

- Is there a linear transform that can transform a scatter like this

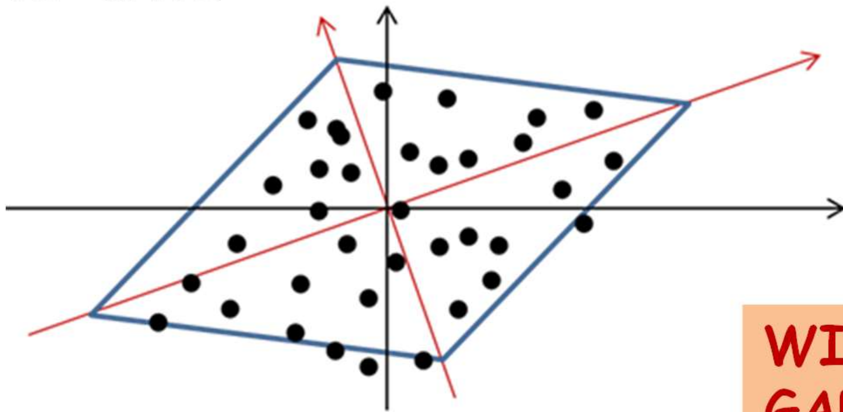


- To something like this:

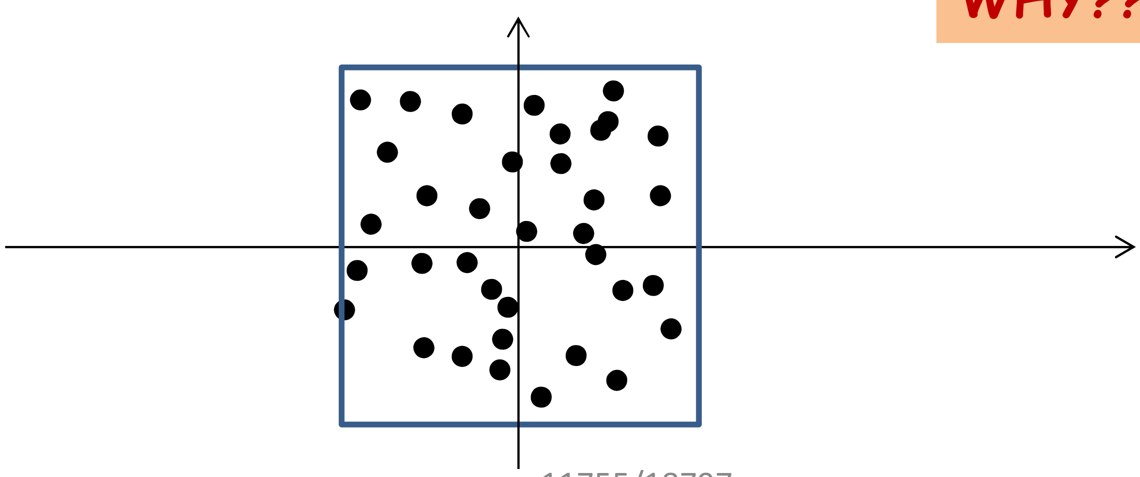


Actual question

- Is there a linear transform that can transform a scatter like this



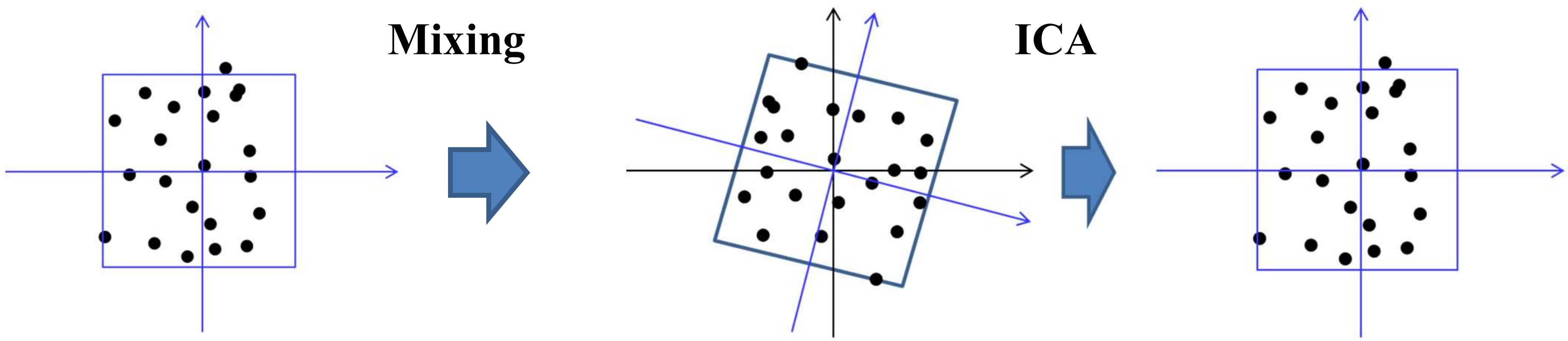
- To something like this:



WILL NOT WORK FOR GAUSSIAN DATA

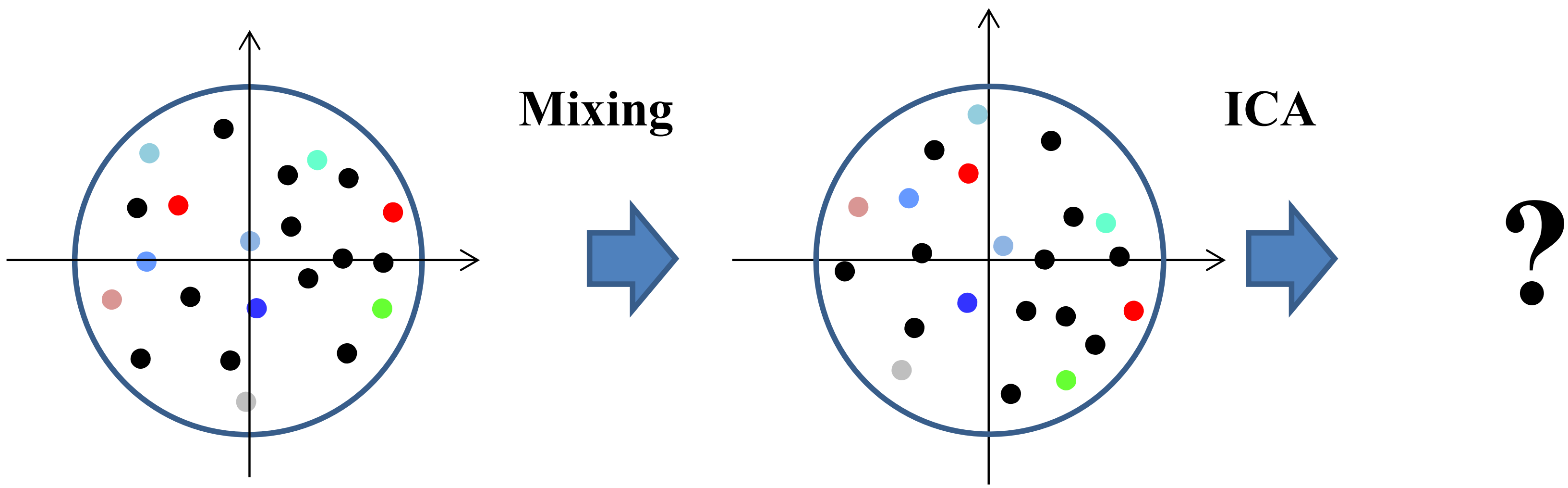
WHY??

Will not work for Gaussian data



- Concept behind ICA:
  - Original sources had some independent distribution
    - Assume all had identical variance
  - "Mixing" rotated the joint distribution
  - ICA finds the axes that "unmixes" the distribution
    - In principle, searches through all rotations such that the distribution is axis parallel again
      - This should give us back the original independent distribution

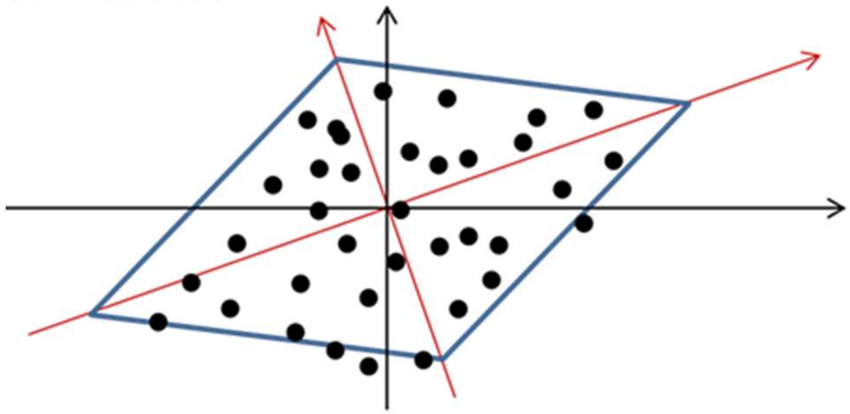
Will not work for Gaussian data



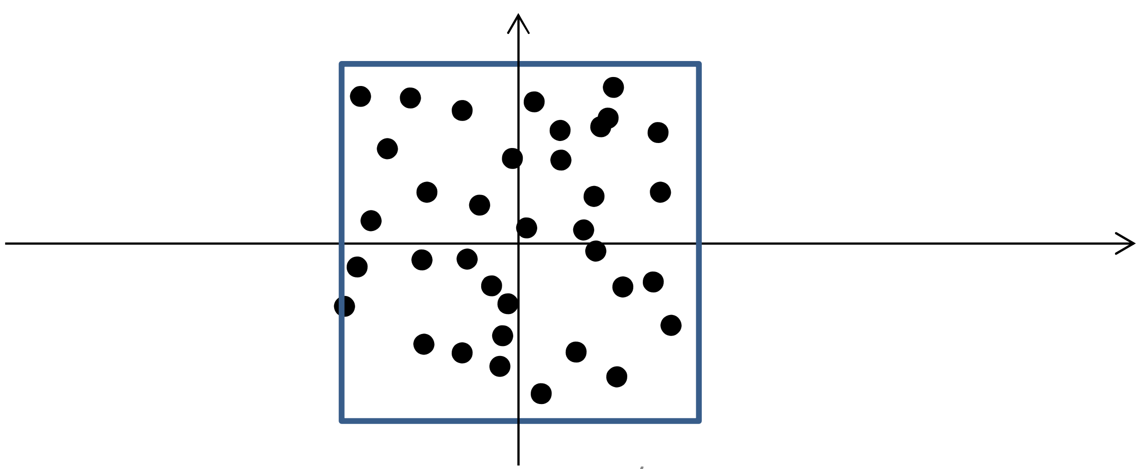
- For independent Gaussian RVs of equal variance, a mixing rotation results in an effectively unchanged distribution
  - The unmixing rotation cannot be determined through inspection of the distribution

Returning to our problem

- Is there a linear transform that can transform a scatter like this



- To something like this:



Zero Mean

- Usual to assume *zero mean* processes
  - Otherwise, some of the math doesn't work well

- $\mathbf{M} = \mathbf{W}\mathbf{H} \qquad \mathbf{H} = \mathbf{A}\mathbf{M}$

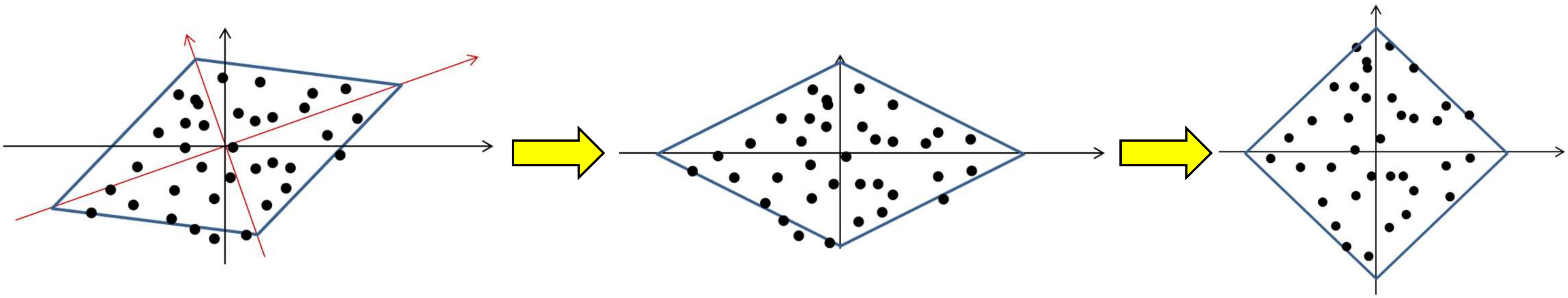
- If mean($\mathbf{M}$) = 0 => mean($\mathbf{H}$) = 0
  - $\mathrm{E}[\mathbf{H}] = \mathbf{A}.\mathrm{E}[\mathbf{M}] = \mathbf{A0} = \mathbf{0}$
  - First step of ICA: Set the mean of $\mathbf{M}$ to 0

$$\mu_{\mathbf{m}} = \frac{1}{cols(\mathbf{M})} \sum_i \mathbf{m}_i$$

$$\mathbf{m}_i = \mathbf{m}_i - \mu_{\mathbf{m}} \qquad \forall i$$

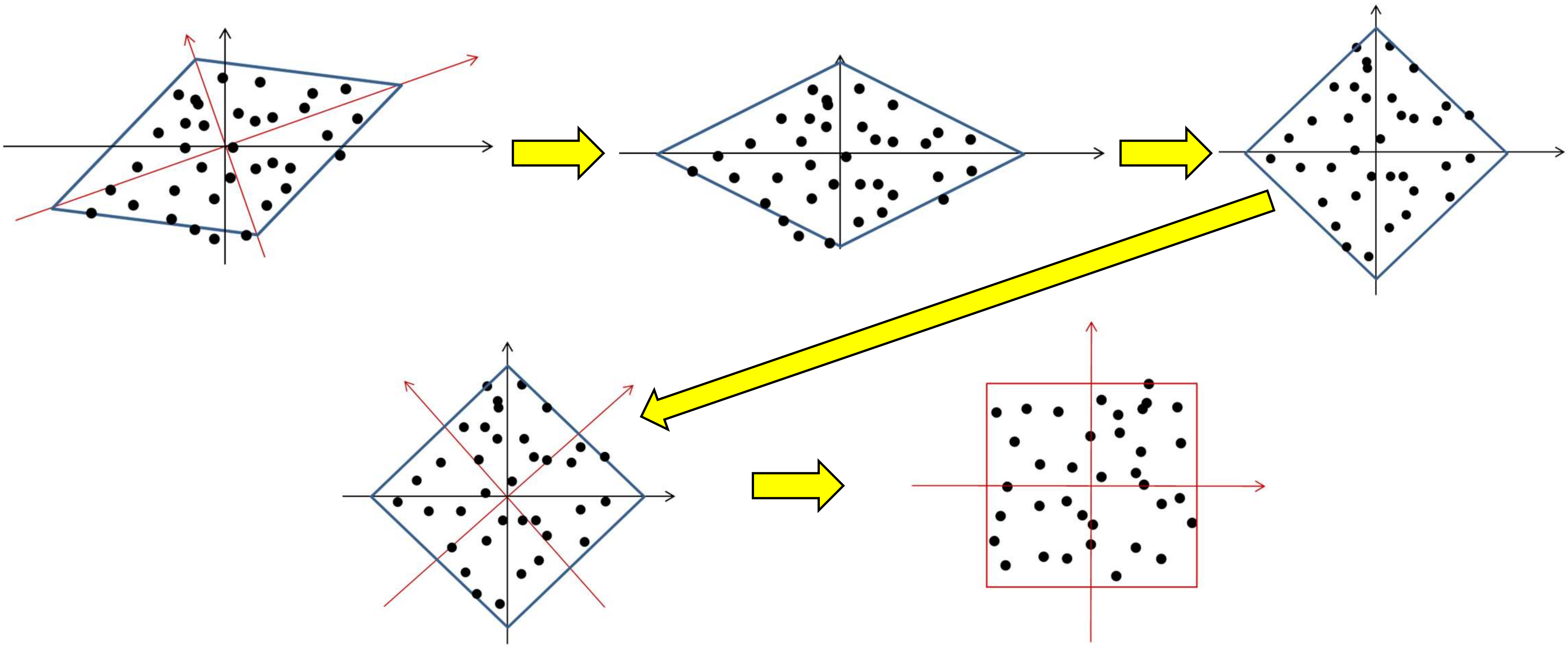
  - $\mathbf{m}_i$ are the columns of $\mathbf{M}$

Actual process



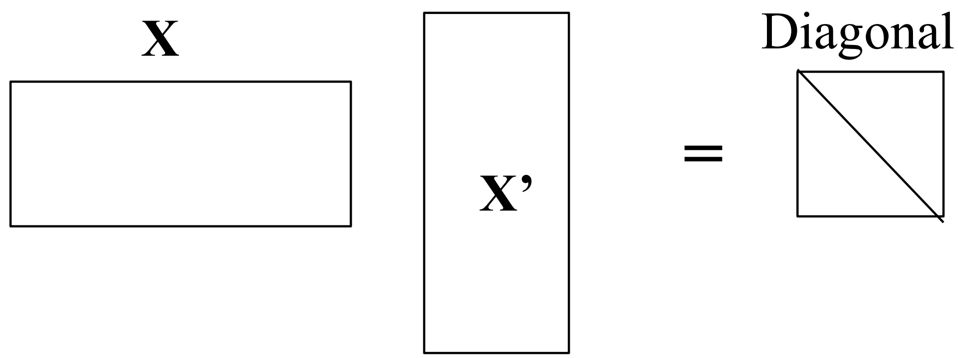
- To simplify the process, we will first *decorrelate* the data and *whiten* it
  - So that the variance is the same along all dimensions

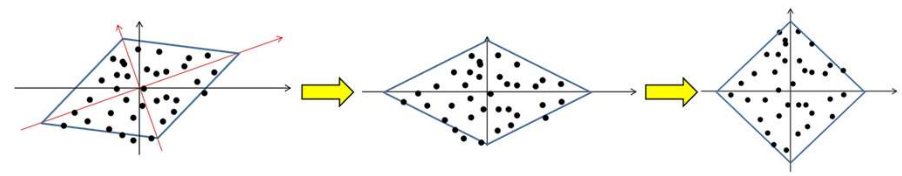
Actual process



- To simplify the process, we will first *decorrelate* the data and *whiten* it
  - So that the variance is the same along all dimensions
- *Then* we search for the axes that make the data independent

Decorrelating and Whitening





- Eigen decomposition $\mathbf{MM}^T = \mathbf{E\Lambda E}^T$
- $\mathbf{C} = \Lambda^{-1/2}\mathbf{E}^T$
- $\mathbf{X} = \mathbf{CM}$

- Not merely decorrelated but ***whitened***
  - $\mathbf{XX}^T = \mathbf{CMM}^T\mathbf{C}^T = \Lambda^{-1/2}\mathbf{E}^T\,\mathbf{E\Lambda E}^T\mathbf{E}\Lambda^{-1/2} = \mathbf{I}$

- $\mathbf{C}$ is the ***whitening matrix***

Uncorrelated != Independent

- Whitening merely ensures that the resulting signals are uncorrelated, i.e.

$$E[\mathbf{x}_i\mathbf{x}_j] = 0 \text{ if } i \mathrel{!}= j$$

- This does not ensure higher order moments are also decoupled, e.g. it does not ensure that

$$E[\mathbf{x}_i^2\mathbf{x}_j^2] = E[\mathbf{x}_i^2]E\,[\mathbf{x}_j^2]$$

- This is *one* of the signatures of independent RVs
- Lets explicitly decouple the fourth order moments

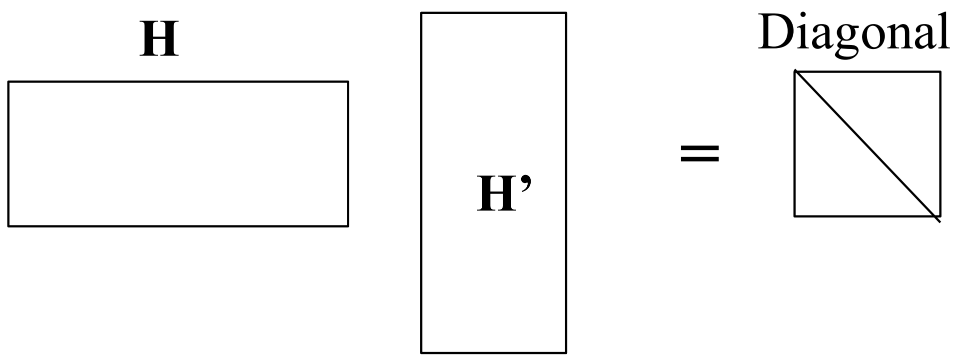
Decorrelating

$\mathbf{H{=}BX}$

$\mathbf{H{=}BCM}$

$\mathbf{A{=}BC}$

$\mathbf{H{=}AM}$



- $\mathbf{X} = \mathbf{CM}$
- $\mathbf{XX}^T = \mathbf{I}$

- **Our objective:** Find the matrix $\mathbf{B}$ that makes the rows of $\mathbf{BX}$ independent
  - $\mathbf{H} = \mathbf{BX}$
- Will multiplying $\mathbf{X}$ by $\mathbf{B}$ *re-correlate* the components?
- Not if $\mathbf{B}$ is *unitary*
  - $\mathbf{BB}^T = \mathbf{B}^T\mathbf{B} = \mathbf{I}$
- $\mathbf{HH}^T = \mathbf{BXX}^T\mathbf{B}^T = \mathbf{BB}^T = \mathbf{I}$
  - Because $\mathbf{XX}^T = \mathbf{I}$
- So we want to find a *unitary* matrix
  - Since the rows of $\mathbf{H}$ are uncorrelated
    - Because they are independent

An ugly algebraic solution

- We *decorrelated* signals by diagonalizing the covariance matrix through Eigen decomposition

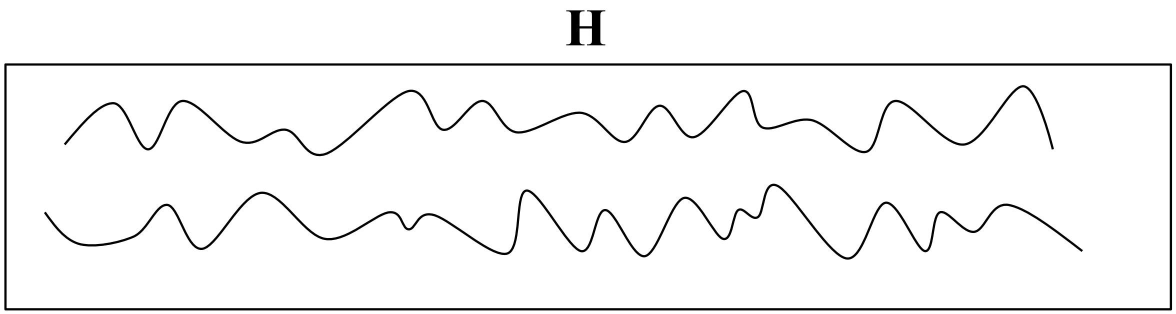
- *Is there a simple matrix we could just similarly diagonalize to make them independent?*
  - Some matrix whose Eigenvector matrix gives us the transform $\mathbf{A}$ such that the rows of $\mathbf{AM}$ are independent

An ugly algebraic solution

- We *decorrelated* signals by diagonalizing the covariance matrix through Eigen decomposition

- *Is there a simple matrix we could just similarly diagonalize to make them independent?*
  - Not really, but there is a matrix we can diagonalize to make *fourth-order* moments independent
    - Just as decorrelation made second-order moments independent

Emulating Independence



- The rows of $\mathbf{H}$ are uncorrelated
  - $E[\mathbf{h}_i\mathbf{h}_j] = E[\mathbf{h}_i]E[\mathbf{h}_j]$
  - $\mathbf{h}_i$ and $\mathbf{h}_j$ are the i^th^ and j^th^ components of any vector in $\mathbf{H}$

- The fourth order moments are independent
  - $E[\mathbf{h}_i\mathbf{h}_j\mathbf{h}_k\mathbf{h}_l] = E[\mathbf{h}_i]E[\mathbf{h}_j]E[\mathbf{h}_k]E[\mathbf{h}_l]$
  - $E[\mathbf{h}_i^2\mathbf{h}_j\mathbf{h}_k] = E[\mathbf{h}_i^2]E[\mathbf{h}_j]E[\mathbf{h}_k]$
  - $E[\mathbf{h}_i^2\mathbf{h}_j^2] = E[\mathbf{h}_i^2]E[\mathbf{h}_j^2]$
  - Etc.

FOBI: Freeing Fourth Moments

- Find $\mathbf{B}$ such that the rows of $\mathbf{H} = \mathbf{BX}$ are independent

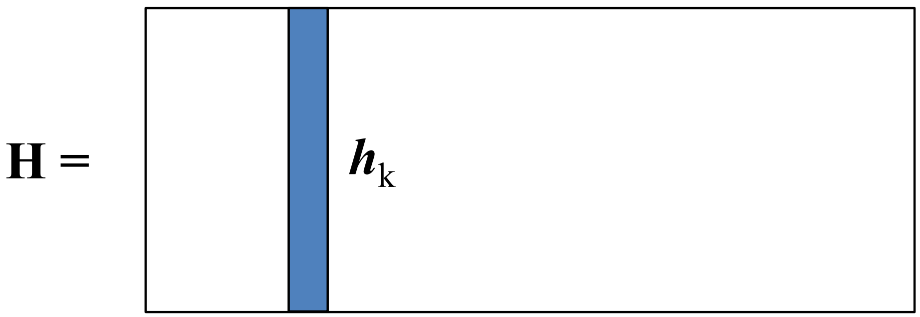
- The fourth moments of $\mathbf{H}$ have the form:
  $\mathrm{E}[\mathbf{h}_i\ \mathbf{h}_j\ \mathbf{h}_k\ \mathbf{h}_l]$

- If the rows of $\mathbf{H}$ were independent
  $\mathrm{E}[\mathbf{h}_i\ \mathbf{h}_j\ \mathbf{h}_k\ \mathbf{h}_l] = \mathrm{E}[\mathbf{h}_i]\ \mathrm{E}[\mathbf{h}_j]\ \mathrm{E}[\mathbf{h}_k]\ \mathrm{E}[\mathbf{h}_l]$

- Solution: Compute $\mathbf{B}$ such that the fourth moments of $\mathbf{H} = \mathbf{BX}$ are decoupled
  - While ensuring that $\mathbf{B}$ is Unitary

- **FOBI: Fourth Order Blind Identification**

ICA: Freeing Fourth Moments



Objective: Find a matrix B such that the rows of H=BX are statistically independent

Define a matrix D that *would* be diagonal if the rows of BX are independent

Compute B such that this matrix becomes diagonal

- Create a matrix of fourth moment terms that would be diagonal if the rows of $\mathbf{H}$ were independent, and diagonalize it

- A good candidate: the weighted correlation matrix of $\mathbf{H}$

$$\boldsymbol{D} = E\left[\|\boldsymbol{h}\|^2 \boldsymbol{h}\boldsymbol{h}^{\mathrm{T}}\right] = \sum_k \|\boldsymbol{h}_k\|^2 \boldsymbol{h}_k \boldsymbol{h}_k^{\mathrm{T}}$$

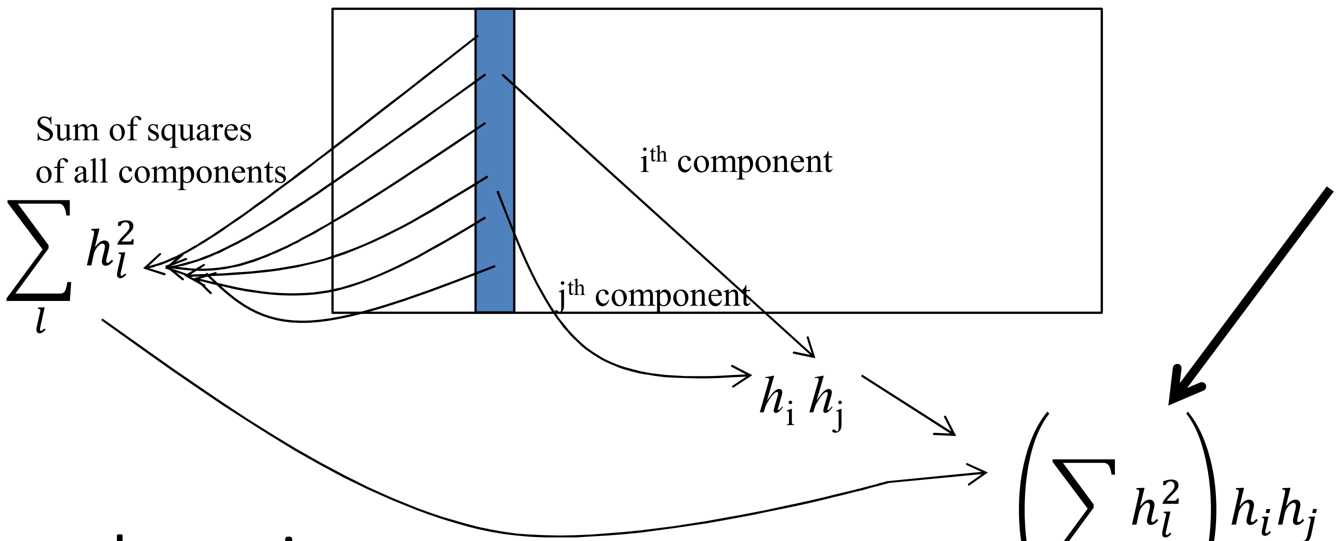
  - $\boldsymbol{h}$ are the columns of $\mathbf{H}$
  - Assuming $\boldsymbol{h}$ is real, else replace transposition with Hermitian

ICA: The D matrix

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & .. \\ d_{21} & d_{22} & d_{23} & .. \\ .. & .. & .. & .. \end{bmatrix}$$

$$\boldsymbol{D} = \boldsymbol{E}\left[\|\boldsymbol{h}\|^2 \boldsymbol{h}\boldsymbol{h}^{\mathrm{T}}\right]$$

$$d_{ij} = \boldsymbol{E}\left[\left(\sum_l h_l^2\right) h_i h_j\right]$$



$$\left(\sum_l h_l^2\right) h_i h_j$$

On the actual matrix

$$\boldsymbol{D} = \frac{\mathbf{1}}{cols(\boldsymbol{H})} \sum_{\boldsymbol{k}} \|\boldsymbol{h}_k\|^2 \boldsymbol{h}_k \boldsymbol{h}_k^{\mathrm{T}}$$

$$d_{ij} = \frac{1}{cols(\mathbf{H})} \sum_k \left(\sum_l h_{kl}^2\right) h_{ki} h_{kj}$$

ICA: The D matrix

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & .. \\ d_{21} & d_{22} & d_{23} & .. \\ .. & .. & .. & .. \end{bmatrix}$$

$$d_{ij} = \frac{1}{cols(\mathbf{H})} \sum_k \left(\sum_l h_{kl}^2 \right) h_{ki} h_{kj}$$

- If the h_i terms were independent and zero mean
- For i != j (off-diagonal elements)

$$E\left[h_i h_j \sum_l h_l^2\right] = E[h_i^3]E[h_j] + E[h_i]E\left[h_j^3\right] + E[h_i]E[h_j] \sum_{l \neq i, l \neq j} E[h_l^2] = \mathbf{0}$$

- For $i = j$ (diagonal elements)
  - $E[h_i h_j \sum_l h_l^2] = E[h_i^4] + E[h_i^2] \sum_{l \neq i} E[h_l^2] \neq \mathbf{0}$

- i.e., if h_i were independent, D would be a diagonal matrix
  - **Let us diagonalize *D***

Diagonalizing D

- Recall: $\mathbf{H} = \mathbf{BX}$
  - $\mathbf{B}$ is what we're trying to learn to make $\mathbf{H}$ independent
  - Assumption: $\mathbf{B}$ is unitary, i.e. $\mathbf{B}^T\mathbf{B} = \mathbf{I}$

Objective: Find a matrix B such that the rows of H=BX are statistically independent

Define a matrix D that *would* be diagonal if the rows of BX are independent

Compute B such that this matrix becomes diagonal

- Note: if $\mathbf{H} = \mathbf{BX}$, then each vector $\mathbf{h} = \mathbf{Bx}$
- The fourth moment matrix of $\mathbf{H}$ is
- $\mathbf{D} = E[\mathbf{h}^T\mathbf{h}\ \mathbf{h}\ \mathbf{h}^T] = E[\mathbf{x}^T\mathbf{B}^T\mathbf{Bx}\ \mathbf{Bx}\ \mathbf{x}^T\ \mathbf{B}^T]$

$$= E[\mathbf{x}^T\mathbf{x}\ \mathbf{Bx}\ \mathbf{x}^T\ \mathbf{B}^T]$$

$$= \mathbf{B}\ E[\mathbf{x}^T\mathbf{x}\ \mathbf{xx}^T]\mathbf{B}^T$$

$$= \mathbf{B}\ E[||\mathbf{x}||^2\ \mathbf{xx}^T]\mathbf{B}^T$$

Diagonalizing D

- Objective: Estimate $\mathbf{B}$ such that the fourth moment of $\mathbf{H} = \mathbf{BX}$ is diagonal

- Compose $\mathbf{D_x} = \sum_{\boldsymbol{k}} \|\mathbf{x}_k\|^2 \mathbf{x_k x}_k^{\mathrm{T}}$

- Diagonalize $\mathbf{D_x}$ via Eigen decomposition

  $\mathbf{D_x} = \mathbf{U}\Lambda_{\mathrm{H}}\mathbf{U}^{\mathrm{T}}$

- $\mathbf{B} = \mathbf{U}^{\mathrm{T}}$
  - **That's it!!!!**

B frees the fourth moment

$$\mathbf{D_x} = \mathbf{U}\Lambda\mathbf{U}^T \;\; ; \quad \mathbf{B} = \mathbf{U}^T$$

- $\mathbf{U}$ is a unitary matrix, i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$ (identity)
- $\mathbf{H} = \mathbf{B}\mathbf{X} = \mathbf{U}^T\mathbf{X}$
  - $\mathbf{h} = \mathbf{U}^T\mathbf{x}$

- The fourth moment matrix of $\mathbf{H}$ is

$$\mathbf{D} = E[||\mathbf{h}||^2 \, \mathbf{h}^T]$$

$$\begin{aligned}\mathbf{D} &= \mathbf{U}^T E[||\mathbf{x}||^2 \, \mathbf{x}\mathbf{x}^T]\mathbf{U} \\ &= \mathbf{U}^T \mathbf{D_x} \mathbf{U} \\ &= \mathbf{U}^T \mathbf{U} \Lambda_H \mathbf{U}^T \mathbf{U} = \Lambda_H\end{aligned}$$

- The fourth moment matrix of $\mathbf{H} = \mathbf{U}^T\mathbf{X}$ is Diagonal!!

Overall Solution

- Objective: Estimate A such that the rows of $\mathbf{H} = \mathbf{AM}$ are independent
- Step 1: *Whiten M*
  - $\mathbf{C} = \Lambda^{-1/2}\mathbf{E}^T$ where Λ and $\mathbf{E}$ are the eigen value and eigen vector matrices of $\mathbf{MM}^T$
  - $\mathbf{X} = \mathbf{CM}$
- Step 2: Free up fourth moments on $\mathbf{X}$
  - $\mathbf{B}$ is the (transpose of the) matrix of Eigenvectors of $\mathbf{X}.diag(\mathbf{X}^T\mathbf{X}).\mathbf{X}^T$
  - $\mathbf{A} = \mathbf{BC}$

FOBI for ICA

- Goal: to derive a matrix $\mathbf{A}$ such that the rows of $\mathbf{AM}$ are independent
- Procedure:
  1. "Center" $\mathbf{M}$
  2. Compute the autocorrelation matrix $\boldsymbol{R}_{MM}$ of $\mathbf{M}$
  3. Compute whitening matrix $\mathbf{C}$ via Eigen decomposition
     $\boldsymbol{R}_{\mathrm{MM}} = \mathbf{E}\Lambda\mathbf{E}^{\mathrm{T}}, \quad \mathbf{C} = \Lambda^{-1/2}\mathbf{E}^{\mathrm{T}}$
  4. Compute $\mathbf{X} = \mathbf{CM}$
  5. Compute the fourth moment matrix $\mathbf{D}' = E[\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^{\mathrm{T}}]$
  6. Diagonalize $\mathbf{D}'$ via Eigen decomposition
  7. **$\mathbf{D}' = \mathbf{U}\Lambda_{\mathrm{H}}\mathbf{U}^{\mathrm{T}}$**
  8. Compute $\mathbf{A} = \mathbf{U}^{\mathrm{T}}\boldsymbol{C}$
- The fourth moment matrix of $\mathbf{H{=}AM}$ is diagonal
  – **Note that the autocorrelation matrix of $\mathbf{H}$ will also be diagonal**

ICA by diagonalizing moment matrices

- FOBI is not perfect
  - Only a subset of fourth order moments are considered
    - Diagonalizing the particular fourth-order moment matrix we have chosen is not guaranteed to diagonalize every other fourth-order moment matrix
- JADE: (Joint Approximate Diagonalization of Eigenmatrices), J.F. Cardoso
  - Jointly diagonalizes multiple fourth-order cumulant matrices

Poll 3

- Which of the following statements are true of FOBI
  - It computes a transform that makes *all* fourth-order moments independent
  - It requires a first pre-whitening step
  - The transform is the Eigenvector matrix of the fourth-order moment matrix
  - The transform is the product of the Eigenvector matrix of the fourth-order moment matrix of the whitened data, and the whitening matrix obtained through PCA

Poll 3

- Which of the following statements are true of FOBI
  - It computes a transform that makes *all* fourth-order moments independent
  - **It requires a first pre-whitening step**
  - The transform is the Eigenvector matrix of the fourth-order moment matrix
  - **The transform is the product of the Eigenvector matrix of the fourth-order moment matrix of the whitened data, and the whitening matrix obtained through PCA**

Lets try a different tack

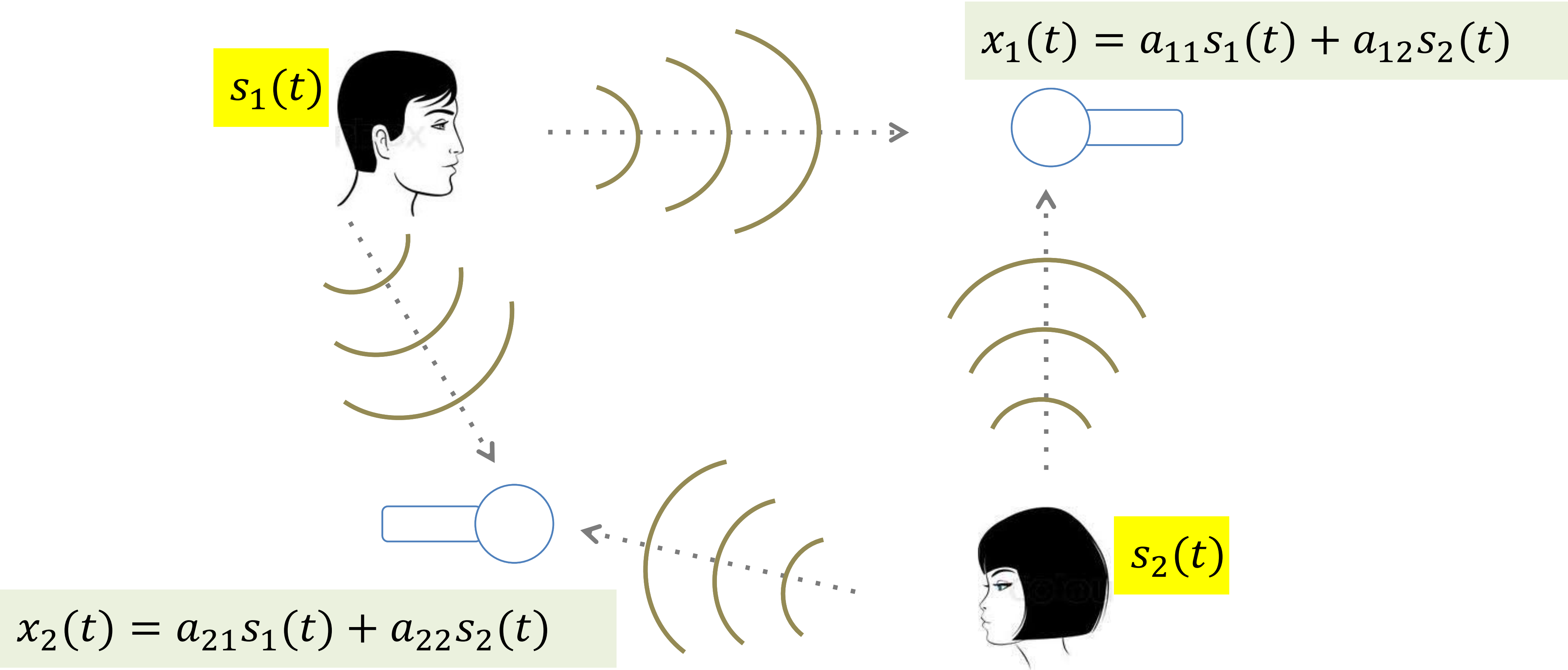
- Use the statistical properties of mixing…

The Central Limit Theorem

- Sum of independent random variables will tend toward a Gaussian distribution
- Even if the independent random variables don't have a Gaussian distribution!

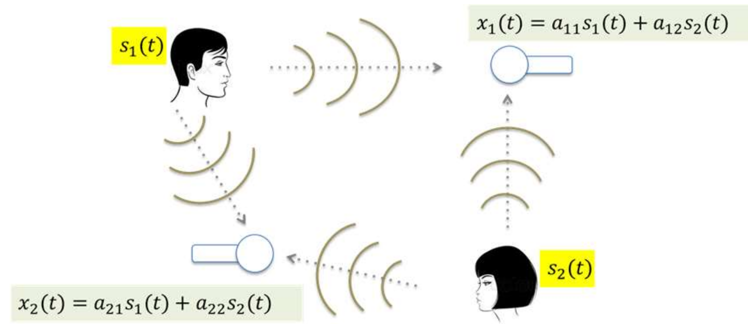
- The sum will *almost always* be *"more"* Gaussian than the component signals
  - Even if the independent RVs are not Gaussian

Changing notation for a bit



- Two people speak simultaneously are recorded by two microphones
  - Each recorded signal is a mixture of both signals
- Find a linear transform that unmixes them

Problem setting and notation



- Independent signals $s_1 \ldots s_N$ (arranged as a vector $\boldsymbol{s}$) have been mixed by mixing matrix $\boldsymbol{A}$ to generate mixed output $\boldsymbol{x}$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{22} & a_{22} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$$

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

- We need to find a matrix $\boldsymbol{W}$ that will unmix $\boldsymbol{x}$ to recover $\boldsymbol{s}$

$$\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x} \qquad \boldsymbol{s.t.}\ \boldsymbol{y} \approx \boldsymbol{x}$$

The Central Limit Theorem & ICA

Let each s_i be identically distributed

Let's obtain one of the sources

$$y = w^T \boldsymbol{x}$$

Here, $\boldsymbol{w}$ is a column of $\boldsymbol{W}$

The Central Limit Theorem & ICA

$$y = w^T \boldsymbol{x}$$

Suppose, w^T is a row of the mixing matrix's inverse ($W^T = A^{-1}$) . Then y would be one of the independent sources:

$$\boldsymbol{x} = A\boldsymbol{s} \ \rightarrow \boldsymbol{s} = A^{-1}\boldsymbol{x}$$

The Central Limit Theorem & ICA

Useful Relations: $\boldsymbol{x} = A\boldsymbol{s}$ $\boldsymbol{y} = W^T\boldsymbol{x}$

$y = w^T\boldsymbol{x}$

Let's define a convenient variable:

$$z = A^T w$$

And let's do some substitutions:

$$y = w^T\boldsymbol{x} \rightarrow y = w^T A\boldsymbol{s} \rightarrow y = (w^T A)\boldsymbol{s} \rightarrow y = (A^T w)^T\boldsymbol{s} \rightarrow y = z^T\boldsymbol{s}$$

The Central Limit Theorem & ICA

Useful Relations:

$$x = A\boldsymbol{s}$$
$$y = w^T\boldsymbol{x}$$
$$y = z^T\boldsymbol{s}$$

What does this last relation mean?

We want y to be ONE OF the independent sources

The Central Limit Theorem & ICA

Useful Relations: $x = A\boldsymbol{s}$

$$y = w^T \boldsymbol{x}$$

$$y = z^T \boldsymbol{s}$$

1. y is a linear combination of sources

What does this do for us?

The Central Limit Theorem & ICA

Useful Relations: $x = A\boldsymbol{s}$

$$y = w^T\boldsymbol{x}$$

$$y = z^T\boldsymbol{s}$$

1. $\boldsymbol{y}$ is a linear combination of sources
2. If $\boldsymbol{y}$ is one of the sources, then $\boldsymbol{z} = [\mathbf{0}, \ldots, \mathbf{1}, \ldots, \mathbf{0}]$.

What does this do for us?

The Central Limit Theorem & ICA

Useful Relations:

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

$$y = w^T \boldsymbol{x}$$

$$y = z^T \boldsymbol{s}$$

1. $\boldsymbol{y}$ is a linear combination of sources
2. If $\boldsymbol{y}$ is one of the sources, then $\boldsymbol{z} = [0, \ldots, 1, \ldots, 0]$.

$$s_3 = z^T \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} \rightarrow s_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$$

What does this do for us?

The Central Limit Theorem & ICA

Useful Relations:

$$x = A\boldsymbol{s}$$
$$y = w^T\boldsymbol{x}$$
$$y = z^T\boldsymbol{s}$$

1. $\boldsymbol{y}$ is a linear combination of sources
2. If $\boldsymbol{y}$ is one of the sources, then $\boldsymbol{z} = [0, \ldots, 1, \ldots, 0]$.
3. Since the sources are independent R.V.'s, any ***mixed*** $\boldsymbol{y}$ is "more Gaussian" than any of the sources

What does this do for us?

The Central Limit Theorem & ICA

Useful Relations:

$$x = A\boldsymbol{s}$$
$$y = w^T\boldsymbol{x}$$
$$y = z^T\boldsymbol{s}$$

1. $\boldsymbol{y}$ is a linear combination of sources
2. If $\boldsymbol{y}$ is one of the sources, then $\boldsymbol{z} = [0, \ldots, 1, \ldots, 0]$.
3. Since the sources are independent R.V.'s, any ***mixed*** $\boldsymbol{y}$ is "more Gaussian" than any of the sources
4. If $\boldsymbol{y}$ is one of the sources, $\boldsymbol{y}$ is the *least Gaussian!*

What does this do for us?

The Central Limit Theorem & ICA

Useful Relations:

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

$$y = w^T\boldsymbol{x} \qquad y = z^T\boldsymbol{s}$$

Recall: we are given *x*.

Recall: we are not given *s*.

Recall: *z* is a variable we defined for convenience

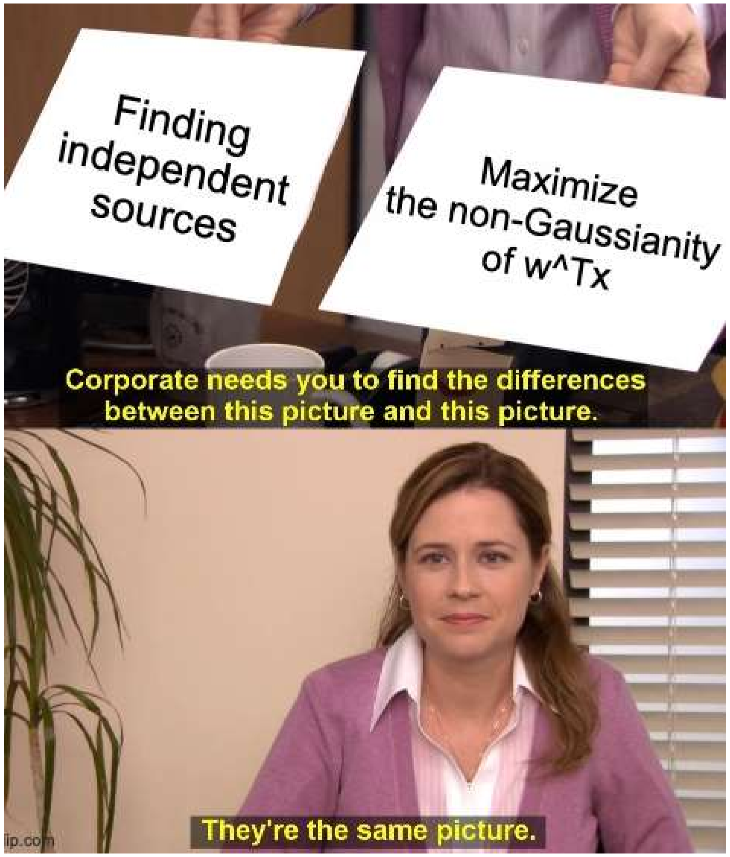
Let's pick a *w* that maximizes the non-Gaussianity of *y*.

This should force *z* to have just one non-zero component

***y* will then be one of the independent sources.**

BIG GOAL™

MAXIMIZE THE NON-GAUSSIANITY OF $y = w^T x$



What they are and what they proxy

CONTRAST FUNCTIONS

“more Gaussian” & “least Gaussian”

- How can we measure Gaussianity
- If we can measure Gaussianity, can we produce a way to optimize over that?
- If we can optimize non-Gaussianity, can we solve ICA?

Fortunately, there are lots of ways to measure non-Gaussianity!

Kurtosis

A very clear formula:

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

$$Kurt[X] = E\left[X^4\right] - 3\left(E\left[X^2\right]\right)^2)$$

Kurtosis

$$Kurt[X] = E[X^4] - 3\left(E[X^2]\right)^2)$$

Note: For a multivariate normal distribution with unit variance, $E[X^4] = 3\left(E[X^2]\right)^2 = 3$.

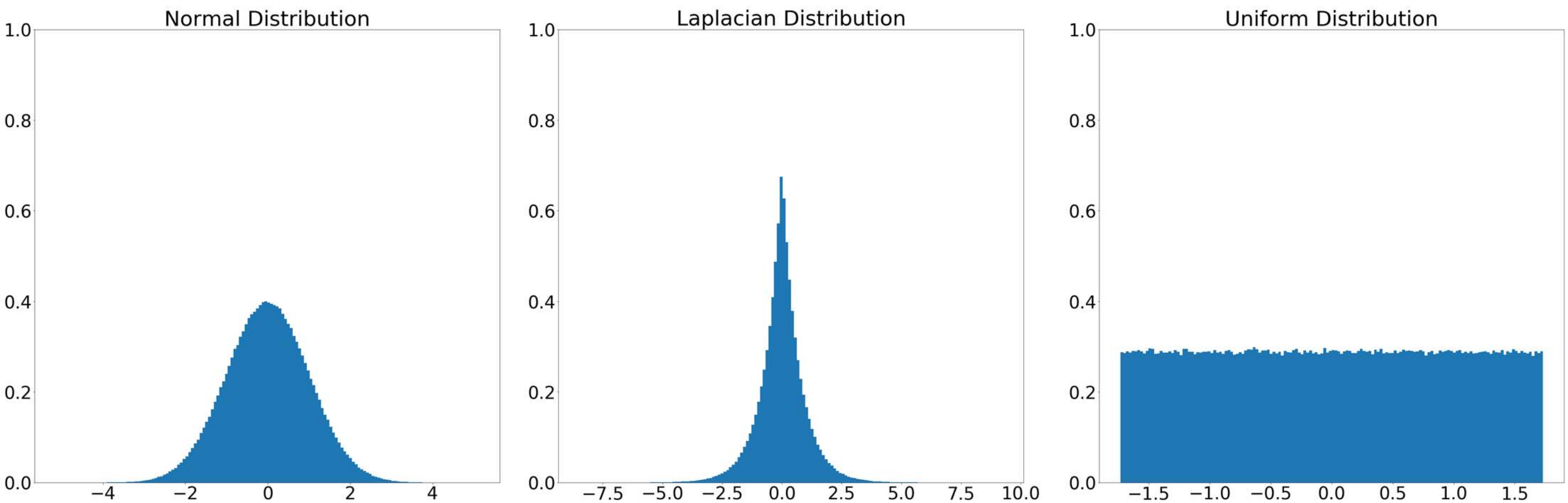
Note: for a multivariate normal distribution with unit variance, $3\left(E[X^2]\right)^2 = 3(1)^2 = 3.$

So, if $X \sim N(0,1)$, $Kurt[X] = 0.$

Kurtosis

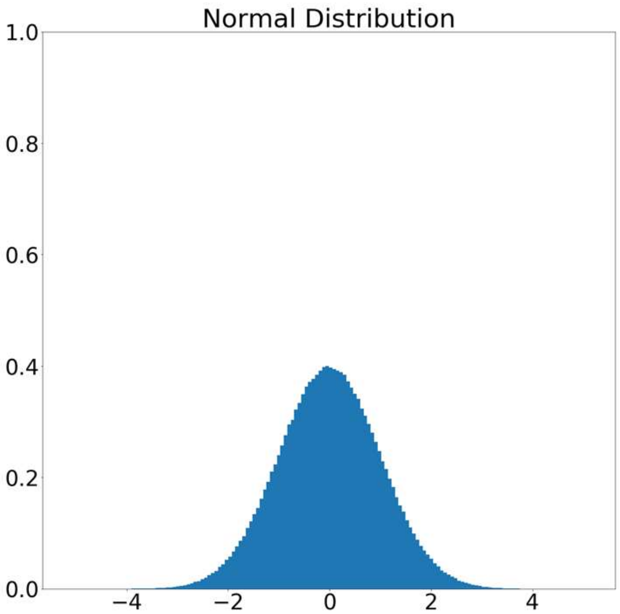
- A measure of how heavy the tails of a distribution are

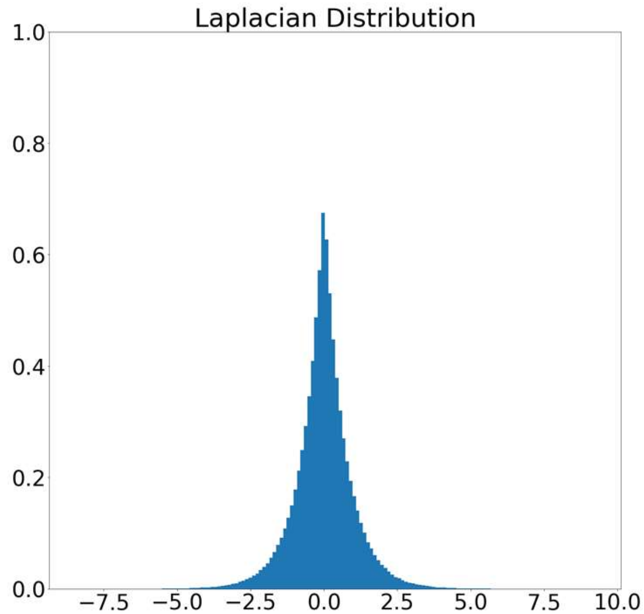
Generated with 1,000,000 samples.

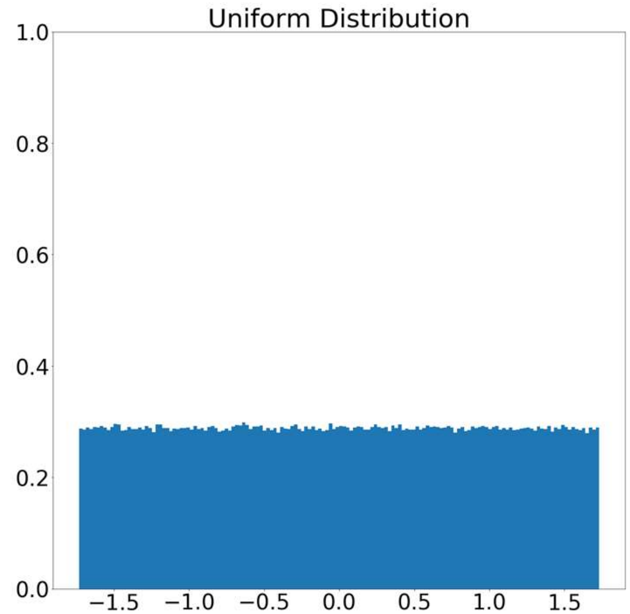


Ground Truth Kurt[X] = 0.0 **Ground Truth Kurt[X] = 3.0** **Ground Truth Kurt[X] = -1.2**

Generated with 1,000,000 samples.







Ground Truth Kurt[X] = 0.0

Calculated Kurt[X] = 0.0

Ground Truth Kurt[X] = 3.0

Calculated Kurt[X] = 3.023

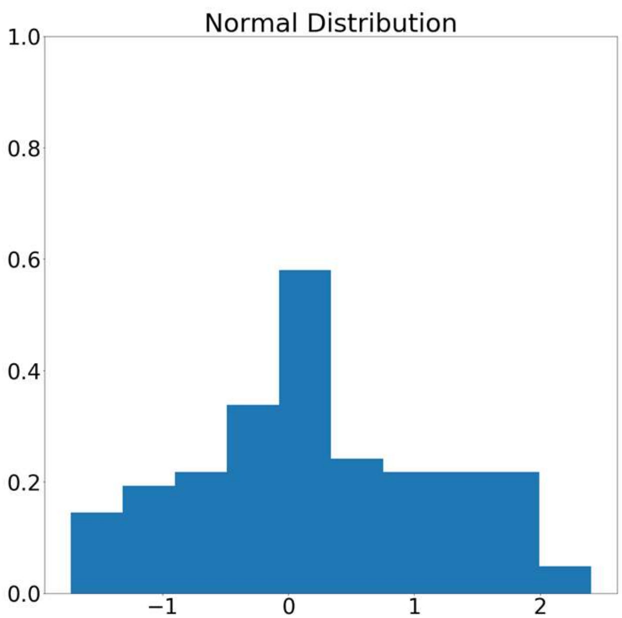
Ground Truth Kurt[X] = -1.2

Calculated Kurt[X] = -1.199

Kurtosis

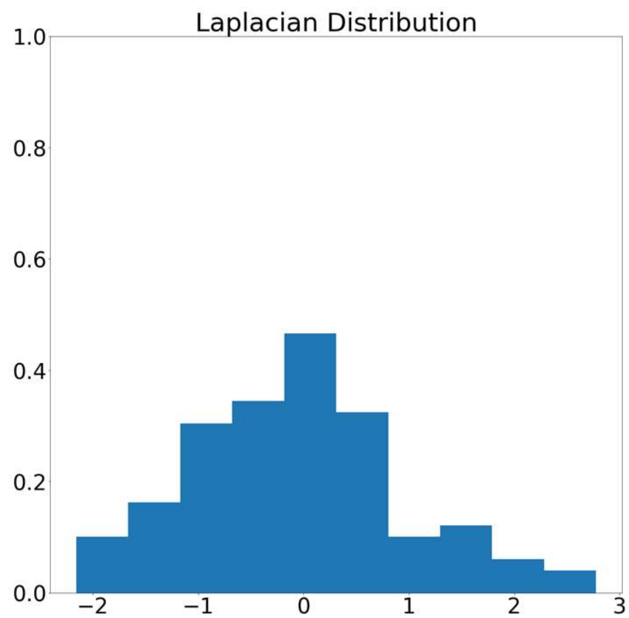
- How would we optimize?
- Use the absolute value of kurtosis
- For a Gaussian R.V., its kurtosis is 0
- Therefore, we want to maximize the kurtosis of the distribution

Generated with 100 samples.



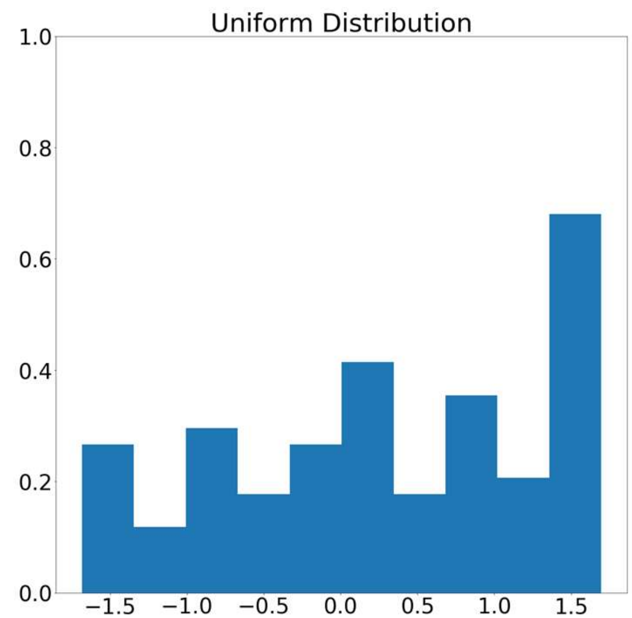
Ground Truth Kurt[X] = 0.0

Calculated Kurt[X] = -0.54



Ground Truth Kurt[X] = 3.0

Calculated Kurt[X] = 0.121



Ground Truth Kurt[X] = -1.2

Calculated Kurt[X] = 1.15

Kurtosis

- Benefits
  - computationally easy
  - some nice linearity properties
  - widely used!
- Disadvantages
  - Susceptible to outliers
  - Few data points leads to bad estimate

Not a robust measure of Gaussianity!

Negentropy

- Entropy:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

From last lecture: minimal number of bits sent for an optimal code

Negentropy

- Entropy: a measure of surprise
- R.V. that is “more random” will have a larger entropy
  - More bits needed to send
- R.V. that is “less random” will have a smaller entropy
  - Fewer bits needed to send
  - Spiky PDFs

What is the entropy of a Gaussian random variable?

Negentropy

- Entropy of a Gaussian: depends but it's the largest possible value of any distribution with equal variance

How does this help us?

Negentropy

Define:

$$J(X) = H(X_{gauss}) - H(X)$$

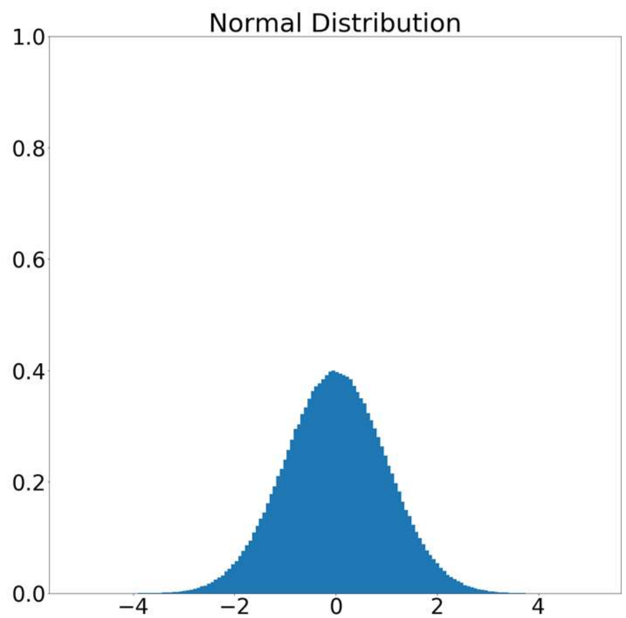
X_{gauss} is a Gaussian with the same covariance matrix as ***X.***

With this definition: ***J(X)*** > 0 and ***J(X) = 0*** if ***X*** is Gaussian

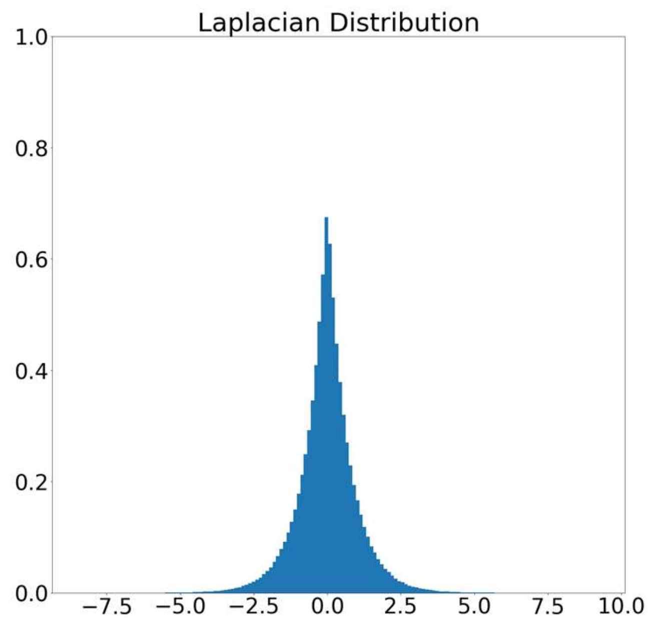
So, to minimize Gaussianity, we want to maximize negentropy!

Negentropy

Generated with 1,000,000 samples.



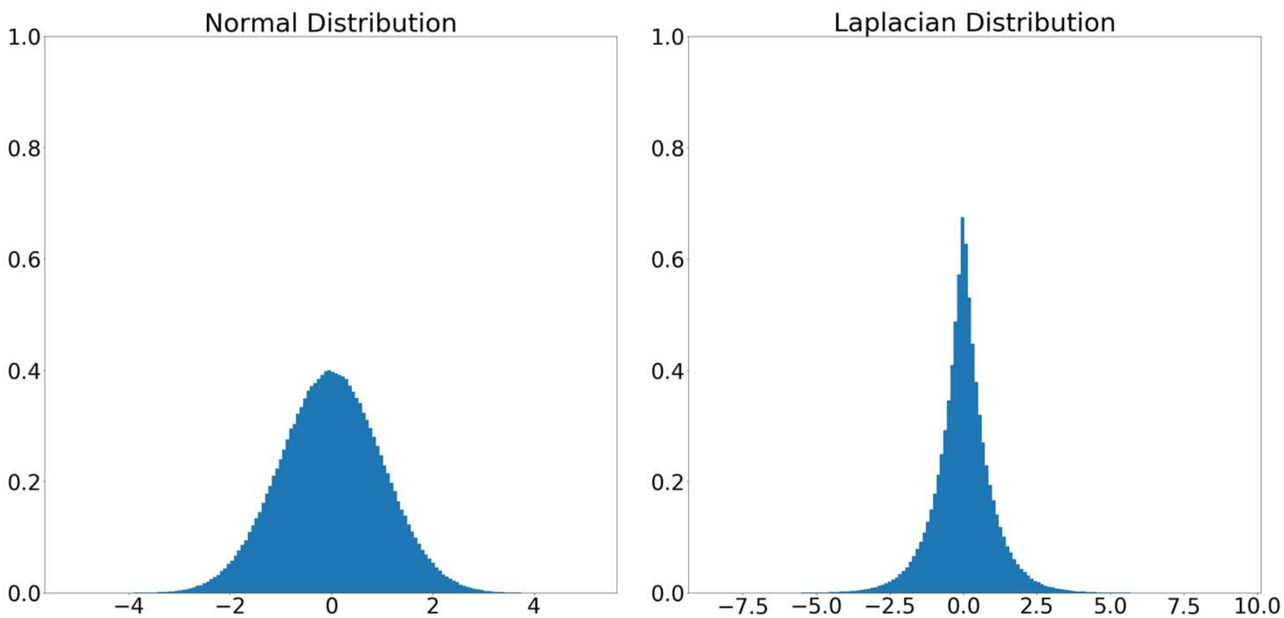
Ground Truth J[X] = 0.0



Ground Truth J[X] = 1.07

Negentropy

Generated with 1,000,000 samples.



Ground Truth J[X] = 0.0
Calculated J[X] = 0.08

Ground Truth J[X] = 1.07
Calculated J[X] = 0.717

Negentropy

- Advantages:
  - Very well justified measure of Gaussianity
  - *Optimal* measure of Gaussianity
- Disadvantages
  - Computationally hard
  - Must estimate the PDF of a R.V.: *always a fun thing to do :/*

We will usually approximate negentropy and maximize over that

When you’re tired of looking at math slides and want to build something

ALGORITHMS

Maximizing an approximation to negentropy.

FASTICA

General principle

- Want to maximize $H(\nu) - H(X)$
  - Where ν is a 0 mean unit variance Gaussian RV and the variance of X is 1 (whitened)

$$\max(E[-\log P(\nu)] - E[-\log P(X)])$$
$$= \max(E[\log P(X)] - E[\log P(\nu)])$$

- Taking expectations requires knowledge of $\log P(X)$
  - Which we do not know

- Instead we will take a different approach to maximize the difference between $P(X)$ and a Gaussian
- Ensure that the expected value of *every moment of X is maximally different from the corresponding moment of ν*
  - $\max div(E[X^n], E[\nu^n])$ for every n

Maximizing the gap

$$\max div(E[X^n], E[\nu^n]) \,\forall\, n \;\approx$$
$$\max \mathrm{a}_1 div(E[X], E[\nu]) + \mathrm{a}_2 div\left(E[X]^2, E\left[\nu^2\right]\right) + \mathrm{a}_3 div\left(E\left[X^3\right], E\left[\nu^3\right]\right) + \cdots$$

- Not tractable: will require explicit computation or estimation of all inifinite moments
  - Or at least a whole lot of high-order moments
  - Which are verrrrrrrrrrrry noisy to estimate

- Instead do the following

$$\max div\left(E\left[\mathrm{a}_1 X + a_2 X^2 + a_3 X^3 + \cdots\right], E\left[a_1 \nu + a_2 \nu^2 + a_3 \nu^3 \ldots.\right]\right)$$

- Or alternately

$$\max \;\; div(G(X), G(\nu))$$

- Where $G(X)$ is any function that has a fast convergent Power series expansion:

$$G(X) = \sum_{n=0}^{\infty} (x - x_0)^n$$

  - The power series must include at least four terms to be meaningful
- Using the squared L2 divergence we get
  - $\max J(X) \;\; where \;\; J(X) \propto \left[E[G(X)] - E[G(\nu)]\right]^2$

FastICA

- Hyvärinen 2000
- Uses an approximation of negentropy:

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

ν is a Gaussian variable with zero-mean and unit-variance

G are nonquadratic functions

FastICA: the *G* function

- *G* just needs to be non-quadratic
  - Ideally a function whose polynomial expansion includes all higher powers of the argument
    - Maximizing negentropy will "free" up the moments of those higher powers
- Some weird forms:

$$G(u) = \frac{1}{a_1} \log\cosh(a_1 u)$$

$$G(u) = -\frac{1}{a_2} \exp\left(-\frac{a_2 u^2}{2}\right)$$

$$G(u) = \frac{1}{4} u^4$$

FastICA: comments

- Maximize $J(X) = \left[E[G(X)] - E[G(\nu)]\right]^2$ while ensuring $var(X) = 1$
  - Pre-whiten the data
- Taking actual expectations is not possible
- Instead use the empirical average over samples
- Can be performed in online manner

FastICA

1. Pre-whiten the data
2. Choose an initial $\boldsymbol{w}$
3. Let $w^{+} = E\left[xG'\left(w^{T}x\right)\right] - E\left[G''\left(w^{T}x\right)\right]w$
4. Normalize: $w = w^{+}/\ \|w^{+}\|$
5. Check convergence, head back to 3!

- Normalization of w maintains variance = 1

FastICA: Derivation

- Newton's Method
- Maximize:

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

- Constrain:

$$\|w\|^2 = 1$$

FastICA: Industry Standard

- Basically the industry standard implementation of ICA:
  - https://github.com/scikit-learn/scikit-learn/blob/0fb307bf3/sklearn/decomposition/_fastica.py#L304

Poll 4

- Which of the following are true of FastICA
  - It derives a linear transform that frees up fourth moments
  - It finds the independent directions along which the distributions of the data are maximally non-Gaussian
  - It is a *batch* algorithm
  - It is an *online* algorithm

Poll 4

- Which of the following are true of FastICA
  - It derives a linear transform that frees up fourth moments
  - **It finds the independent directions along which the distributions of the data are maximally non-Gaussian**
  - It is a *batch* algorithm
  - **It is an *online* algorithm**

Speech-Music Example

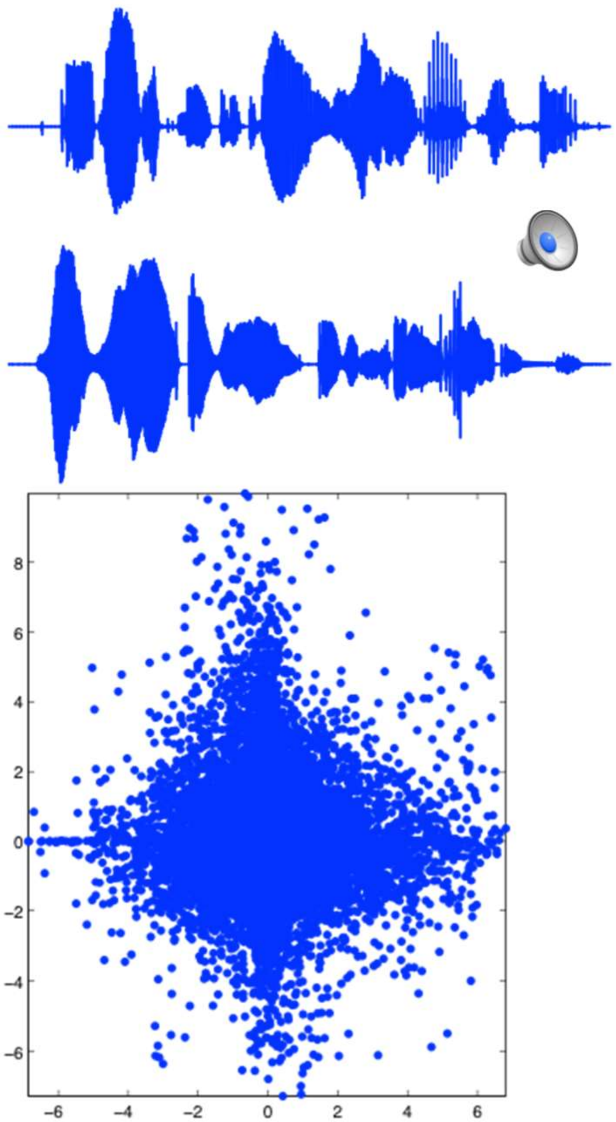
- Te-Won Lee @ UCSD

Mixed

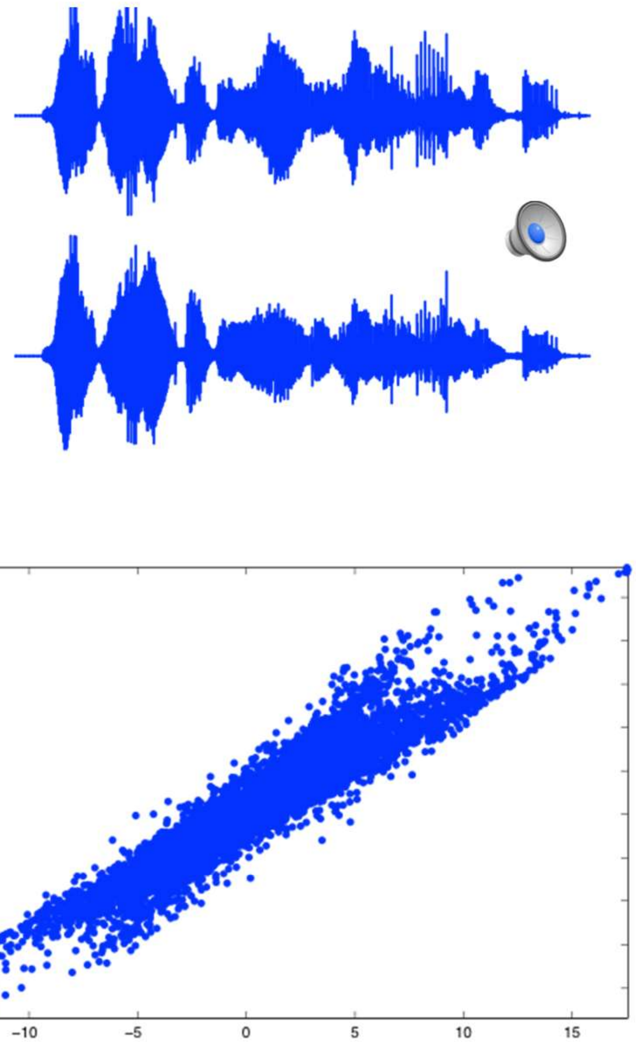
Separated

Another example!

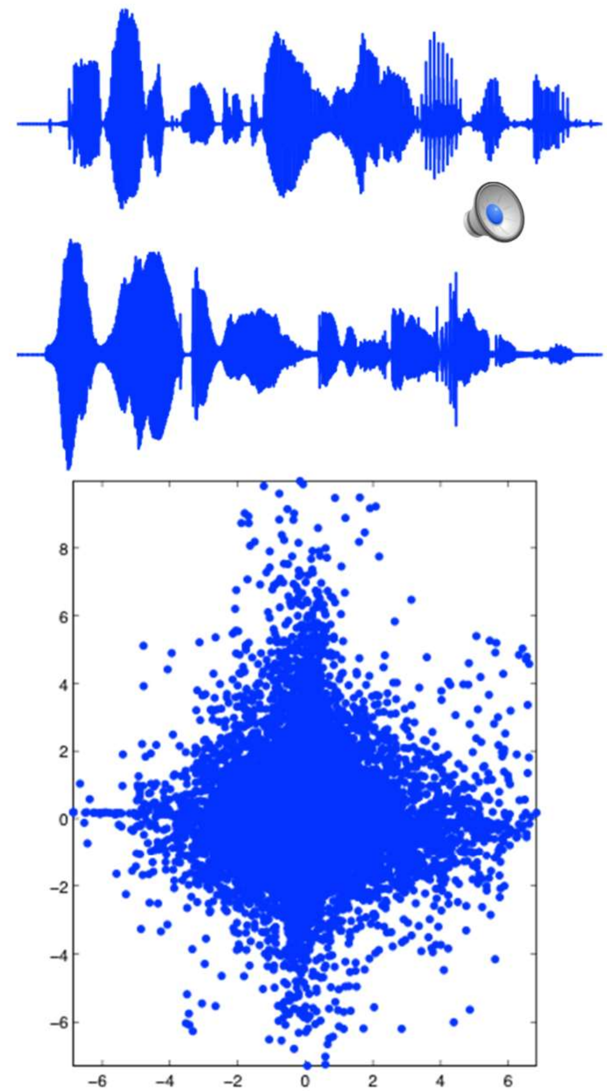
Input



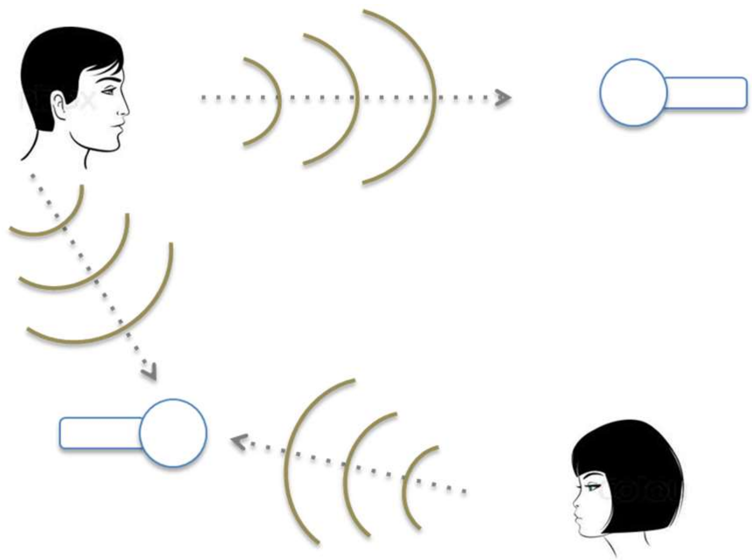
Mix



Output



In Reality



- Mixed signals are not instantaneous mixtures
  - The signals arrive with different delays at the two microphones

$$x_1 = a_{11}s_1(t - t_{11}) + a_{12}s_2(t - t_{12}),$$
$$x_2 = a_{21}s_1(t - t_{21}) + a_{22}s_2(t - t_{22})$$

  - The time-delay issue is hard for ICA to deal with

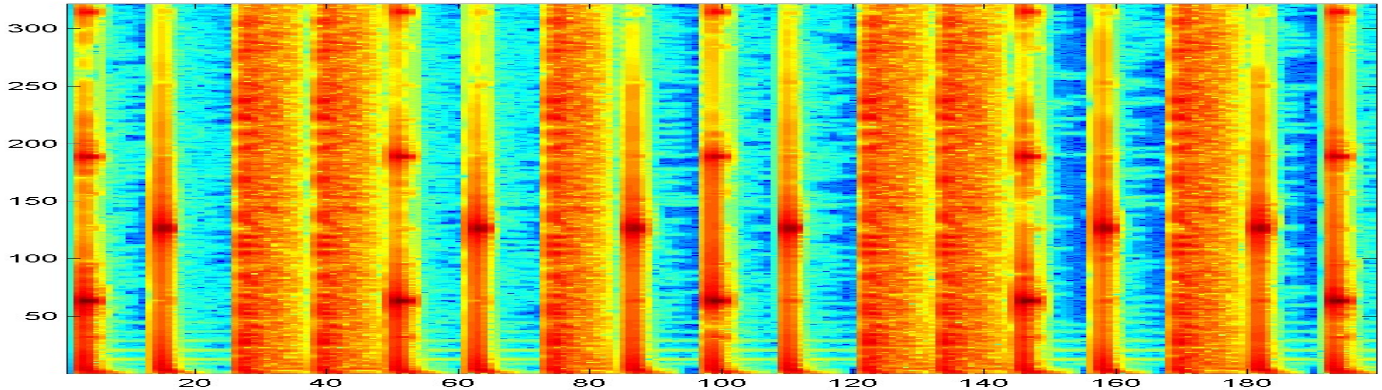
- You must do some clever things for it to work out

Some Explicit Limitations

- ICA is identifiable up to:
  - a sign change (plus or minus)
  - a scaling factor
  - This is just from the model: $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$

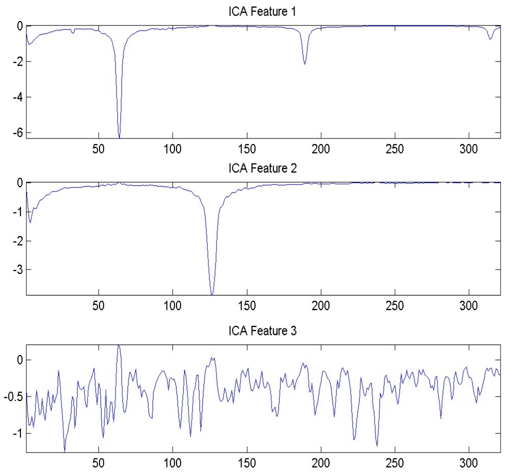
- ICA (unlike PCA) doesn't have a notion of importance
  - The order of the sources doesn't matter.
  - It's unique up to permutation as well.

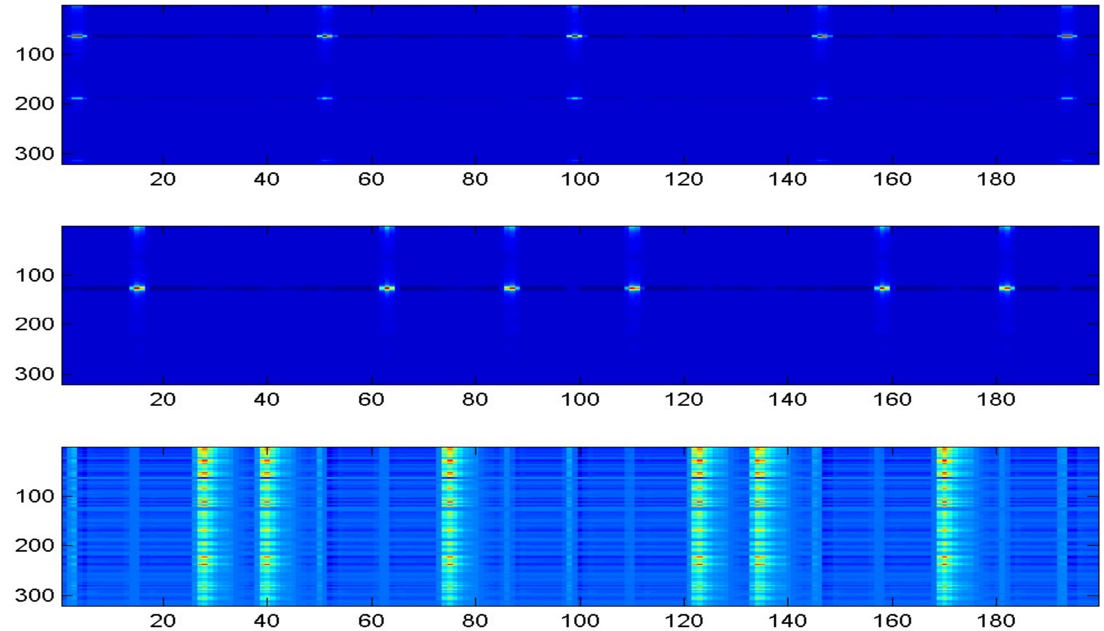
Another Example



- Three instruments..
  - $M = NS$,
  - $S = WM$ (through ICA)
  - $N = W^{+}$

The Notes

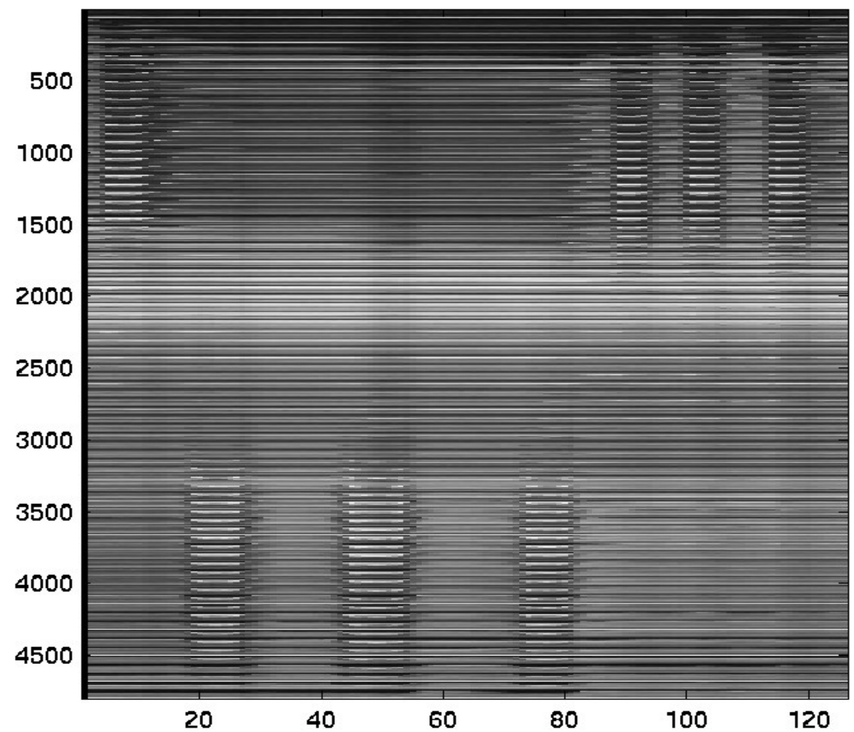




- Three instruments..

ICA for data exploration

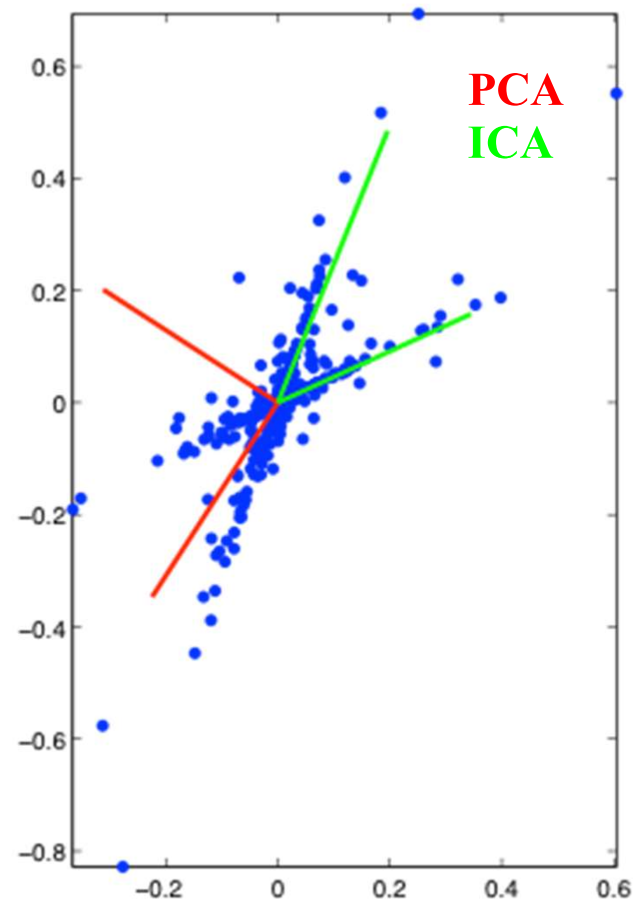
- The "bases" in PCA represent the "building blocks"
  - Ideally notes
- Very successfully used
- So can ICA be used to do the same?



ICA vs PCA bases

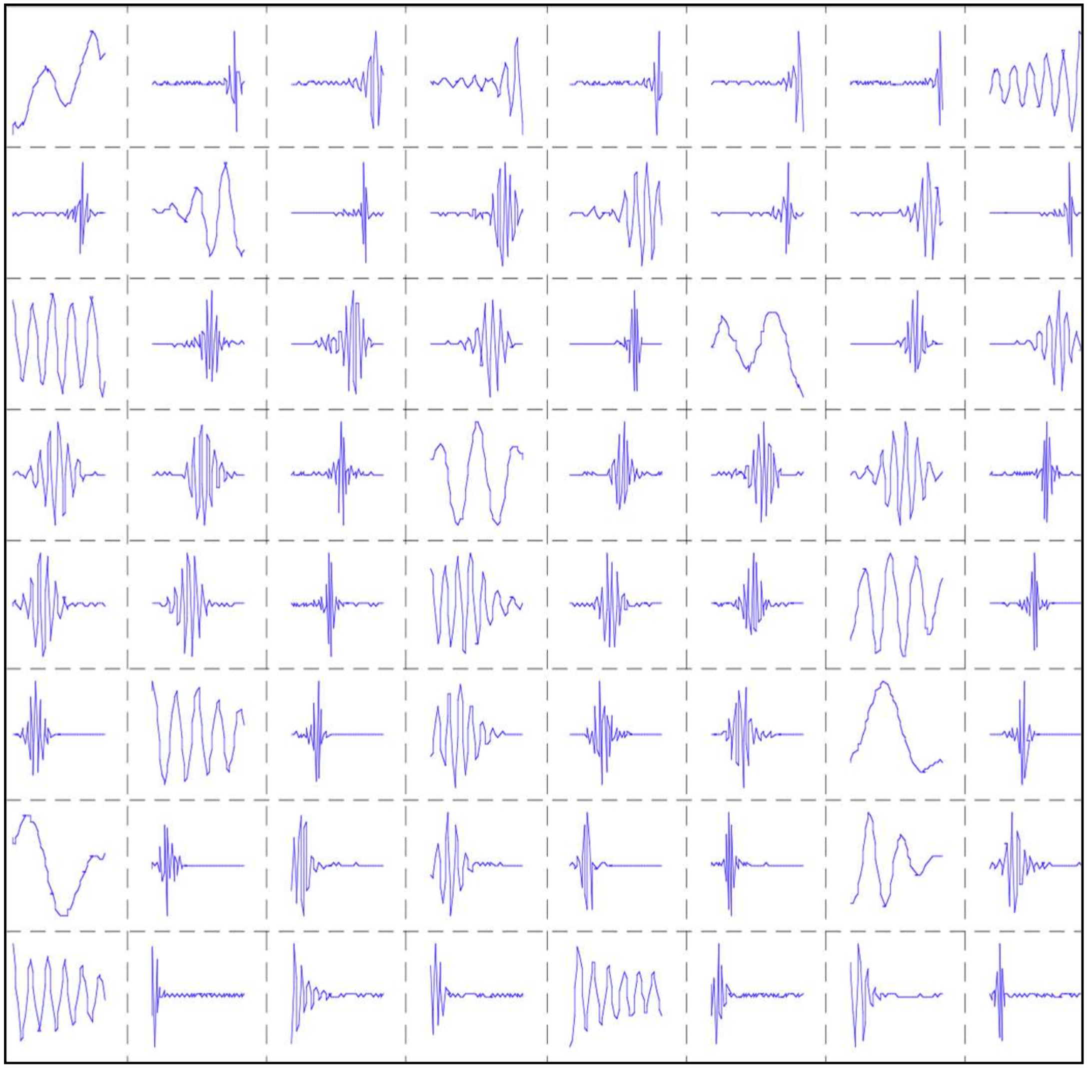
- Motivation for using ICA vs PCA
- PCA will indicate orthogonal directions of maximal variance
  - May not align with the data!
- ICA finds directions that are independent
  - More likely to "align" with the data

Non-Gaussian data

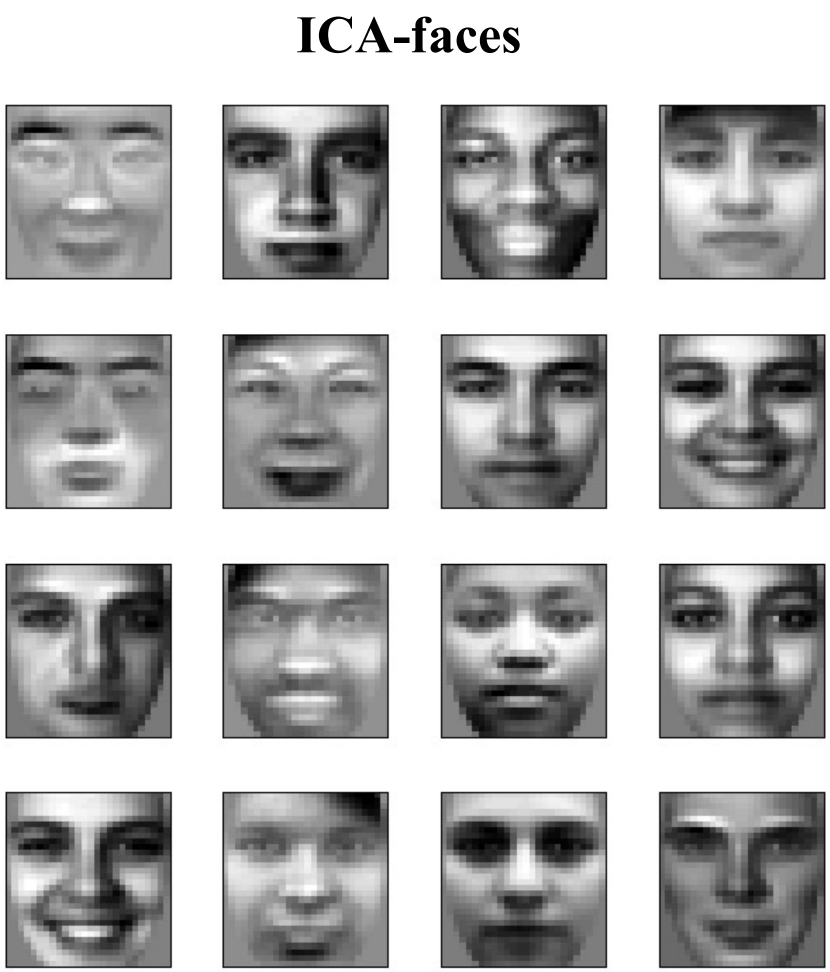


Finding useful transforms with ICA

- Audio preprocessing example
- Take a lot of audio snippets and concatenate them in a big matrix, do component analysis
- PCA results in the DCT bases
- ICA returns time/freq localized sinusoids which is a better way to analyze sounds
- Ditto for images
  - ICA returns localizes edge filters



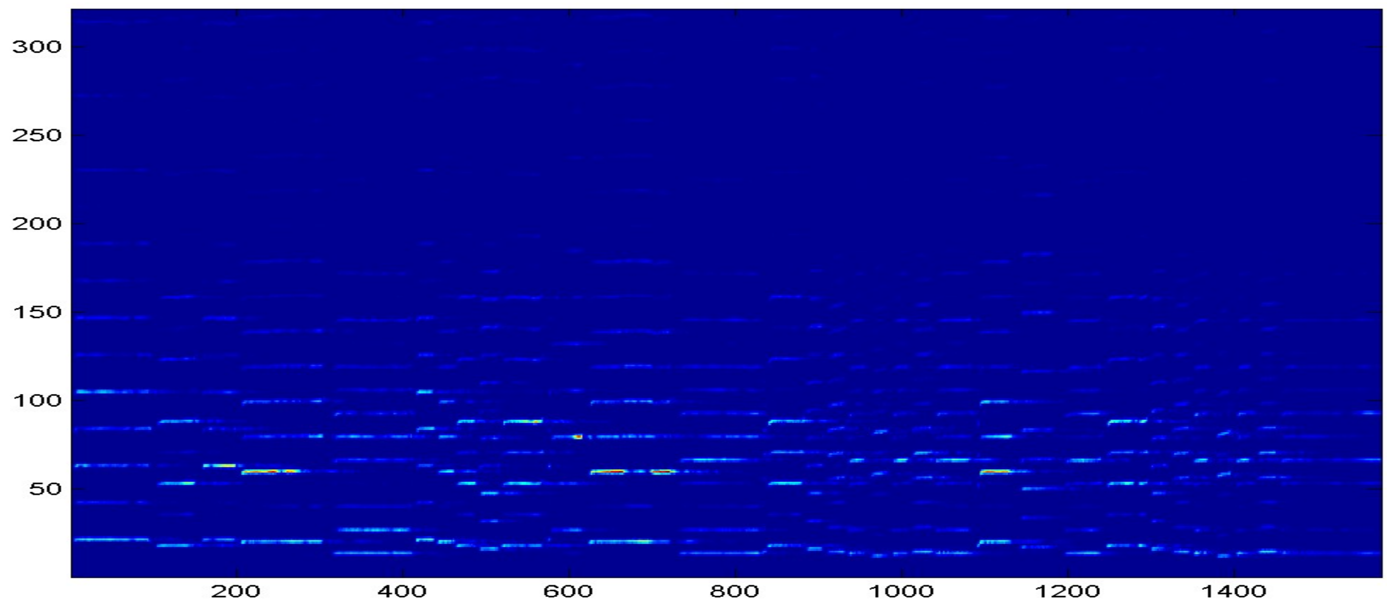
Example case: ICA-faces vs. Eigenfaces



ICA for Signal Enhncement

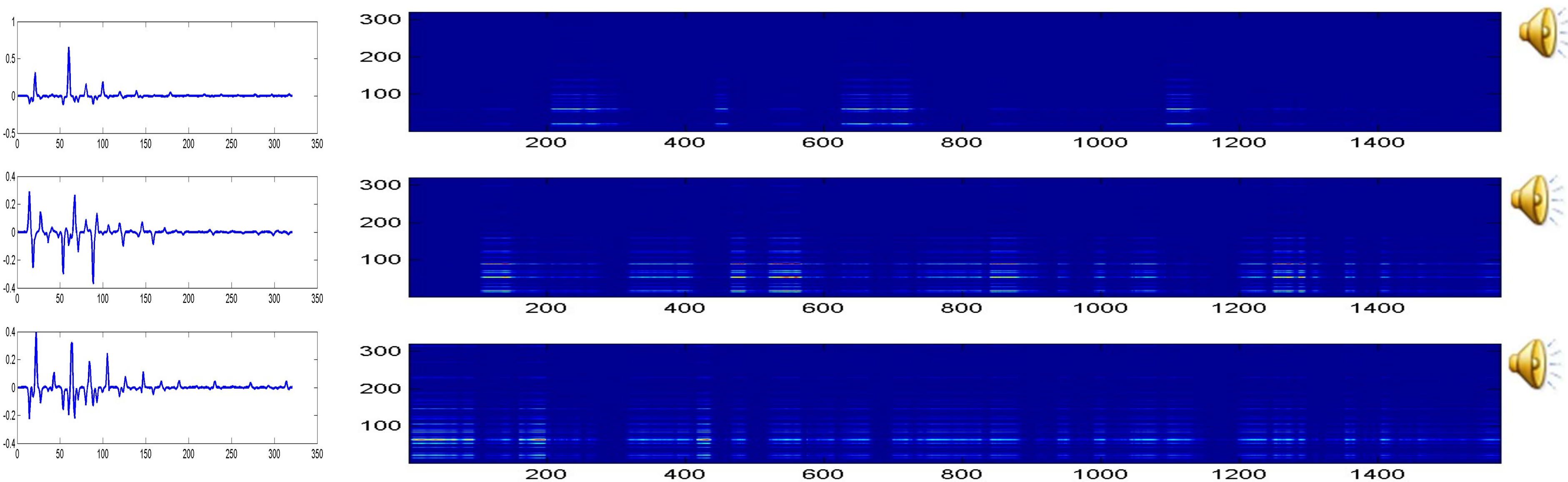
- Very commonly used to enhance EEG signals
- EEG signals are frequently corrupted by heartbeats and biorhythm signals
- ICA can be used to separate them out

So how does that work?



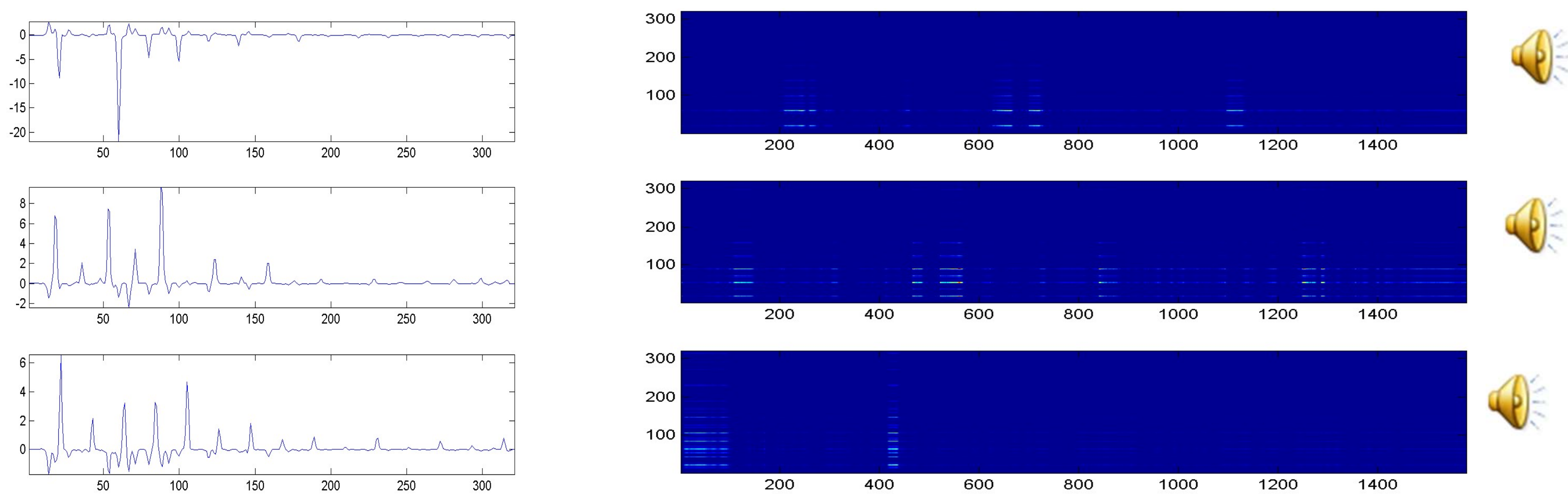
- There are 12 notes in the segment, hence we try to estimate 12 notes..

PCA solution



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

So how does this work: ICA solution



- Better..
  - But not much
- But the issues here?

ICA Issues

- No sense of *order*
  - Unlike PCA
- Get K independent directions, but does not have a notion of the “best” direction
  - So the sources can come in any order
  - *Permutation invariance*
- Does not have sense of *scaling*
  - Scaling the signal does not affect independence
- Outputs are scaled versions of desired signals in permuted order
  - In the best case
  - In worse case, output are not desired signals at all..

What else went wrong?

- *Notes are not independent*
  - Only one note plays at a time
  - If one note plays, other notes are *not* playing

- Will deal with these later in the course..