Machine Learning for Signal Processing

Independent Component Analysis

Instructor: Bhiksha Raj

Revisiting the Covariance Matrix

- **Assuming centered data**

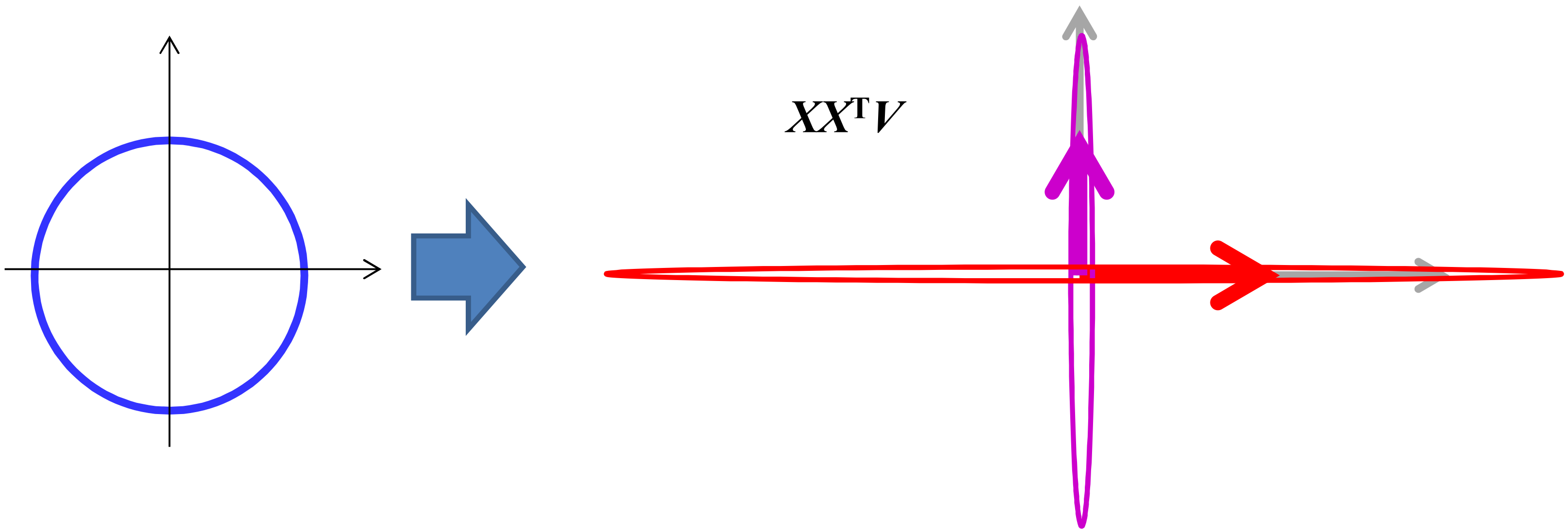
- $C = \Sigma_X XX^T$
- $= X_1X_1^T + X_2X_2^T + ...$

- Let us view C as a transform..

Covariance matrix as a transform

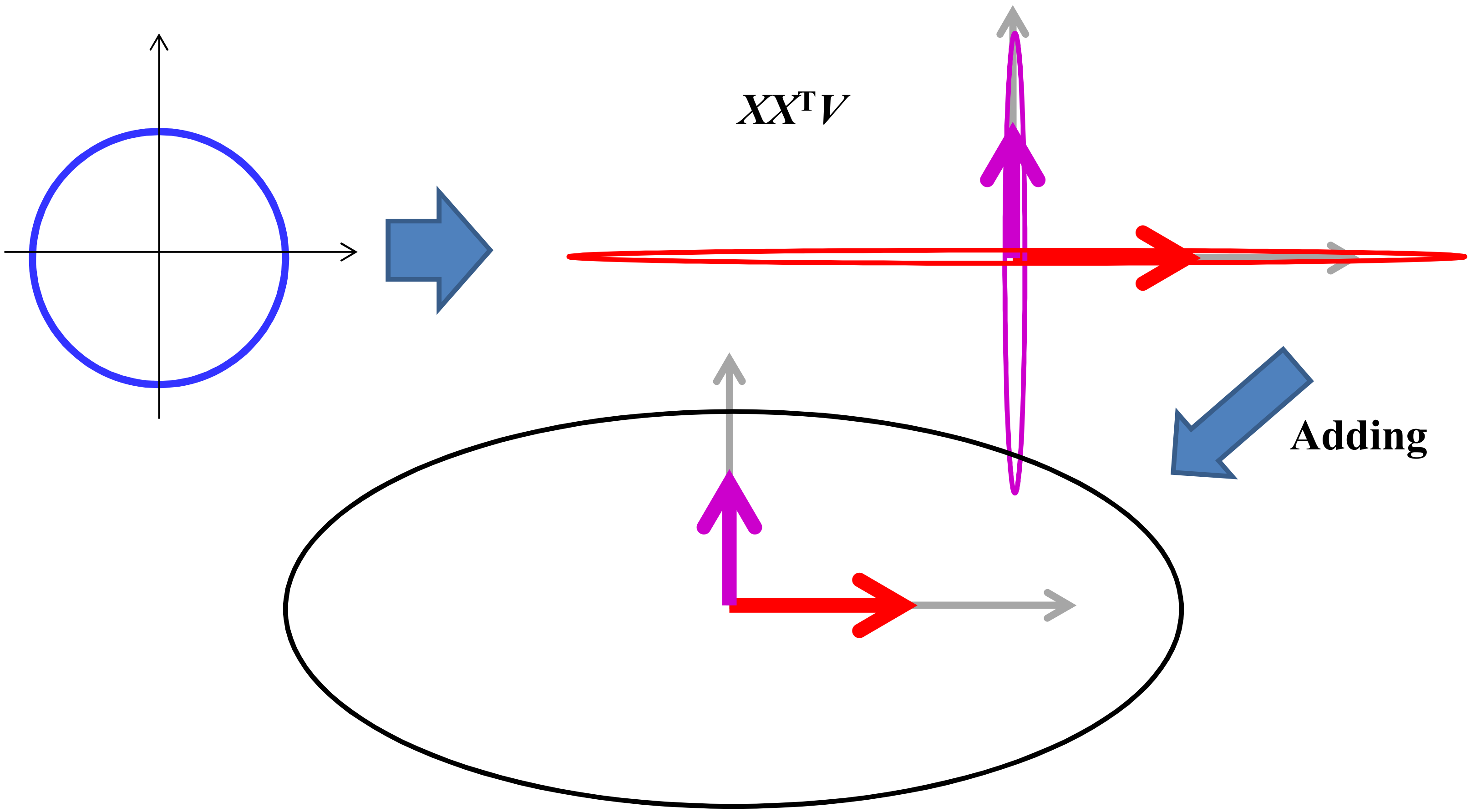
- $(X_1X_1^T + X_2X_2^T + \ldots)\ V = X_1X_1^TV + X_2X_2^TV + \ldots$
- Consider a 2-vector example
  - In two dimensions for illustration

Covariance Matrix as a transform



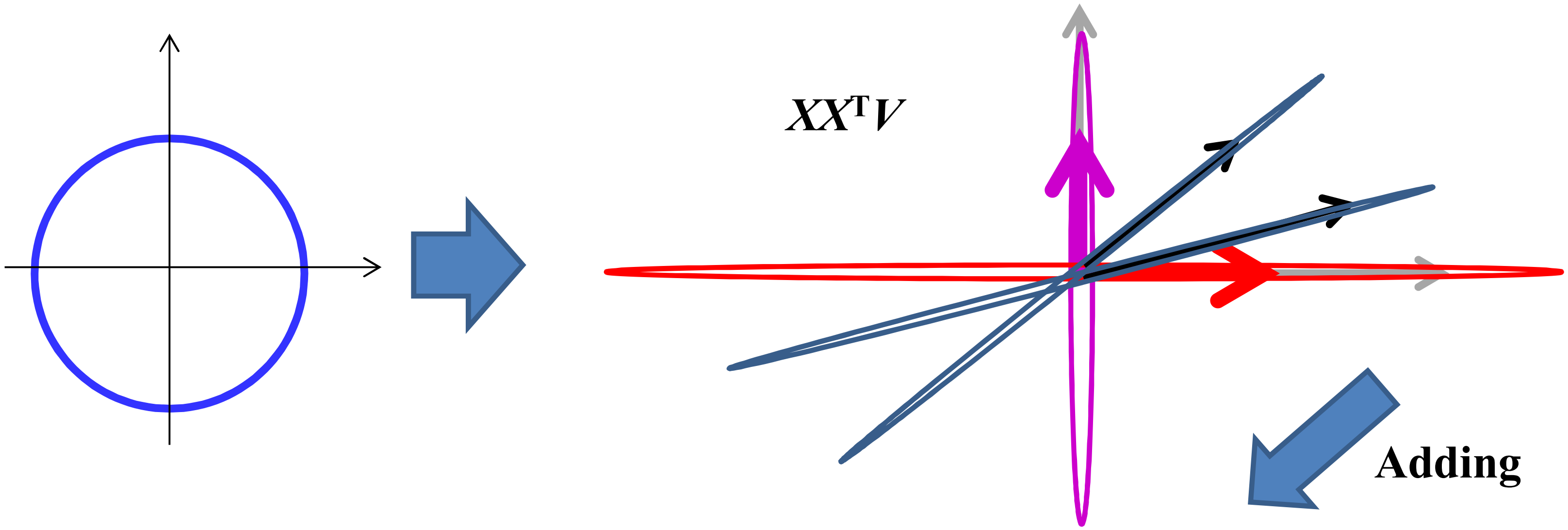
- Data comprises only 2 vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



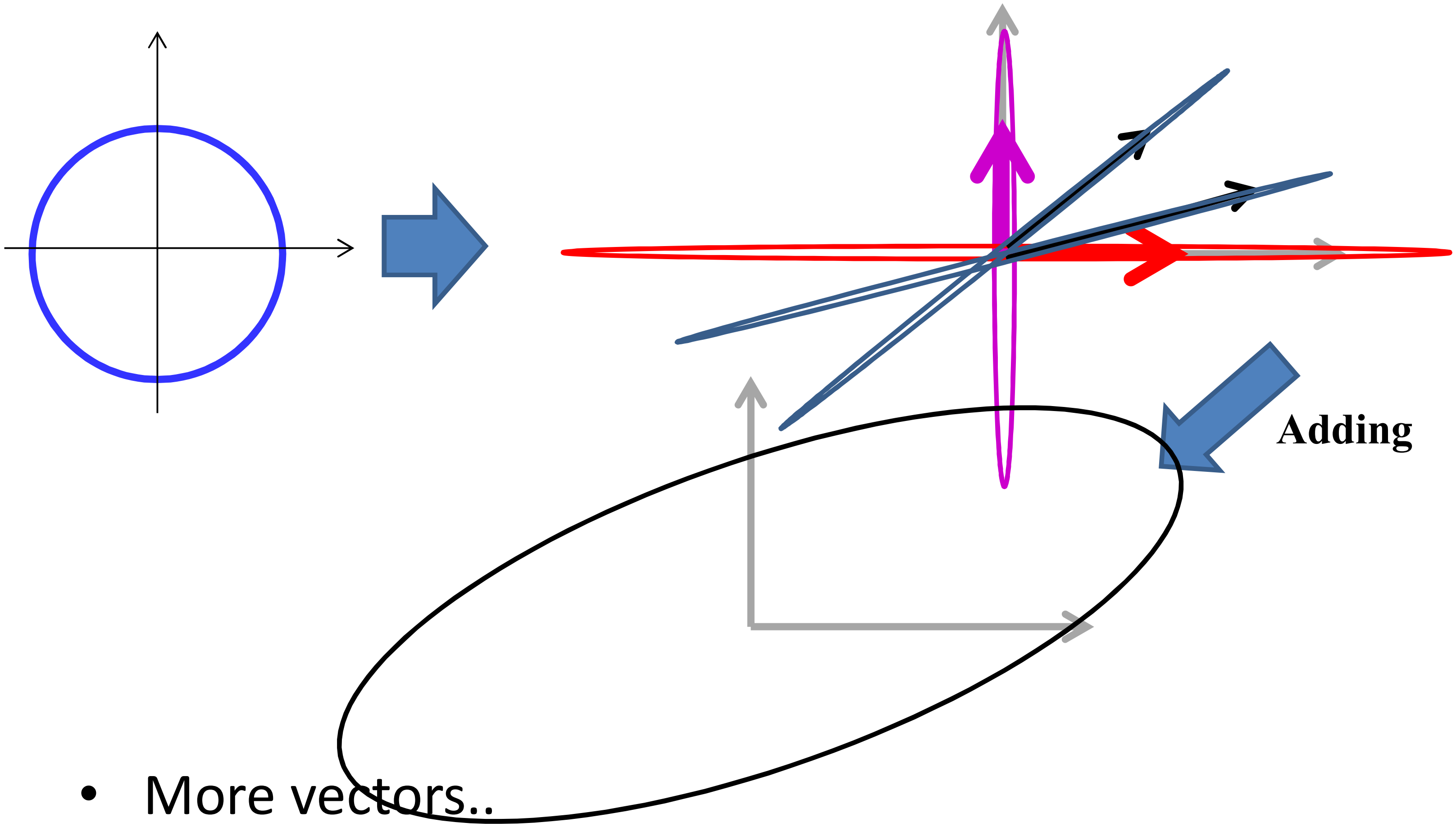
- Data comprises only 2 vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



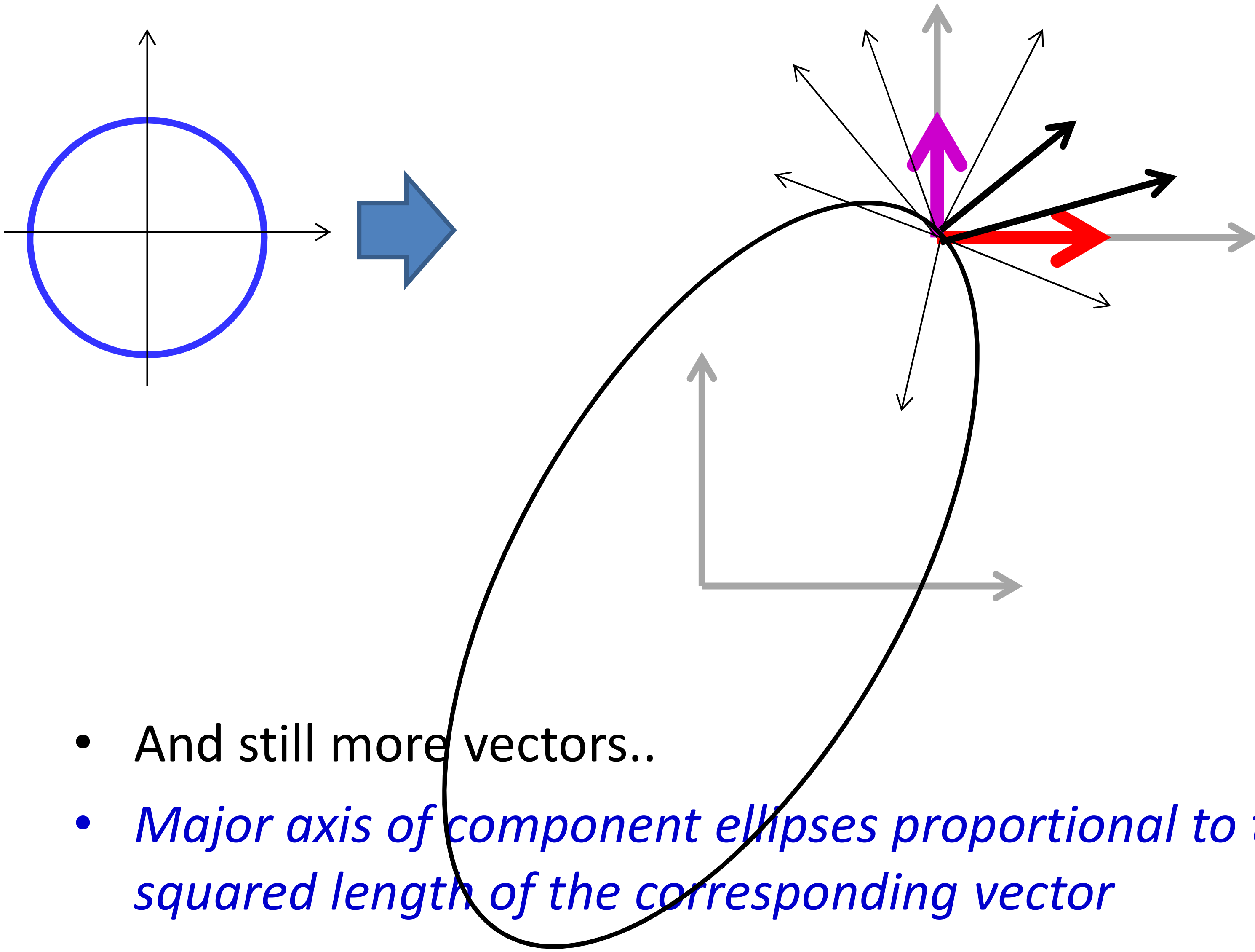
- More vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



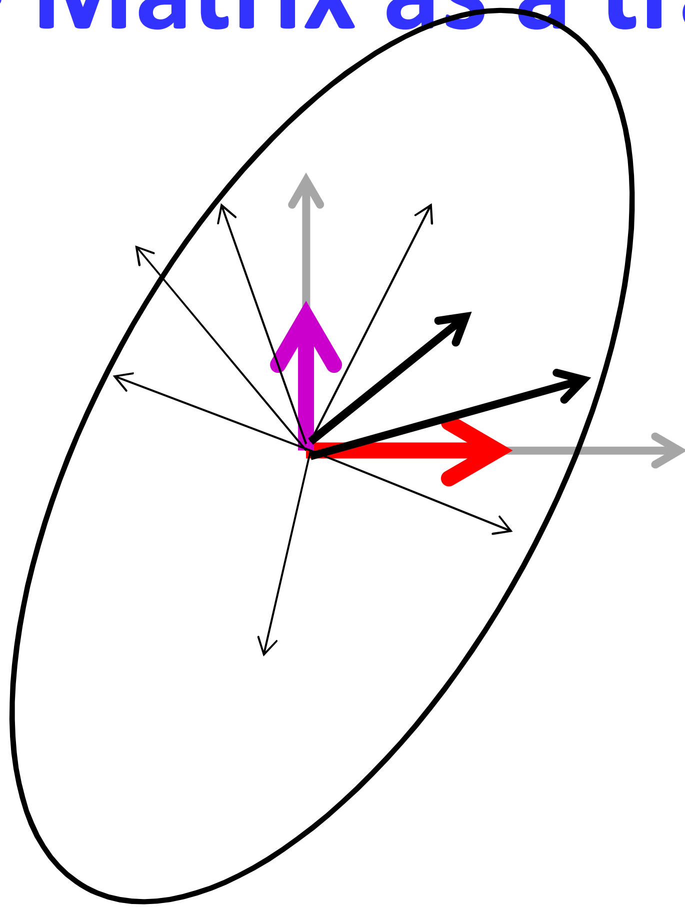
- More vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



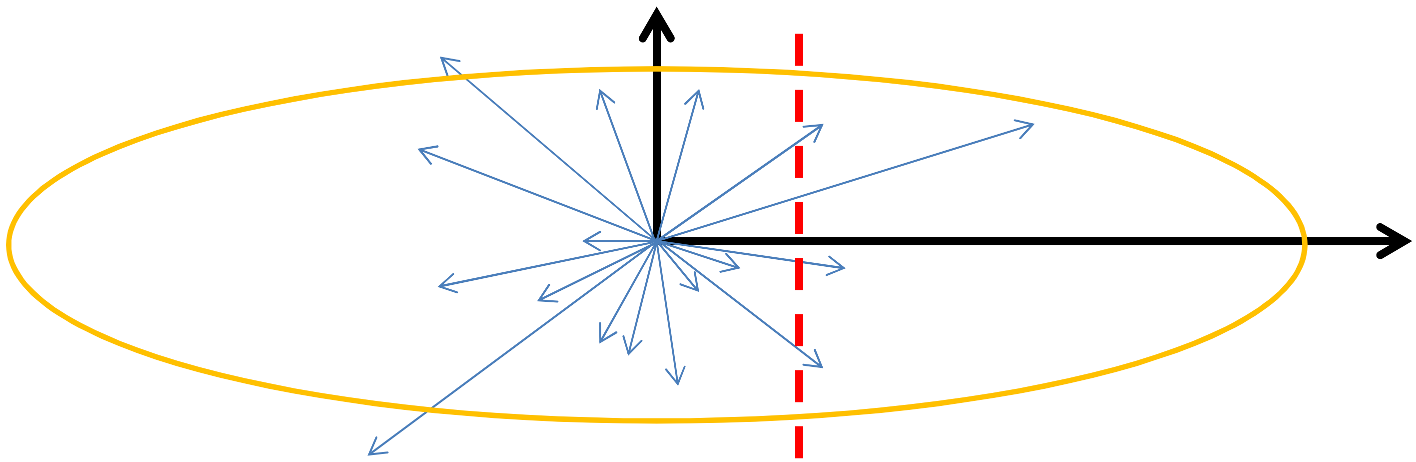
- And still more vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



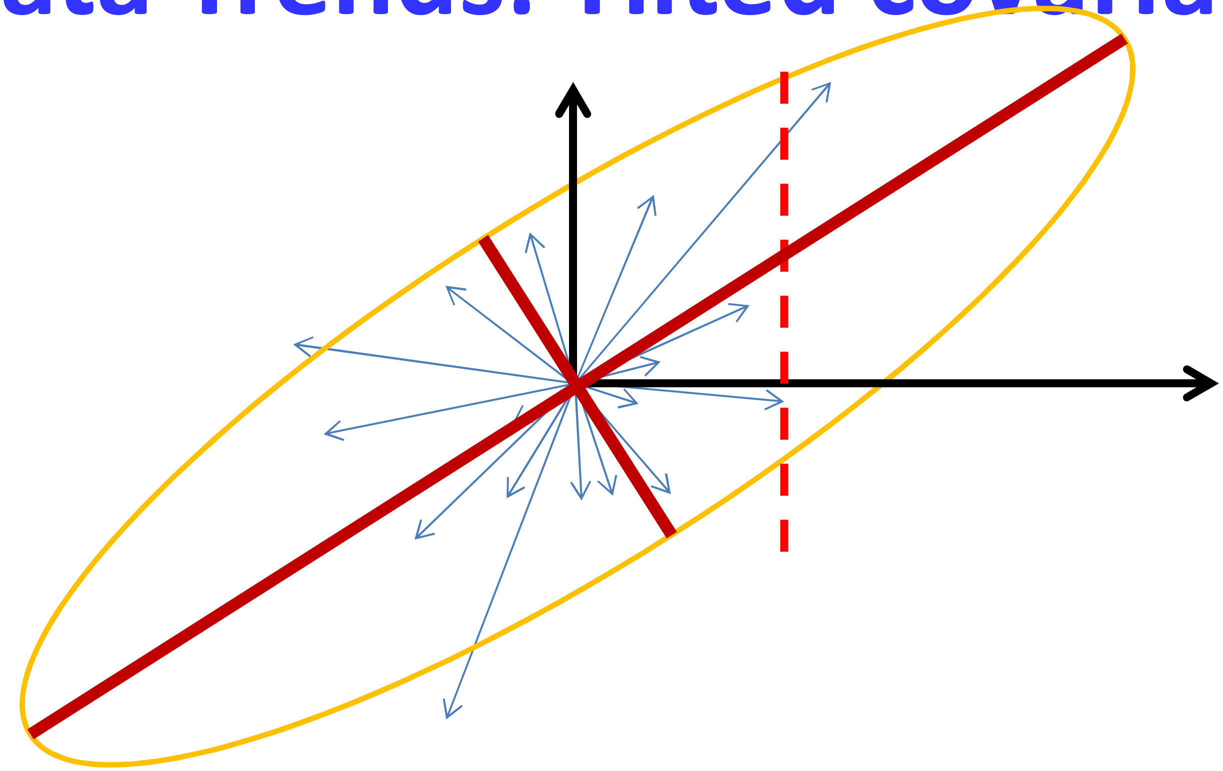
- The covariance matrix captures the directions of maximum variance
- What does it tell us about trends?

Data Trends: Axis aligned covariance



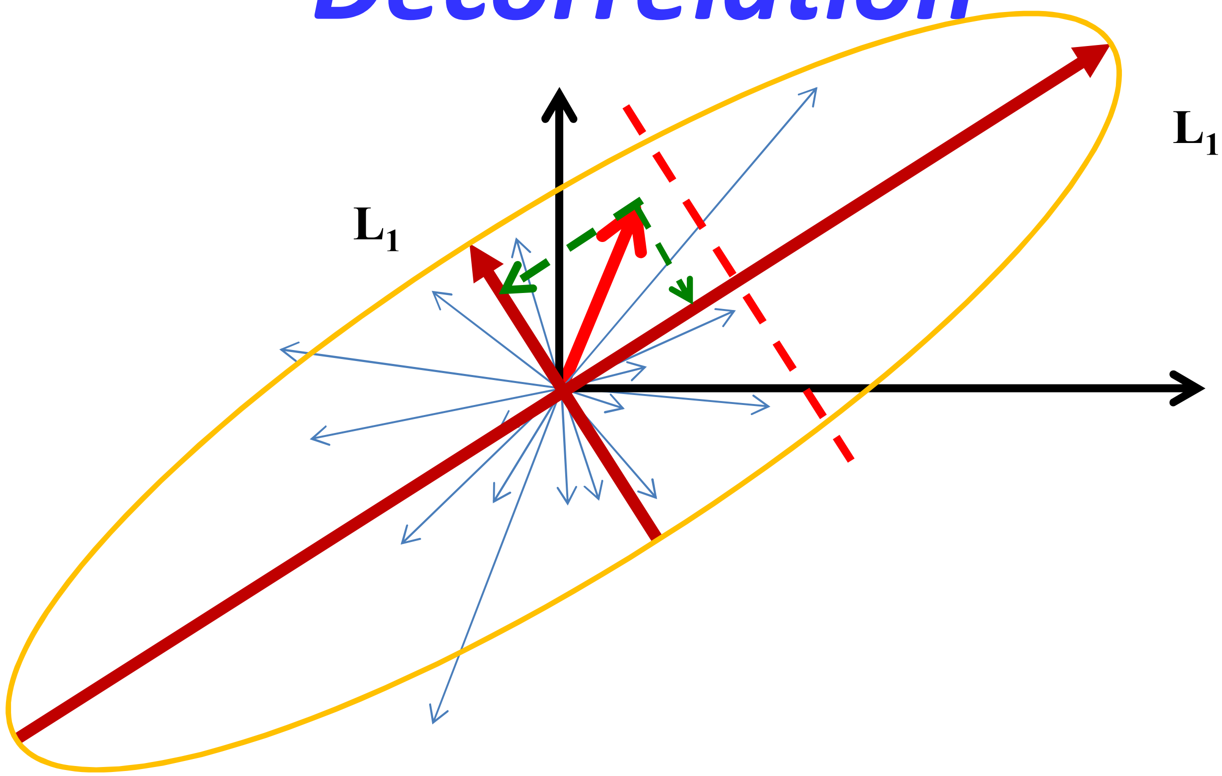
- Axis aligned covariance
- At any X value, the average Y value of vectors is 0
  - X cannot predict Y
- At any Y, the average X of vectors is 0
  - Y cannot predict X
- The X and Y components are ***uncorrelated***

Data Trends: Tilted covariance



- Tilted covariance
- The average Y value of vectors at any X varies with X
  - X predicts Y
- Average X varies with Y
- The X and Y components are ***correlated***

Decorrelation



- Shifting to using the major axes as the coordinate system
  - L_1 does not predict L_2 and vice versa
  - In this coordinate system the data are uncorrelated
- We have ***decorrelated*** the data by rotating the axes

The statistical concept of correlatedness

- Two variables X and Y are correlated if If knowing X gives you an *expected* value of Y

- X and Y are uncorrelated if knowing X tells you nothing about the *expected* value of Y
  - Although it could give you other information
  - How?

Correlation vs. Causation

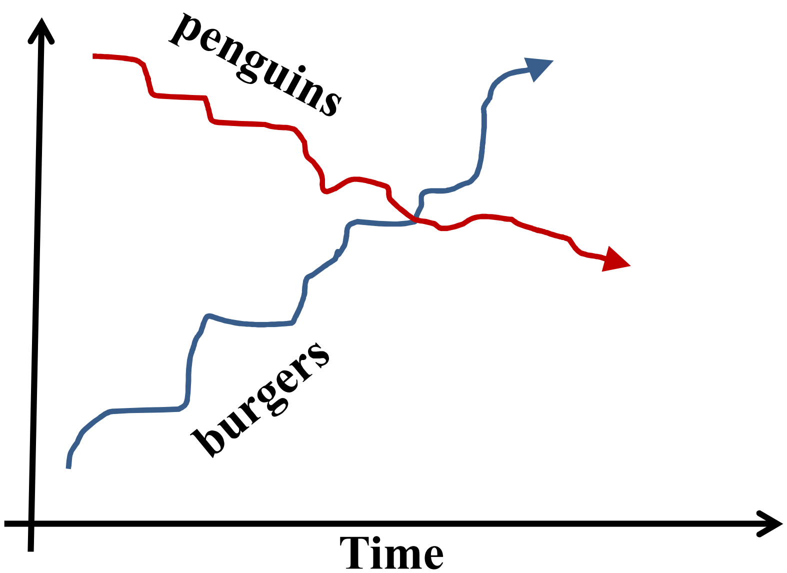
- The consumption of burgers has gone up steadily in the past decade

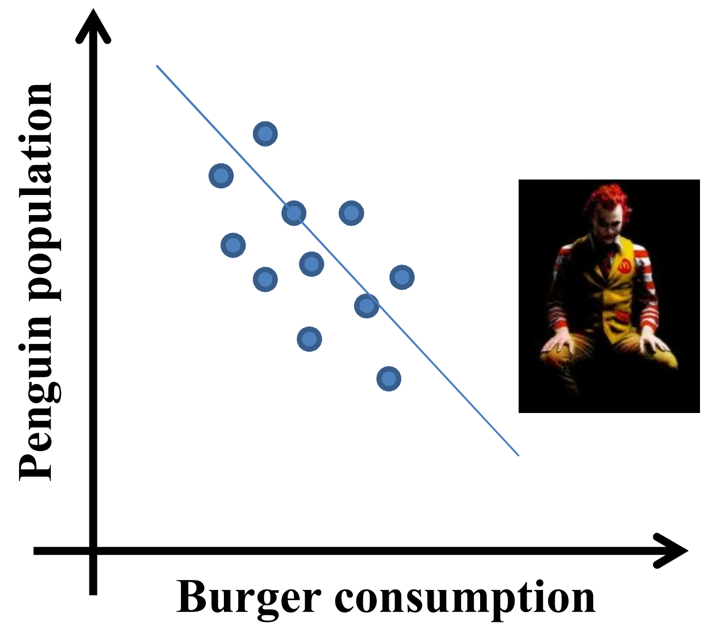
- In the same period, the penguin population of Antarctica has gone down

Correlation, not Causation
(unless McDonalds has a
top-secret Antarctica division)

The concept of *correlation*

- Two variables are correlated if knowing the value of one gives you information about the ***expected value*** of the other





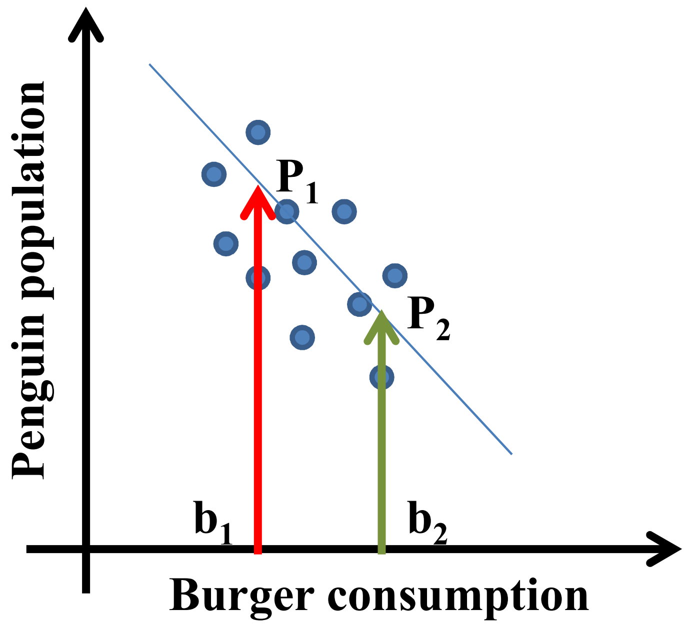
A brief review of basic probability

- *Uncorrelated:* Two random variables X and Y are uncorrelated iff:
  - The *average* value of the product of the variables equals the product of their individual averages

- Setup: Each draw produces one instance of X and one instance of Y
  - I.e one instance of (X,Y)
- E[XY] = E[X]E[Y]

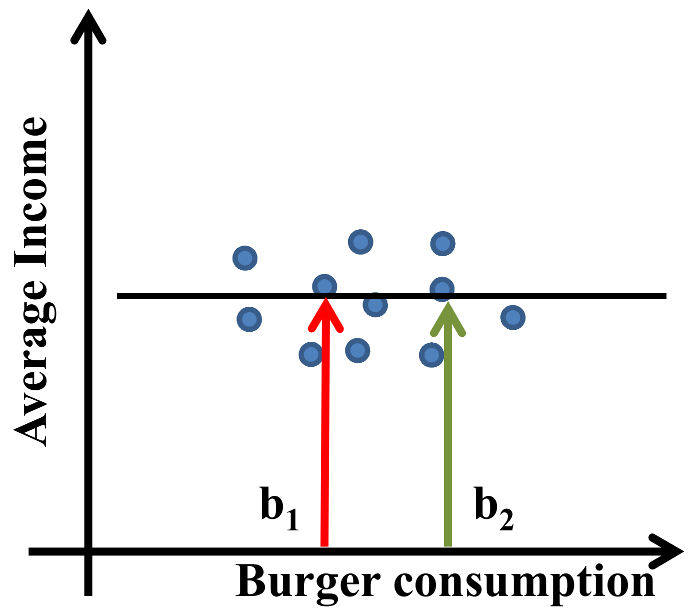
- The average value of Y is the same regardless of the value of X

Correlated Variables



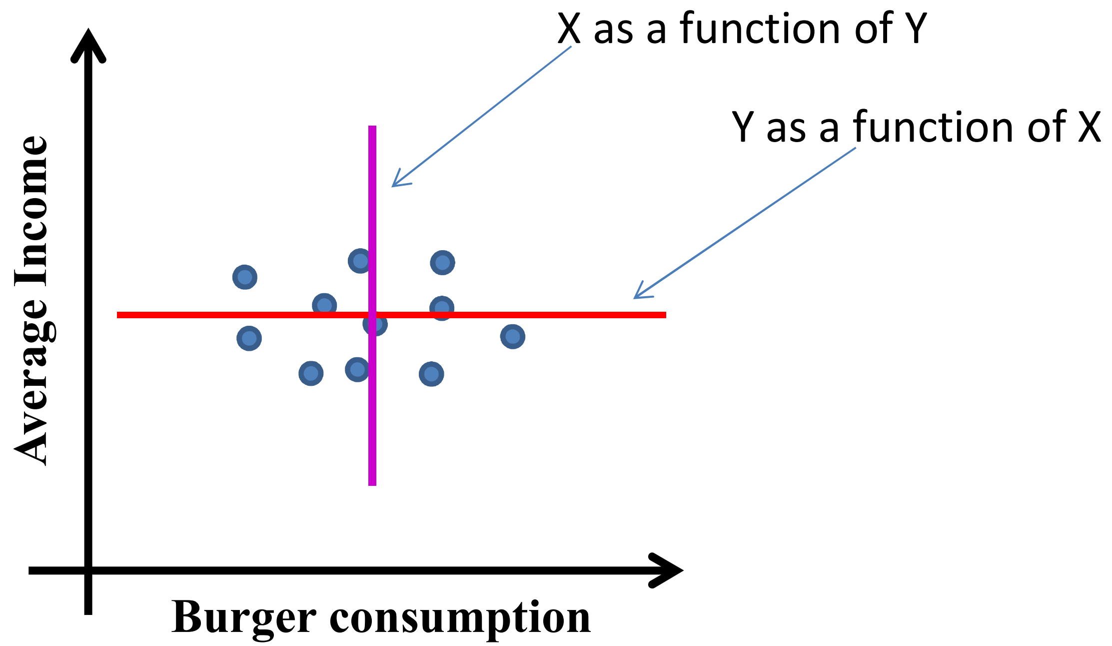
- Expected value of Y given X:
  - Find average of Y values of all samples at (or close) to the given X
  - If this is a function of X, X and Y are correlated

Uncorrelatedness



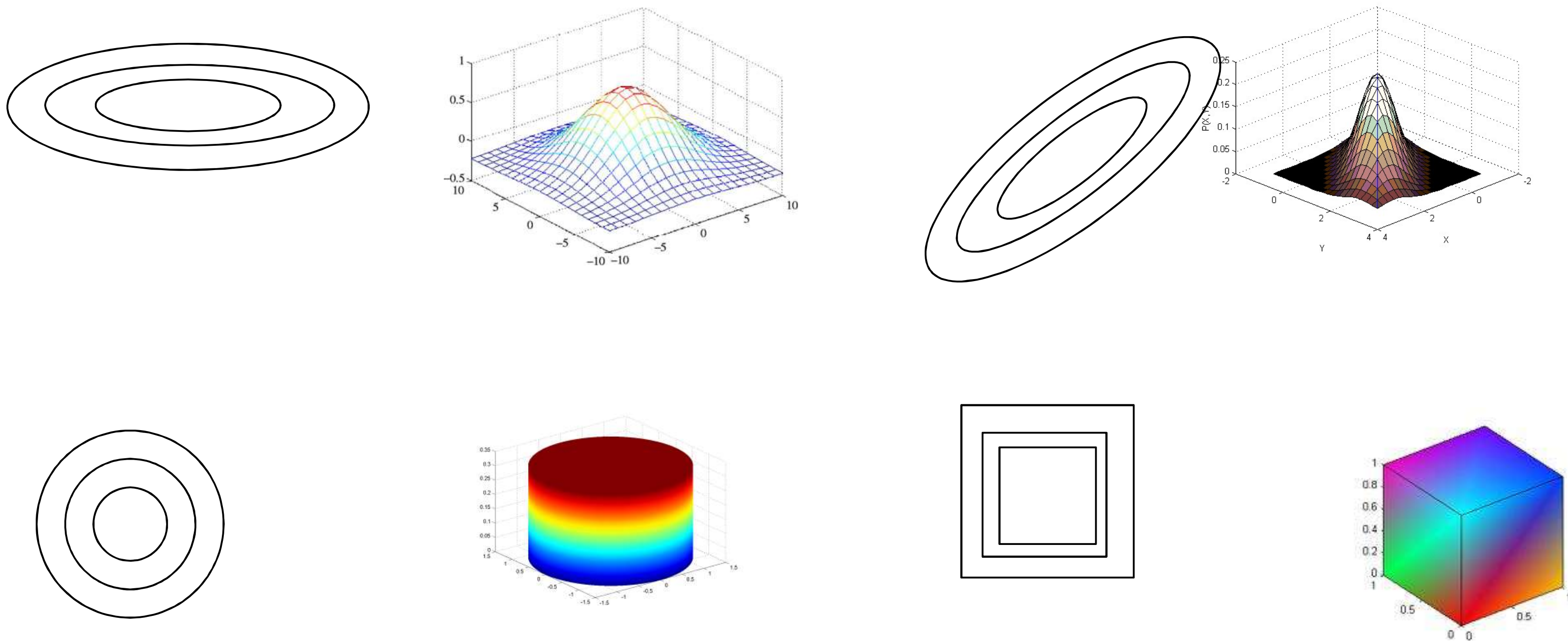
- Knowing X does not tell you what the *average* value of Y is
  – And vice versa

Uncorrelated Variables



- The average value of Y is the same regardless of the value of X and vice versa

Uncorrelatedness in Random Variables



- Which of the above represent uncorrelated RVs?

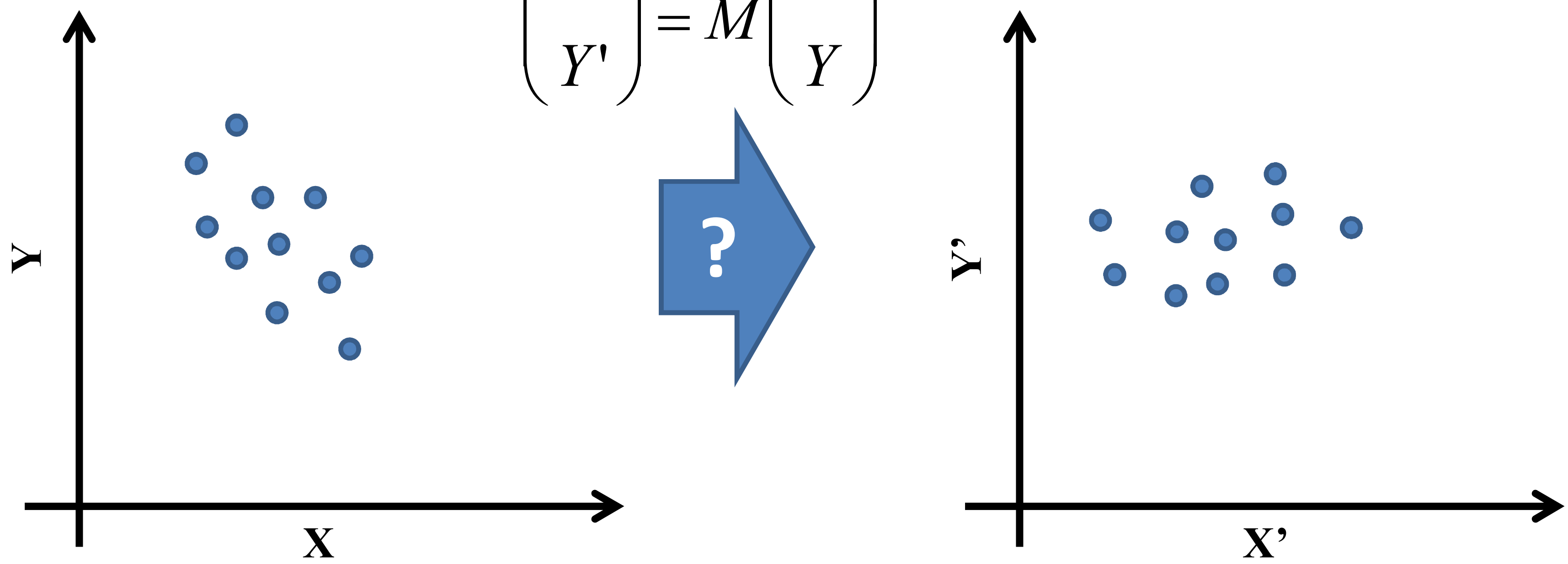
Benefits of uncorrelatedness..

- Uncorrelatedness of variables is generally considered desirable for modelling and analyses
  - For Euclidean error based regression models and probabilistic models, uncorrelated variables can be separately handled
    - Since the value of one doesn't affect the average value of others
    - Greatly reduces the number of model parameters
  - Otherwise their interactions must be considered

- We will frequently transform correlated variables to make them uncorrelated
  - "Decorrelating" variables

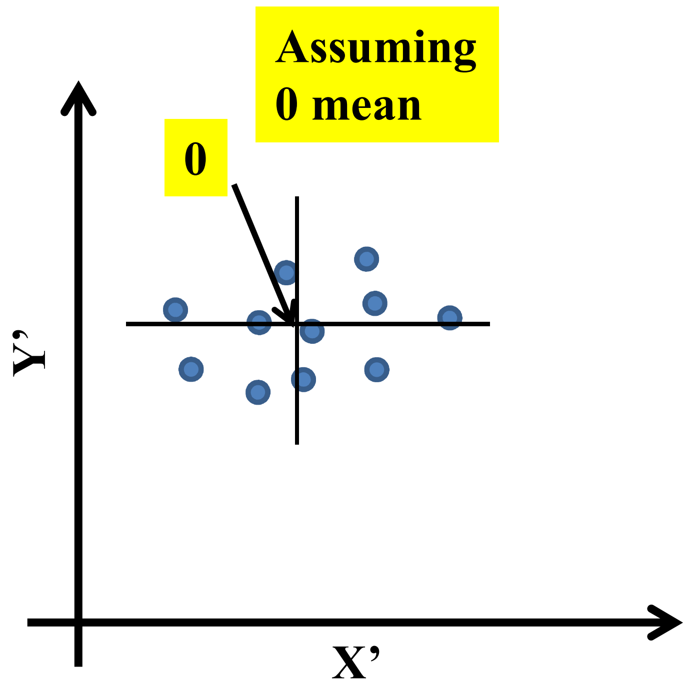
The notion of *decorrelation*

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = M \begin{pmatrix} X \\ Y \end{pmatrix}$$



- So how does one transform the correlated variables (X,Y) to the uncorrelated (X', Y')

What does "uncorrelated" mean



- E[X'] = constant
- E[Y'] = constant
- E[$Y'|X'$] = constant
- E[$X'Y'$] = E[X'] E[Y']
  - All will be 0 for centered data

$$E\left[\begin{pmatrix} X' \\ Y' \end{pmatrix}\begin{pmatrix} X' & Y' \end{pmatrix}\right] = E\begin{pmatrix} X'^2 & X'Y' \\ X'Y' & Y'^2 \end{pmatrix} = \begin{pmatrix} E[X'^2] & 0 \\ 0 & E[Y'^2] \end{pmatrix} = diagonal \quad matrix$$

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

Decorrelation

- 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 $\mathbf{M}$ such that if $\mathbf{Y} = \mathbf{MX}$ such that the covariance of $\mathbf{Y}$ is diagonal
  - $\mathbf{YY}^{\mathrm{T}}$ is the covariance if $\mathbf{Y}$ is zero mean
  - For uncorrelated components, $\mathbf{YY}^{\mathrm{T}} = \mathbf{Diagonal}$

$\Rightarrow \mathbf{MXX}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} = \mathbf{Diagonal}$

$\Rightarrow \mathbf{M}.\mathrm{Cov}(\mathbf{X}).\mathbf{M}^{\mathrm{T}} = \mathbf{Diagonal}$

Decorrelation

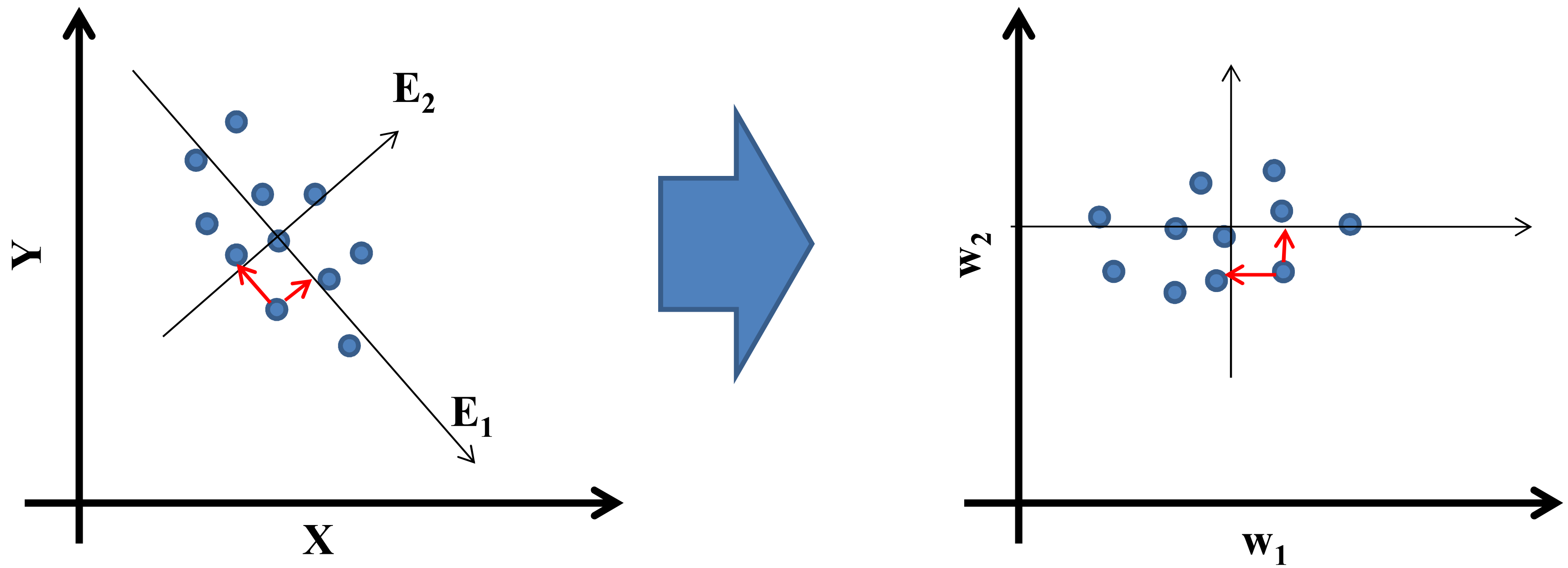
- Easy solution:
  - Eigen decomposition of $\mathrm{Cov}(\mathbf{X})$:

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{E}\Lambda\mathbf{E}^{\mathrm{T}}$$

  - $\mathbf{E}\mathbf{E}^{\mathrm{T}} = \mathrm{I}$
- Let $\mathbf{M} = \mathbf{E}^{\mathrm{T}}$
- $\mathbf{M}\mathrm{Cov}(\mathbf{X})\mathbf{M}^{\mathrm{T}} = \mathbf{E}^{\mathrm{T}}\mathbf{E}\Lambda\mathbf{E}^{\mathrm{T}}\mathbf{E} = \Lambda = \mathrm{diagonal}$
- PCA: $\mathbf{Y} = \mathbf{E}^{\mathrm{T}}\mathbf{X}$
  - Projects the data onto the Eigen vectors of the covariance matrix
  - *Diagonalizes* the covariance matrix
  - "Decorrelates" the data

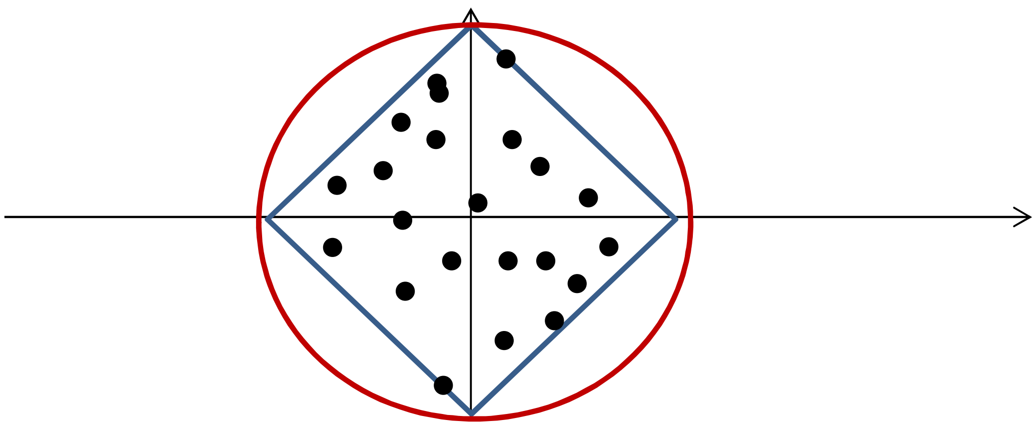
PCA

$$\mathbf{X} = w_1\mathbf{E}_1 + w_2\mathbf{E}_2$$



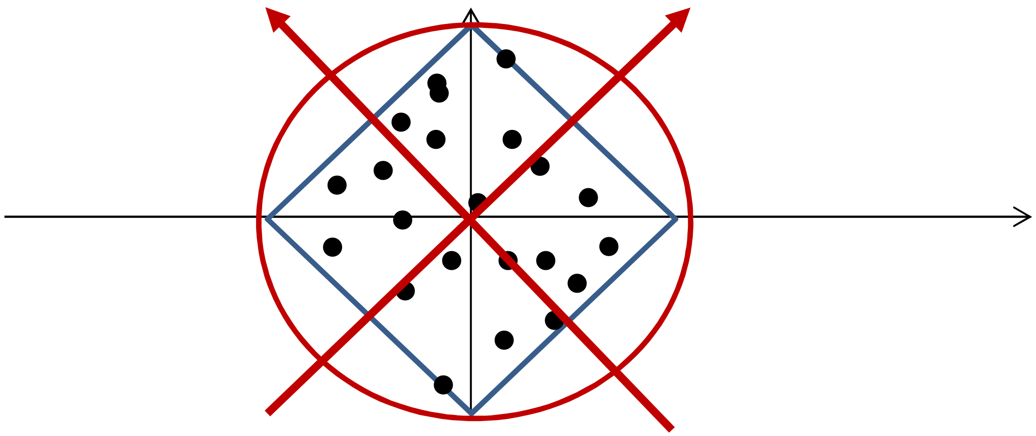
- PCA: $\mathbf{Y} = \mathbf{E}^{\mathrm{T}}\mathbf{X}$
  - Projects the data onto the Eigen vectors of the covariance matrix
    - Changes the coordinate system to the Eigen vectors of the covariance matrix
  - *Diagonalizes* the covariance matrix
  - "Decorrelates" the data

Decorrelating the data



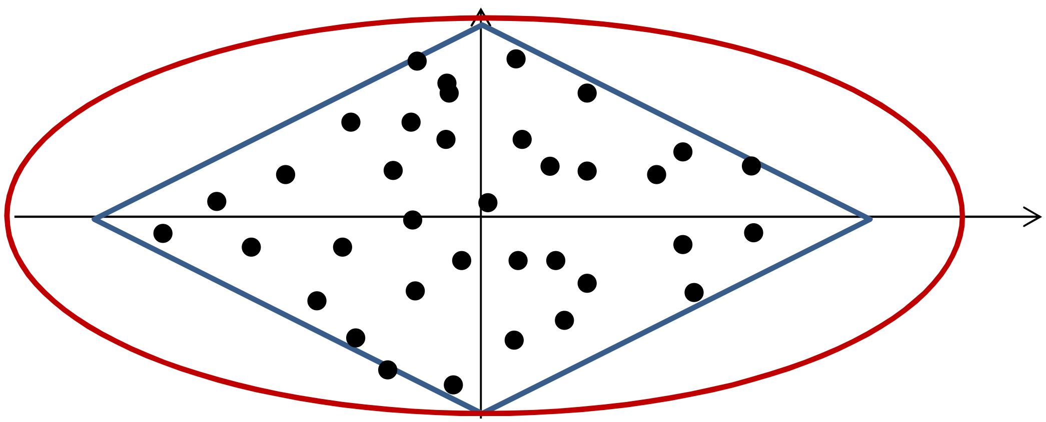
- Are there other decorrelating axes?

Decorrelating the data



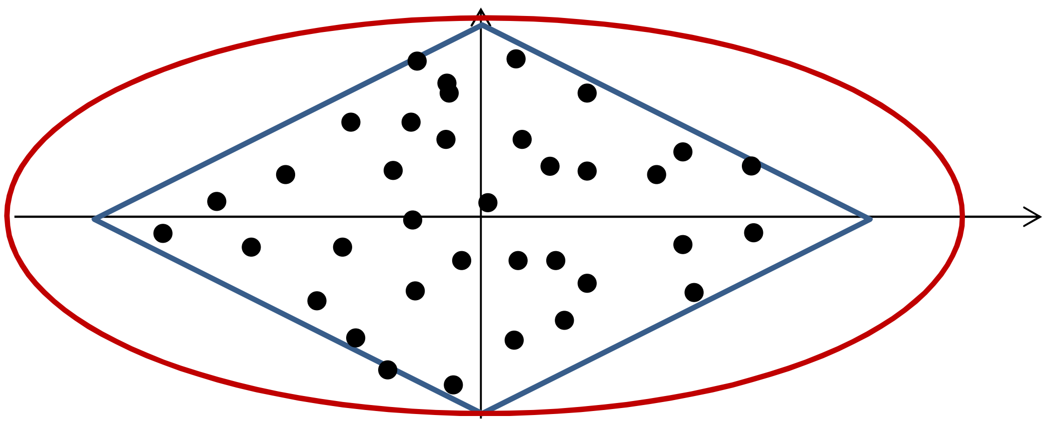
- Are there other decorrelating axes?

Decorrelating the data



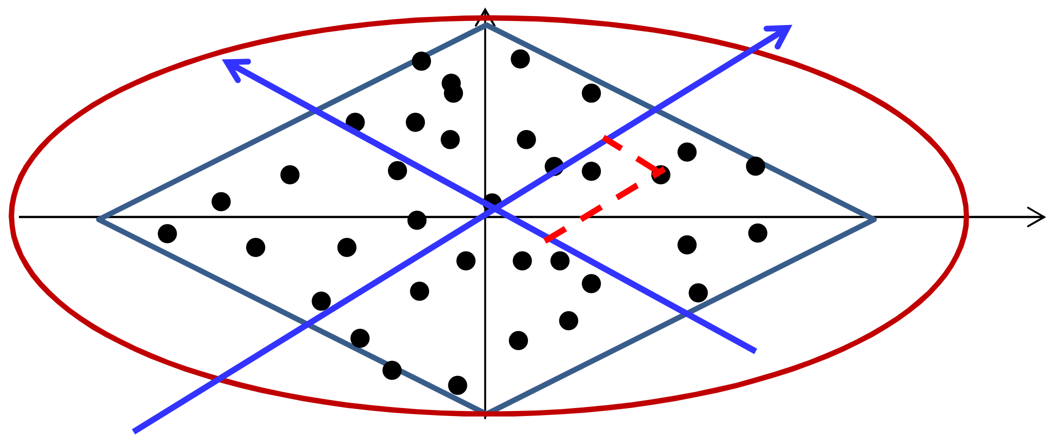
- Are there other decorrelating axes?

Decorrelating the data



- Are there other decorrelating axes?
- What about if we don't require them to be orthogonal?

Decorrelating the data



- Are there other decorrelating axes?
- What about if we don't require them to be orthogonal?
- What is special about these axes?

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

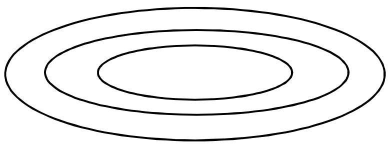
A brief review of basic probability

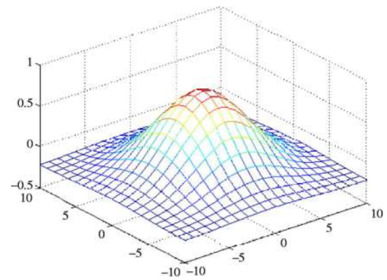
- ***Independence:*** Two random variables X and Y are independent iff:
  - Their joint probability equals the product of their individual probabilities
- $P(X,Y) = P(X)P(Y)$
- Independence implies uncorrelatedness
  - The average value of X is the same regardless of the value of Y
    - $E[X|Y] = E[X]$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

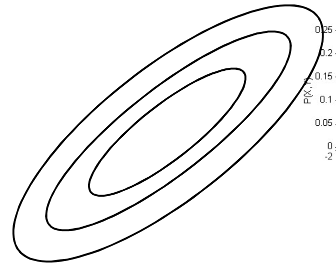
A brief review of basic probability

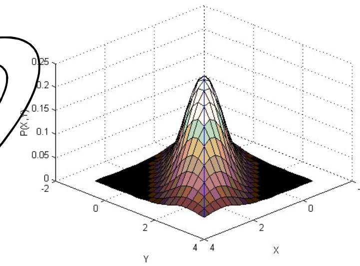
- *Independence:* Two random variables X and Y are independent iff:
- The average value of ***any function*** *of* X is the same regardless of the value of Y
  - Or any function of Y
- **E[f(X)g(Y)] = E[f(X)] E[g(Y)] for all f(), g()**

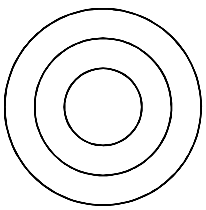
Independence

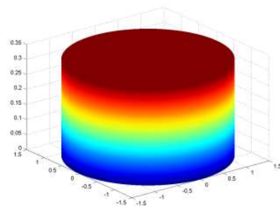


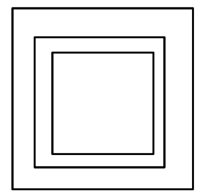


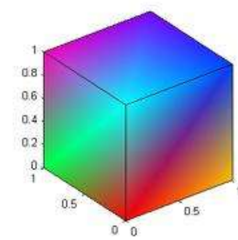






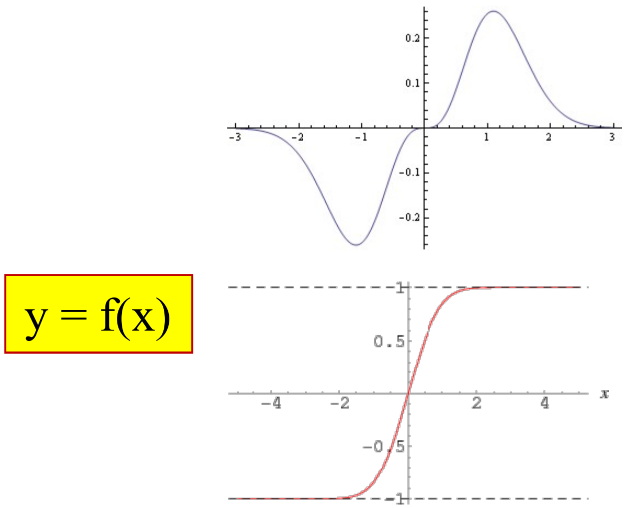


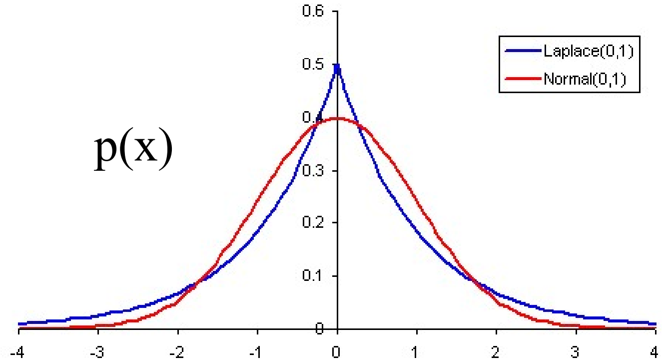




- Which of the above represent independent RVs?
- Which represent uncorrelated RVs?

A brief review of basic probability





- The expected value of an odd function of an RV is 0 if
  - The RV is 0 mean
  - The PDF is of the RV is symmetric around 0
- **E[f(X)] = 0 if f(X) is odd symmetric**

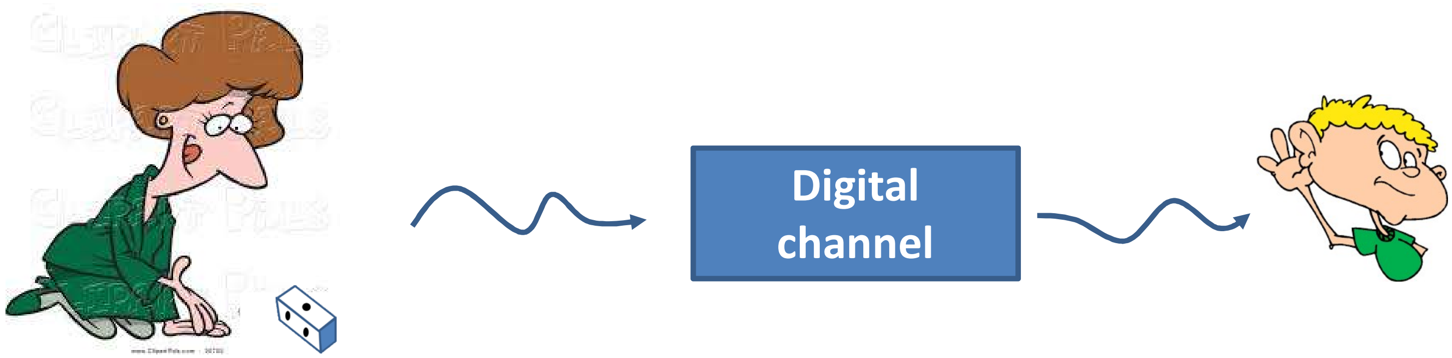
A note on bits..

- You flip a coin. You must inform your friend in the next room about whether the outcome was heads or tails

- How many bits will you have to send?

A note on bits..

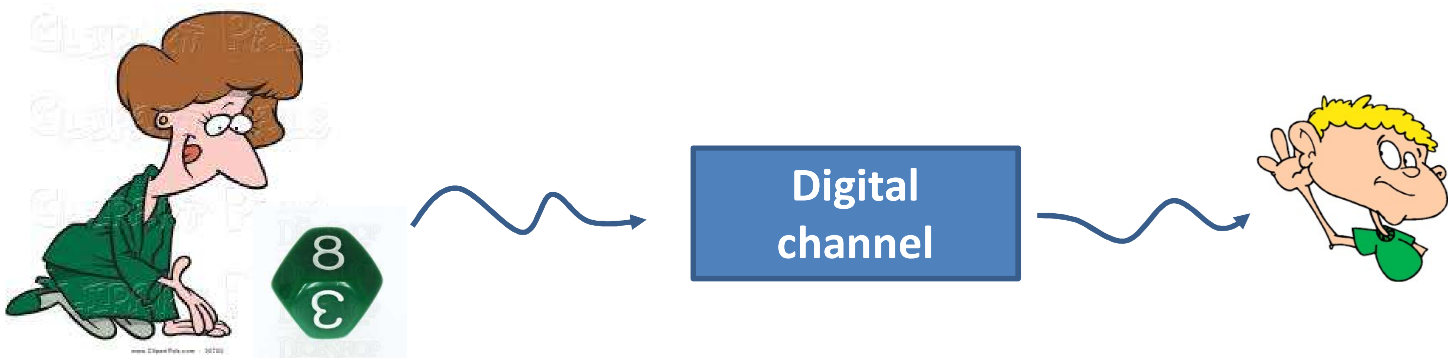
- You roll a four-side dice. You must inform your friend in the next room about the outcome



- How many bits will you have to send?

A note on bits..

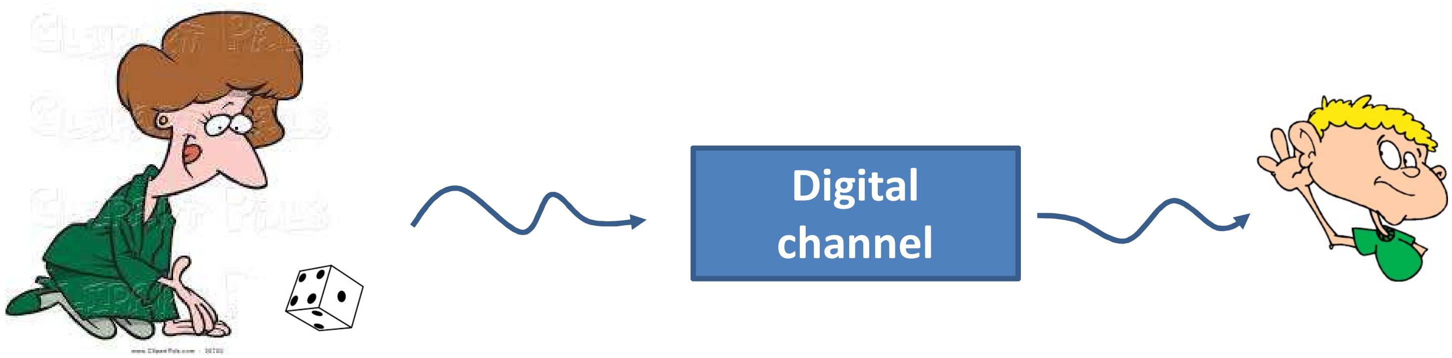
- You roll an *eight-sided polyheldral* dice. You must inform your friend in the next room about the outcome



- How many bits will you have to send?

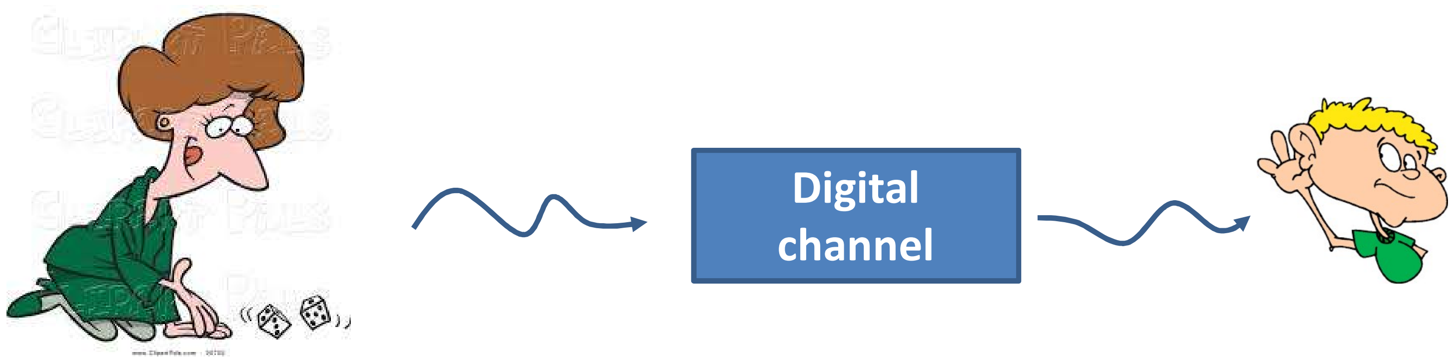
A note on bits..

- You roll a *six-sided* dice. You must inform your friend in the next room about the outcome



- How many bits will you have to send?

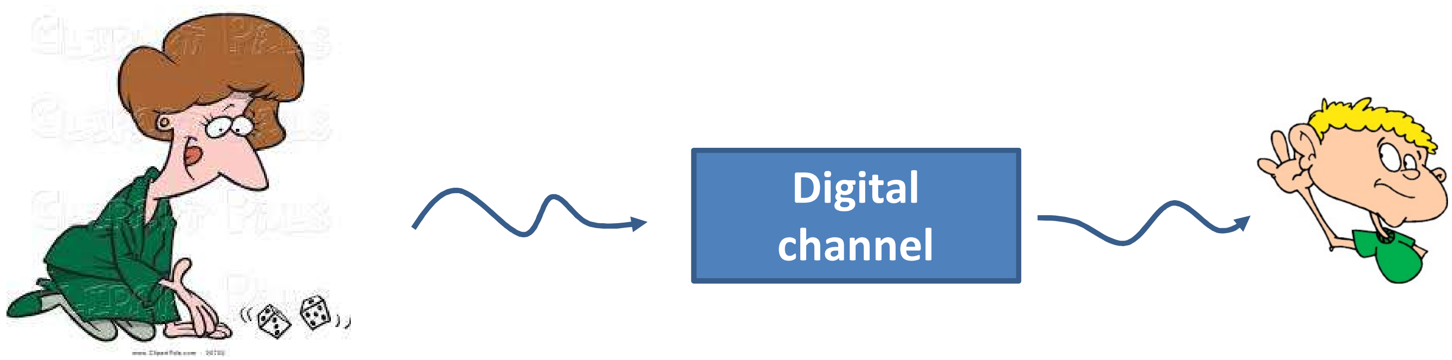
Batching up 6-sided dice rolls



- Instead of sending individual rolls, you roll the dice *twice*
  - And send the *pair* to your friend
- How many bits do you send *per roll?*

1	1
1	2
1	3
..	..
2	1
2	2
..	..
6	6

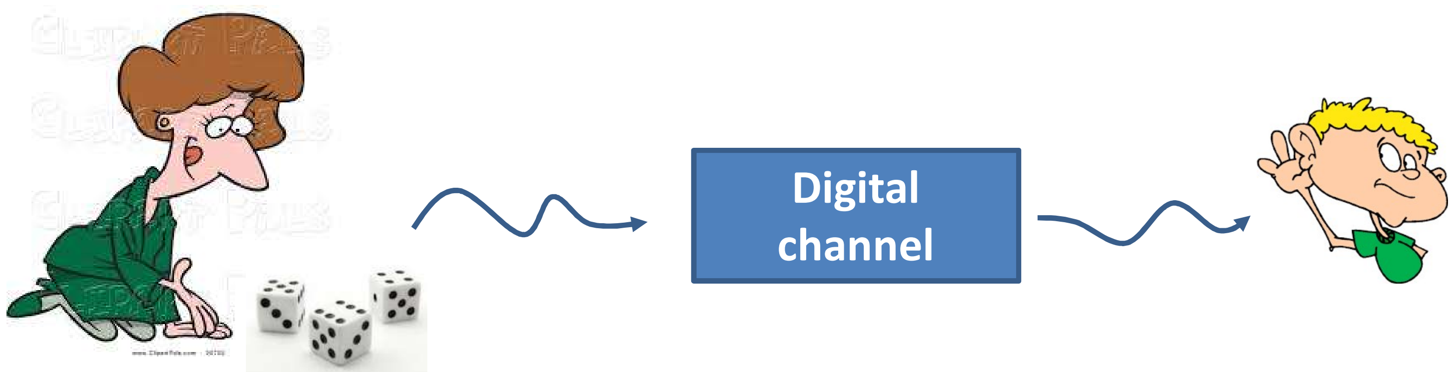
Batching up 6-sided dice rolls



- Instead of sending individual rolls, you roll the dice *twice*
  - And send the *pair* to your friend
- How many bits do you send *per roll?*
- 36 combinations: 6 bits per pair of numbers
  - Still 3 bits per roll

1	1
1	2
1	3
..	..
2	1
2	2
..	..
6	6

Batching up 6-sided dice rolls

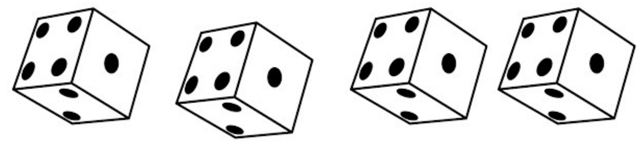


- Instead of sending individual rolls, you roll the dice ***three times***
  - And send the *triple* to your friend
- How many bits do you send *per roll?*
- 216 combinations: 8 bits per triple
  - Still 2.666 bits per roll
  - *Now we're talking!*

1	1	1
1	1	2
..	..	..
1	6	3
..		..
2	1	1
2	1	2
..		..
6	6	6

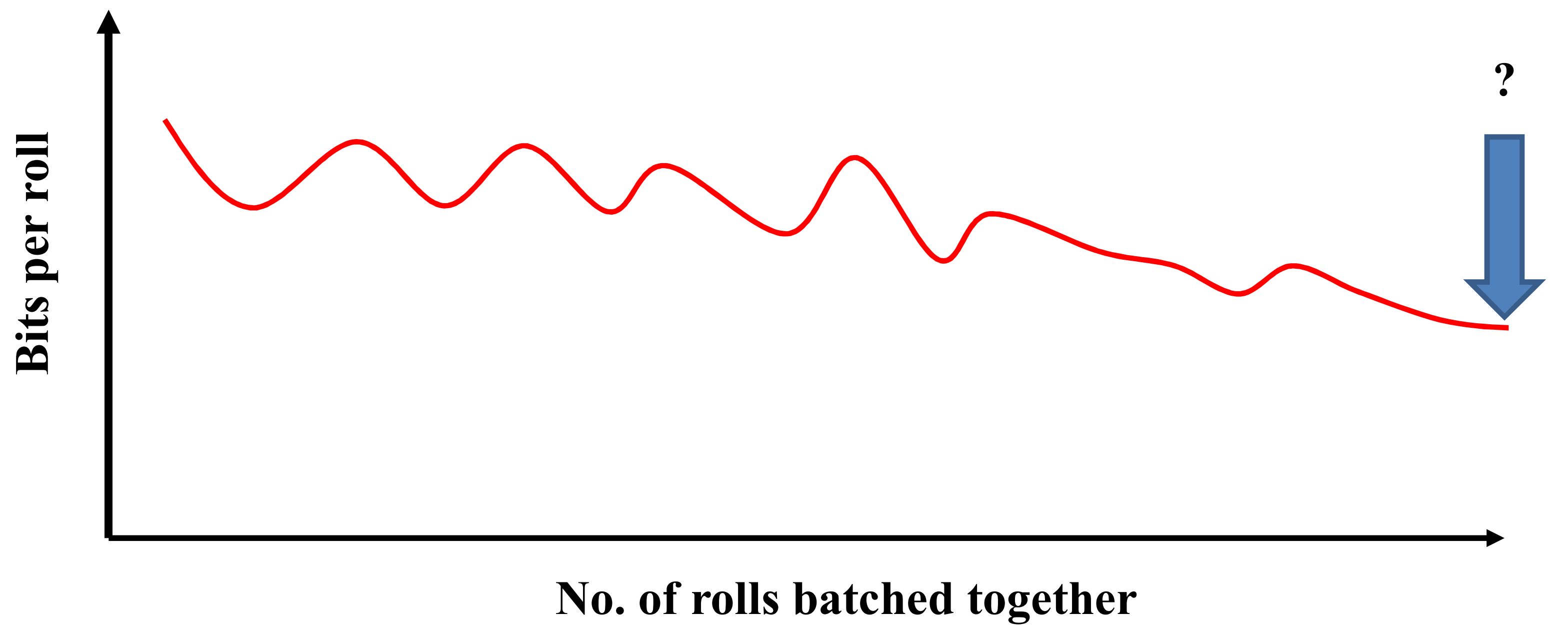
Batching up 6-sided dice rolls

- Batching ***four rolls***
  – 1296 combinations
  – 11 bits per outcome (4 rolls)
  – 2.75 bit per roll



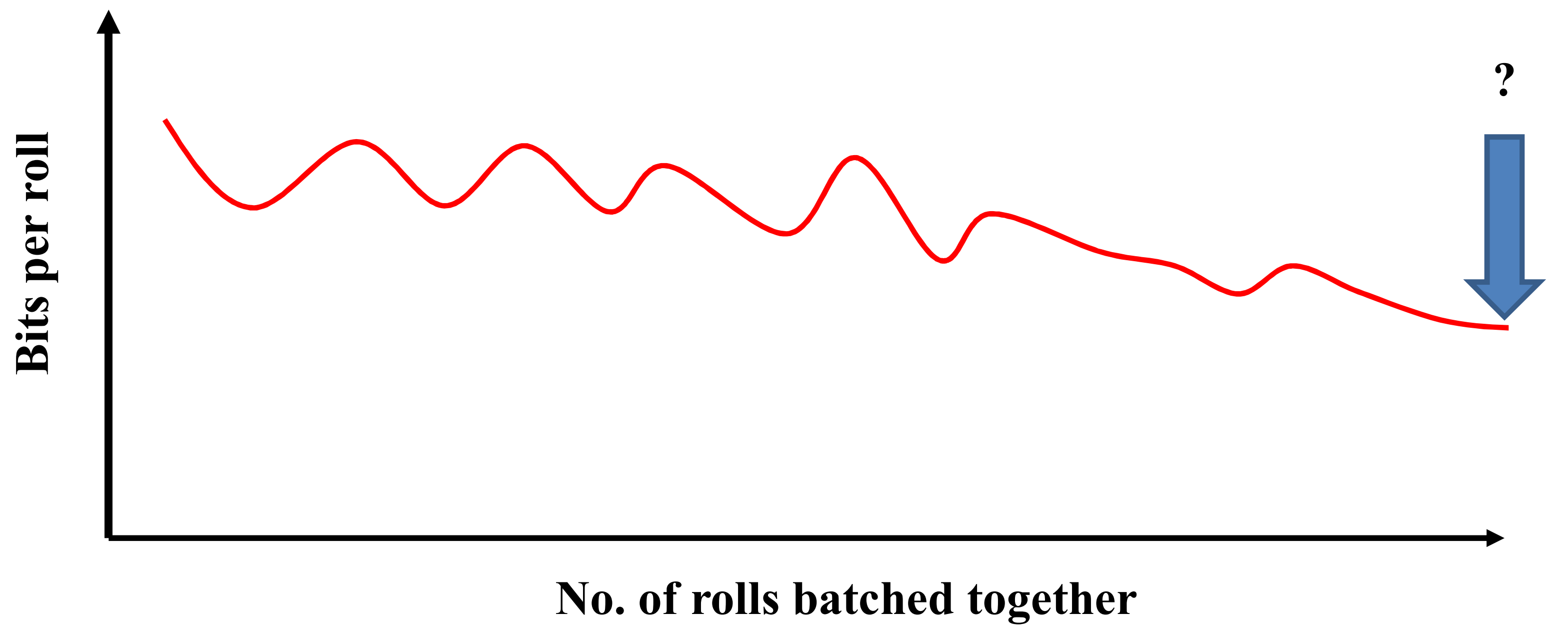
- Batching ***five rolls***
  – 7776 combinations
  – 13 bits per outcome (5 rolls)
  – 2.6 bits per roll

Batching up 6-sided dice rolls



- Where will it end?

Batching up 6-sided dice rolls



- Where will it end?
- $\lim_{k\to\infty}\frac{\lceil k \log2(6) \rceil}{k} = \log2(6)$ bits per roll in the limit
  - This is the absolute minimum – no batching will give you less than these many bits per outcome

Can we do better?

- A four-sided die needs 2 bits per roll
- But then you find not all sides are equally likely

- P(1) = 0.5, P(2) = 0.25, P(3) 0.125, P(4) = 0.125
- *Can you do better than 2 bits per outcome*

Can we do better?

- You have

P(1) = 0.5, P(2) = 0.25, P(3) 0.125, P(4) = 0.125

- You use:

1	0
2	1 0
3	1 1 0
4	1 1 1

  – Note receiver is *never in any doubt as to what they received*

- What is the average number of bits per outcome

Can we do better?

- You have

P(1) = 0.5, P(2) = 0.25, P(3) 0.125, P(4) = 0.125

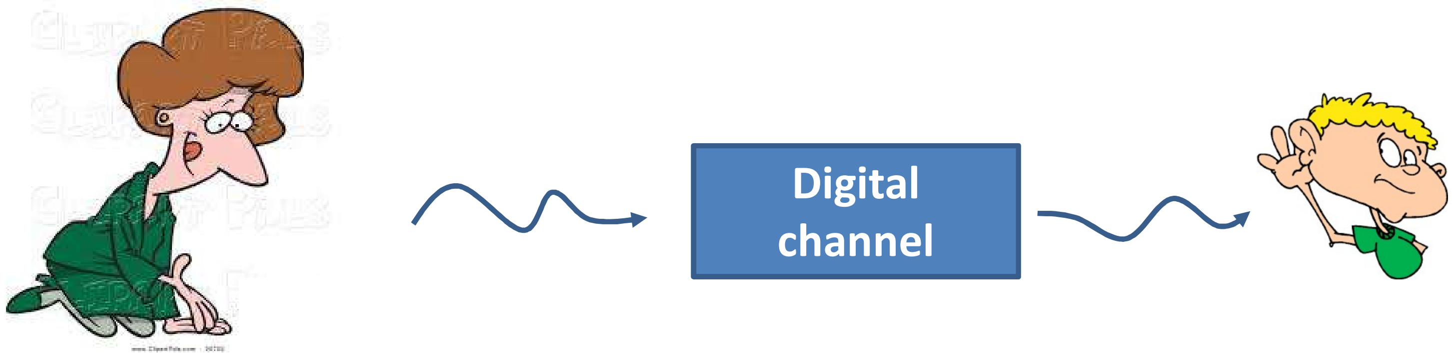
- You use:

1	0
2	1 0
3	1 1 0
4	1 1 1

  – Note receiver is *never in any doubt as to what they received*

- An outcome with probability p is equivalent to obtaining one of $1/p$ equally likely choices
  – Requires $log2(\frac{1}{p})$ bits on average

Entropy



- The average number of bits per symbol required to communicate a random variable over a digitial channel *using an optimal code* is

$$H(p) = \sum_i p_i \log \frac{1}{p_i} = -\sum_i p_i \log p_i$$

- You can't do better
  - Any other code will require more bits
- This is the *entropy of the random variable*

A brief review of basic info. theory

T(all), M(ed), S(hort)…

$$H(X) = \sum_{X} P(X)[-\log P(X)]$$

- Entropy: The *minimum average* number of bits to transmit to convey a symbol

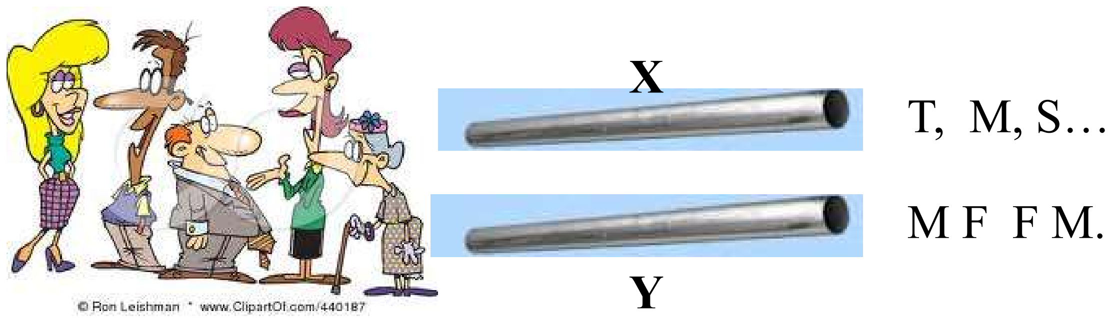
T, M, S…

M F F M..

$$H(X,Y) = \sum_{X,Y} P(X,Y)[-\log P(X,Y)]$$

- Joint entropy: The *minimum average* number of bits to convey sets (pairs here) of symbols

A brief review of basic info. theory



$$H(X \mid Y) = \sum_{Y} P(Y) \sum_{X} P(X \mid Y)[-\log P(X \mid Y)] = \sum_{X,Y} P(X,Y)[-\log P(X \mid Y)]$$

- Conditional Entropy: The *minimum average* number of bits to transmit to convey a symbol X, after symbol Y has already been conveyed
  - Averaged over all values of X and Y

A brief review of basic info. theory

- Conditional entropy of $X|Y = H(X)$ if X is independent of Y

$$H(X|Y) = \sum_Y P(Y) \sum_X P(X|Y)[-\log P(X|Y)] = \sum_Y P(Y) \sum_X P(X)[-\log P(X)] = H(X)$$

- Joint entropy of X and Y is the sum of the entropies of X and Y if they are independent

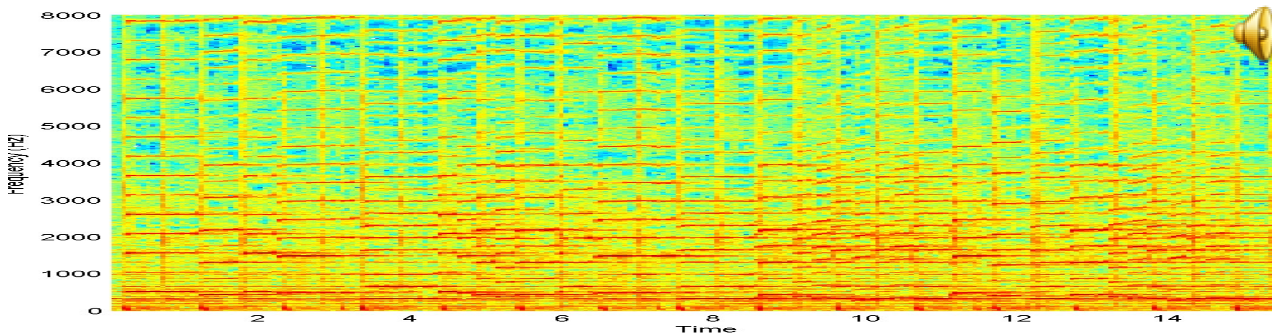
$$H(X,Y) = \sum_{X,Y} P(X,Y)[-\log P(X,Y)] = \sum_{X,Y} P(X,Y)[-\log P(X)P(Y)]$$

$$= -\sum_{X,Y} P(X,Y)\log P(X) - \sum_{X,Y} P(X,Y)\log P(Y) = H(X) + H(Y)$$

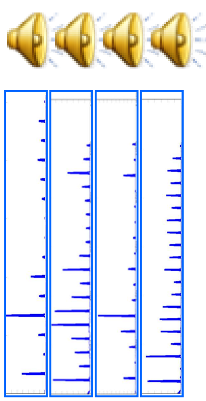
Onward..

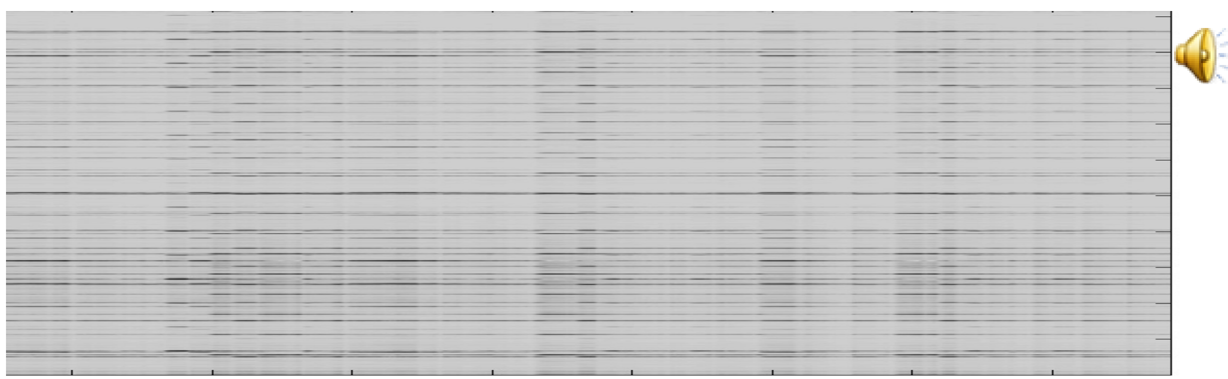
Projection: multiple notes

M =



W =

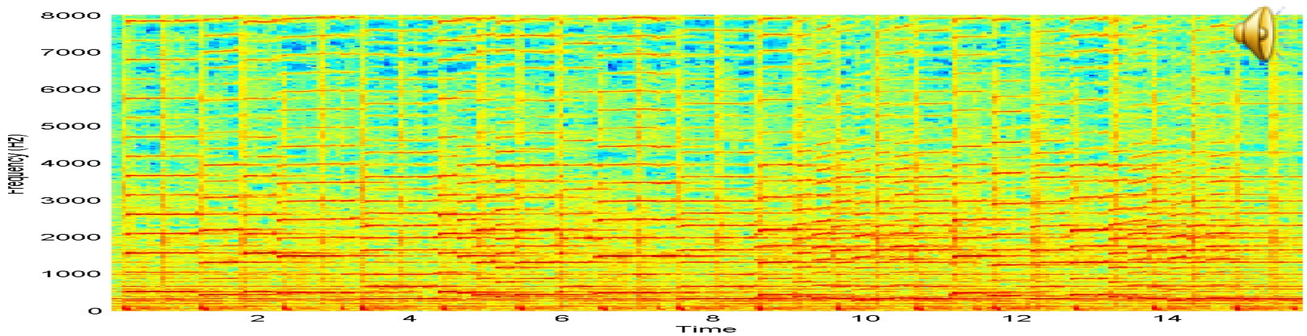




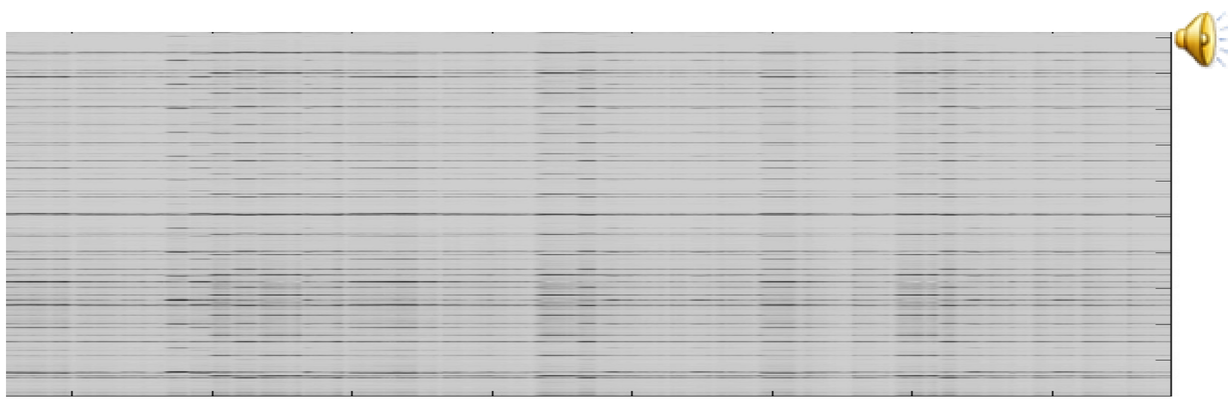
- $\mathbf{P} = \mathbf{W}\,(\mathbf{W}^T\mathbf{W})^{-1}\,\mathbf{W}^T$
- Projected Spectrogram = **PM**

We’re actually computing a score

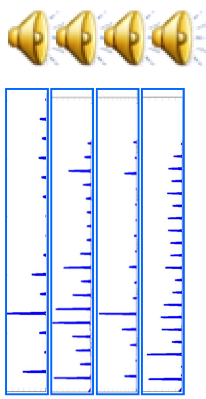
M =



H = ?



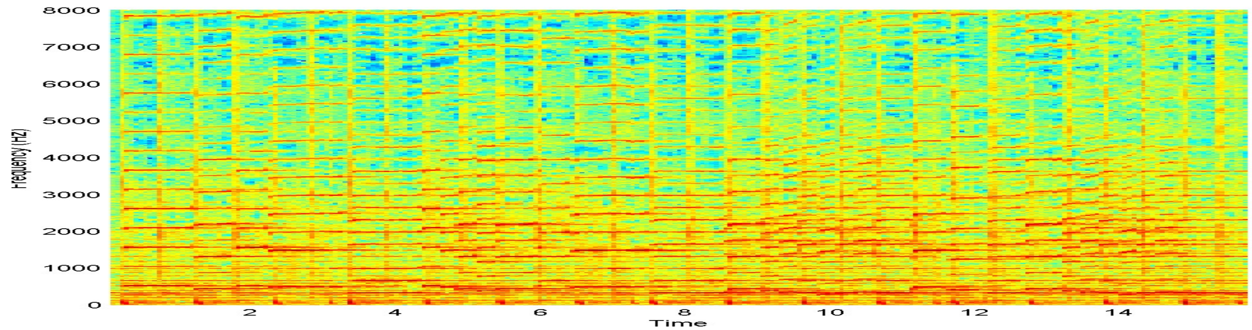
W =



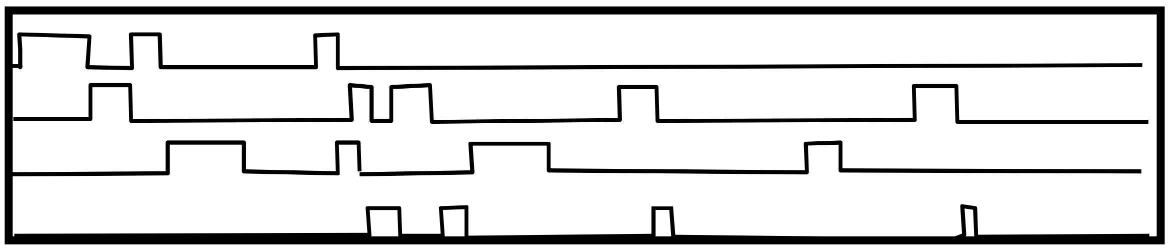
- **M ~ WH**
- **H = pinv(W)M**

How about the other way?

M =



H =

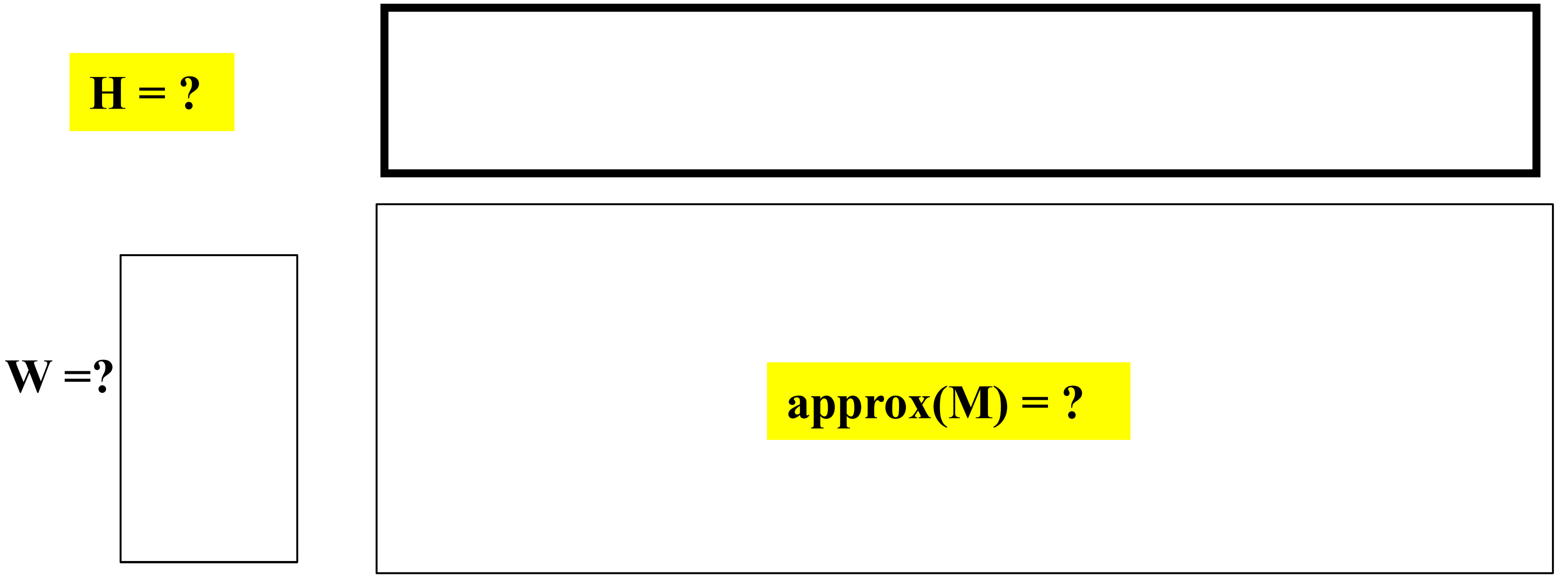


W = ?

U = ?

- M ~ WH W = Mpinv(H) U = WH

When both parameters are unknown



- Must estimate both **H** and **W** to best approximate **M**
- Ideally, must learn *both* the *notes* and *their* transcription!

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 $\mathbf{M}\mathbf{M}^{\mathrm{T}}$
  - PCA!!

- $\mathbf{M} \sim \mathbf{W}\mathbf{H}$ is an approximation
- Also, the rows of $\mathbf{H}$ are *decorrelated*
  - Trivial to prove that $\mathbf{H}\mathbf{H}^{\mathrm{T}}$ is diagonal

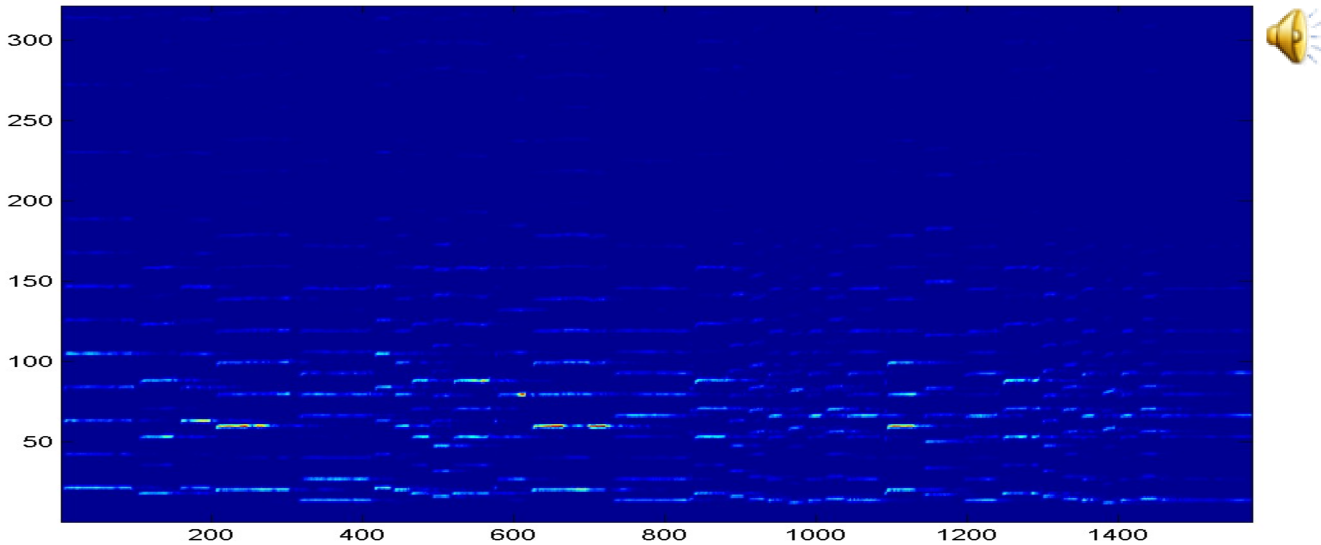
PCA

$$\mathbf{W}, \mathbf{H} = \arg\min_{\overline{\mathbf{W}}, \overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{W}\mathbf{H}} \|_F^2$$

$$\mathbf{M} \approx \mathbf{W}\mathbf{H}$$

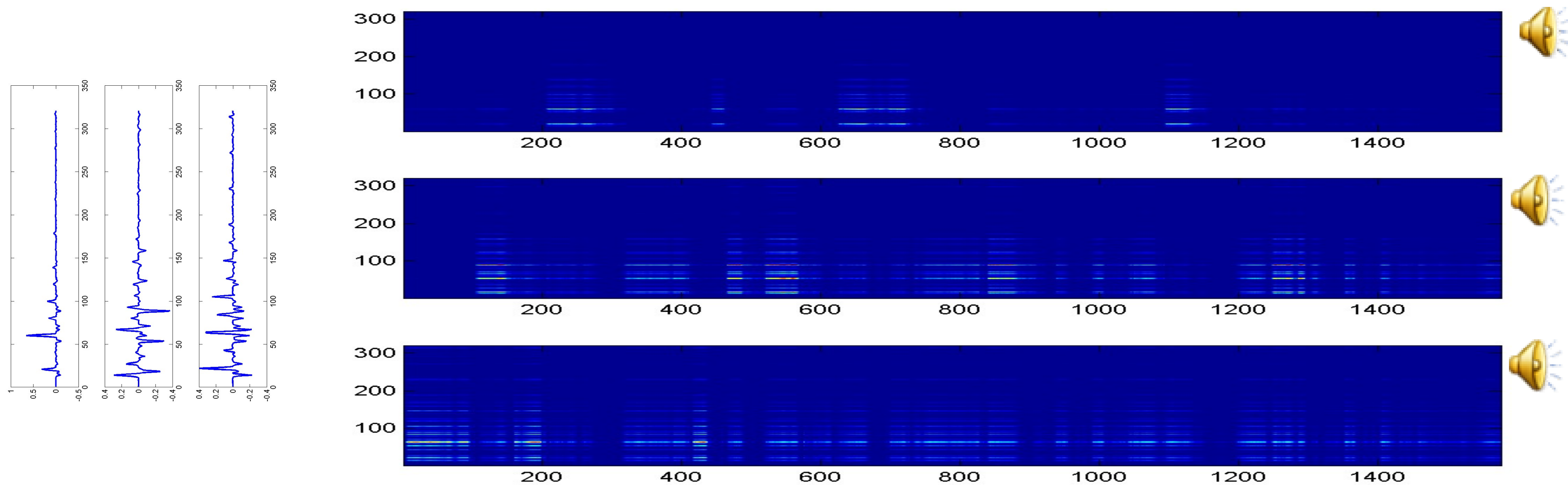
- The columns of W are the bases we have learned
  - The linear "building blocks" that compose the music
- They represent "learned" notes

So how does that work?



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

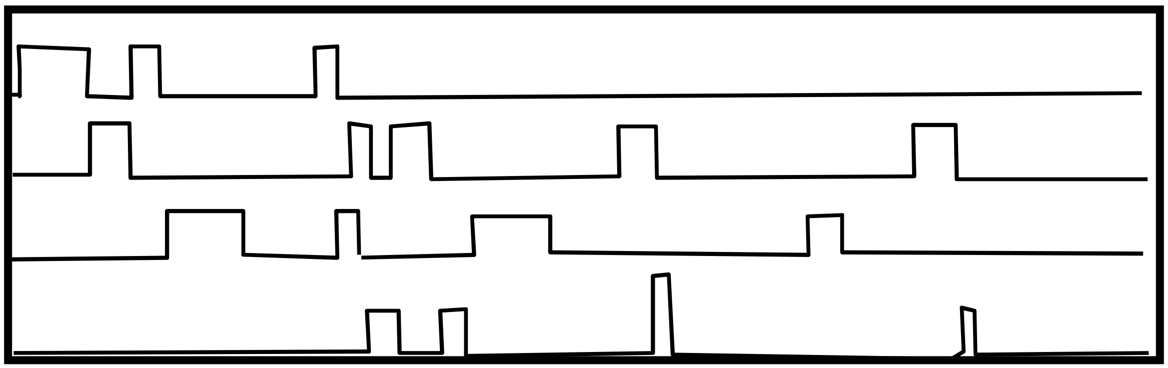
So how does that work?



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

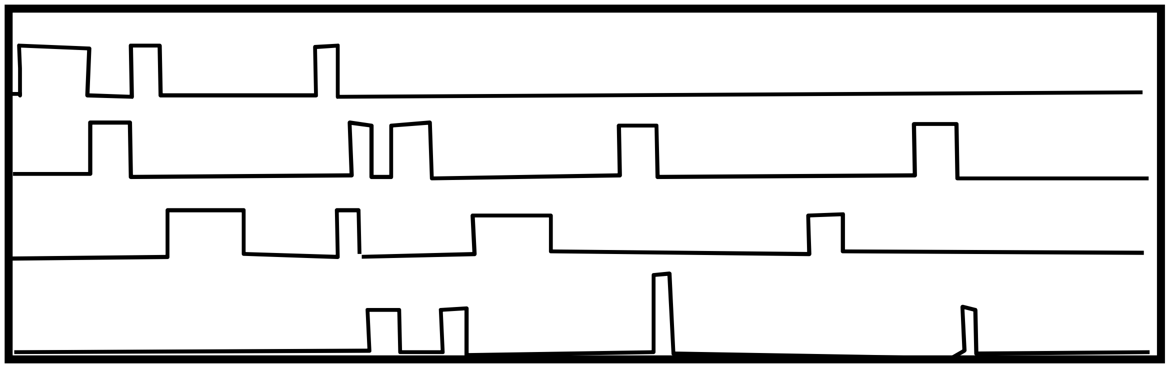
PCA through decorrelation of notes

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



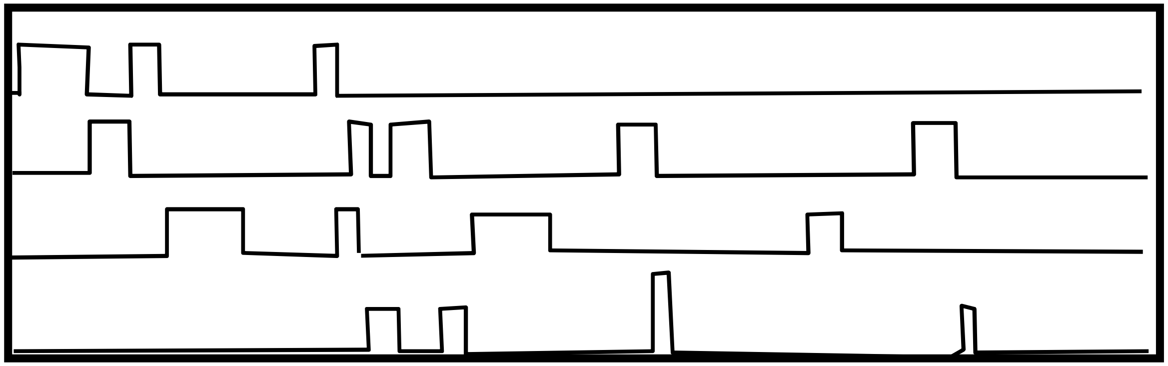
- Different constraint: Constraint $\mathbf{H}$ to be decorrelated
  - $\mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{D}$
- This will result exactly in PCA too
- Decorrelation of $\mathbf{H}$ Interpretation: What does this mean?

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

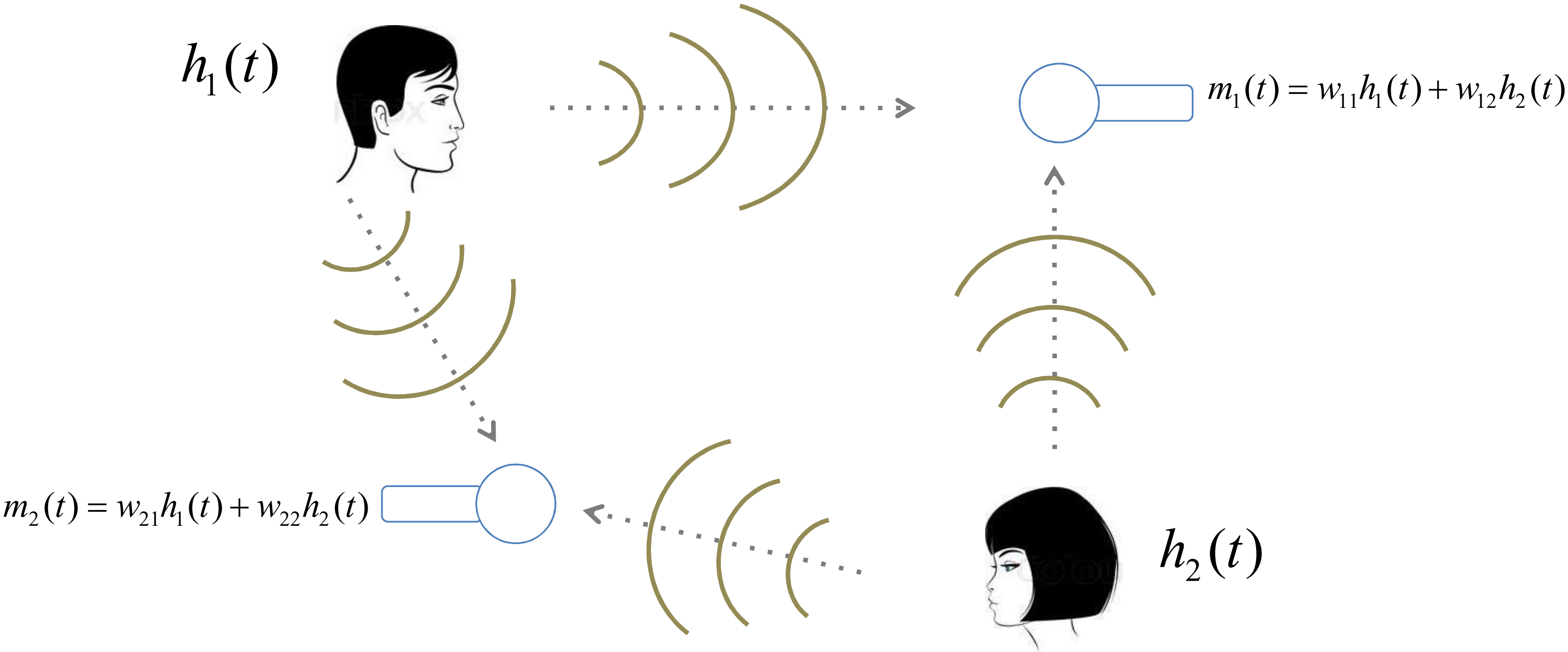
- **Attempting to find statistically independent components of the mixed signal**
  - ***Independent Component Analysis***

Formulating it with Independence

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

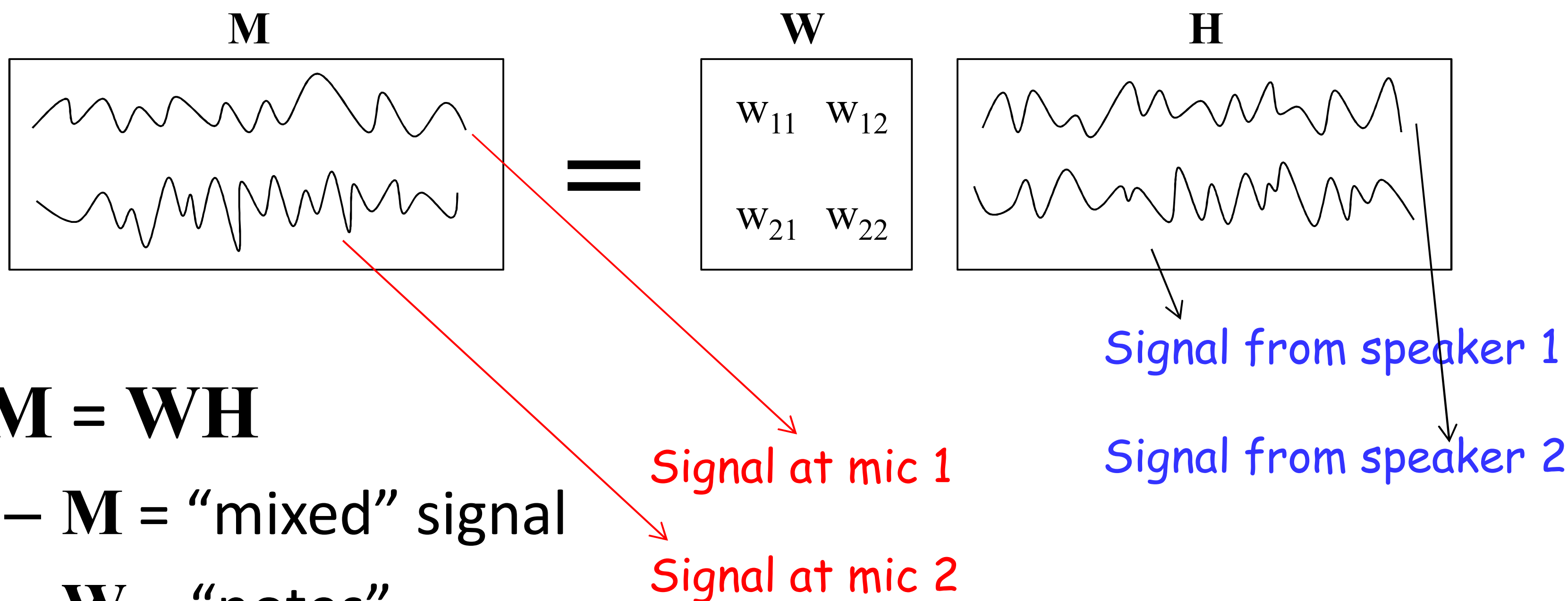
- Impose statistical independence constraints on decomposition

Changing problems for a bit



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

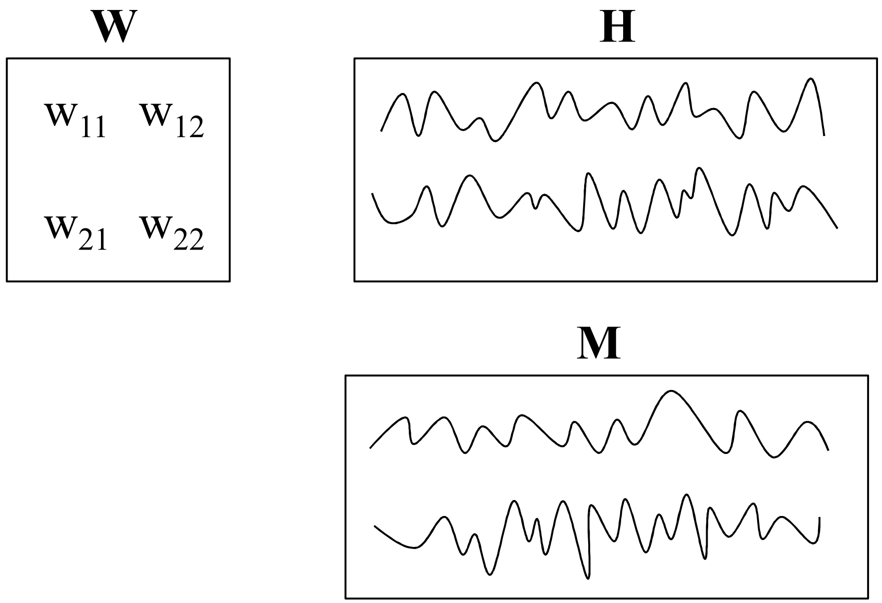
A Separation Problem



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

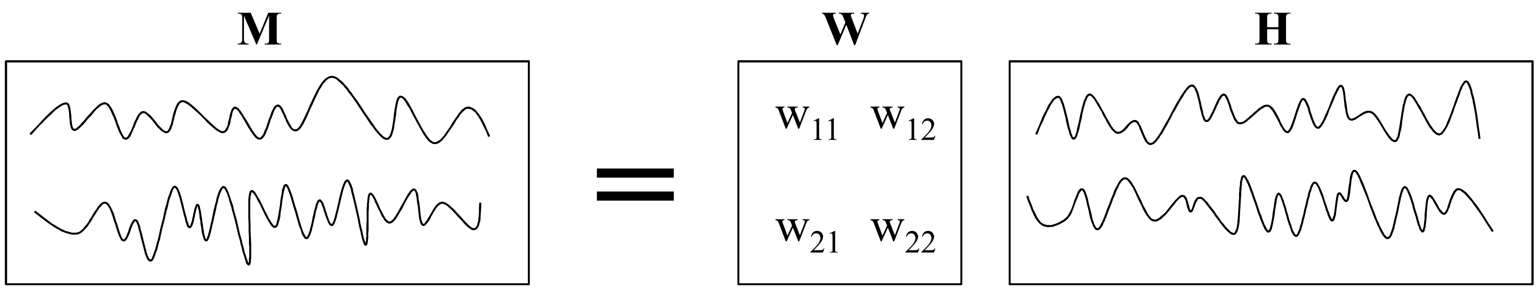
- Separation challenge: Given only **M** estimate **H**
- Identical to the problem of "finding notes"

A Separation Problem



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

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

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

Remember this form

An ugly algebraic solution

$$\mathbf{M} = \mathbf{WH} \quad \ldots\ldots\ldots \quad \mathbf{H} = \mathbf{AM}$$

- We could *decorrelate* signals by algebraic manipulation
  - We know uncorrelated signals have diagonal correlation matrix
  - So we transformed the signal so that it has a diagonal correlation matrix ($\mathbf{HH}^T$)
- 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
- *Is there a simple matrix we could just similarly diagonalize to make them independent?*

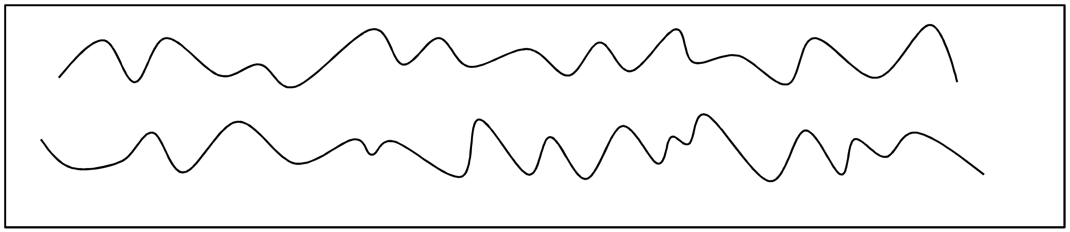
An ugly algebraic solution

- We *decorrelated* signals by diagonalizing the covariance matrix

- *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

H



- 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.

Zero Mean

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

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

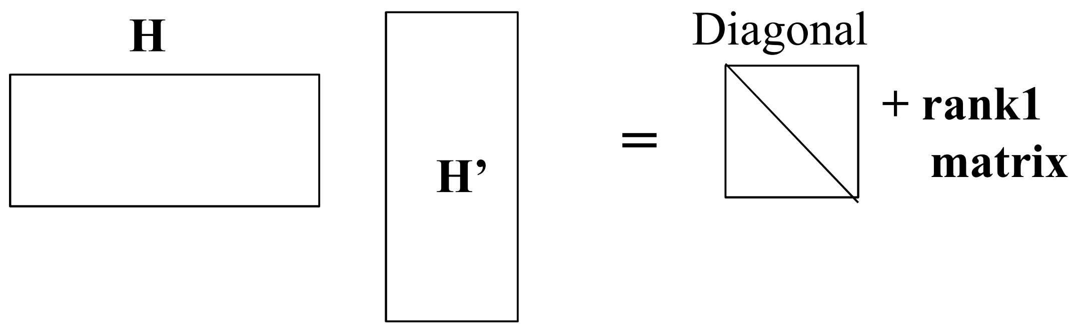
- If mean($\mathbf{M}$) = 0 => mean($\mathbf{H}$) = 0
  - $E[\mathbf{H}] = \mathbf{A}.E[\mathbf{M}] = \mathbf{A0} = \mathbf{0}$
  - First step of ICA: Set the mean of $\mathbf{M}$ to 0

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

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

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

Emulating Independence..



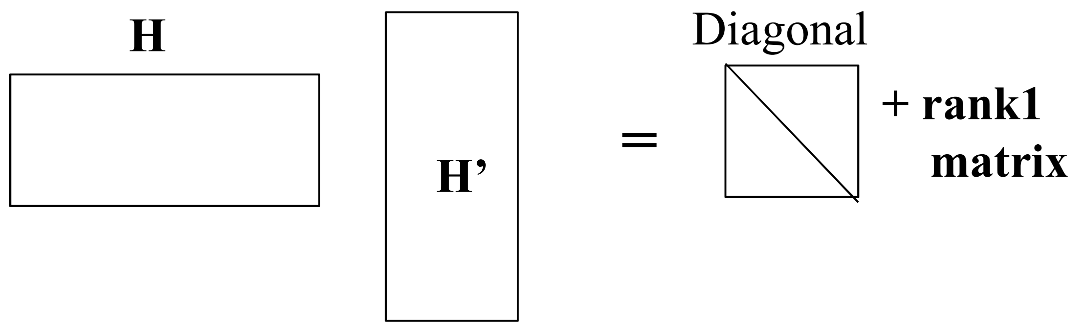
H=AM

A=BC

H=BCM

- Independence → Uncorrelatedness
- Find **C** such that **CM** is decorrelated
  - PCA
- Find **B** such that **B(CM)** is independent
- A little more than PCA

Decorrelating and Whitening



H=AM

A=BC

H=BCM

- Eigen decomposition $\mathbf{MM}^T = \mathbf{ESE}^T$
- $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{E}^T$
- $\mathbf{X} = \mathbf{CM}$
- Not merely decorrelated but ***whitened***
  - $\mathbf{XX}^T = \mathbf{CMM}^T\mathbf{C}^T = \mathbf{S}^{-1/2}\mathbf{E}^T\,\mathbf{ESE}^T\mathbf{ES}^{-1/2} = \mathbf{I}$
- **C** is the ***whitening matrix***

Uncorrelated != Independent

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

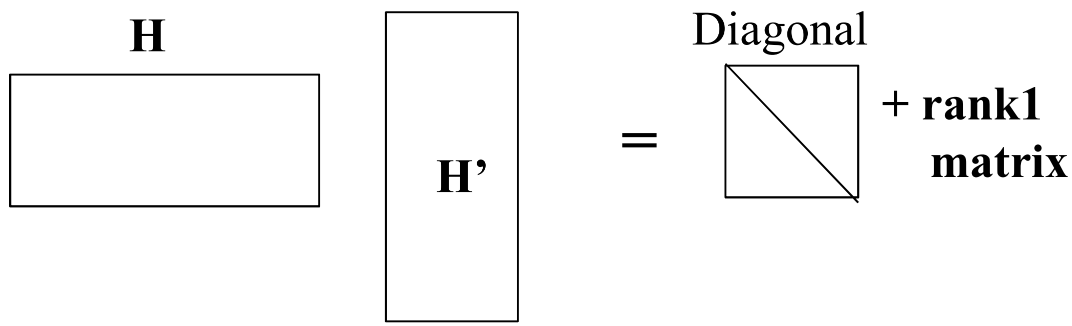
$$E[\mathbf{x}_i \mathbf{x}_j] = 0 \text{ if } i \text{ != } j$$

- 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



H=AM

A=BC

H=BCM

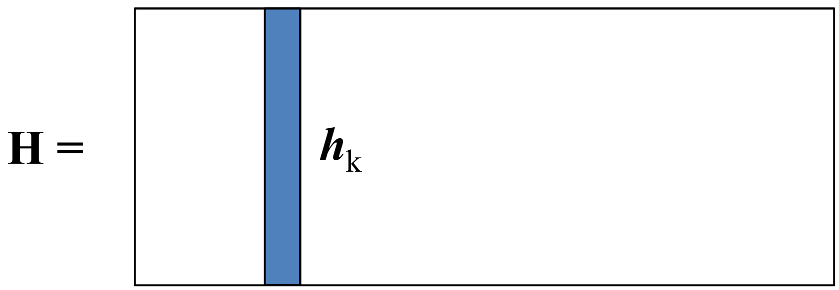
H=BX

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

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

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

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

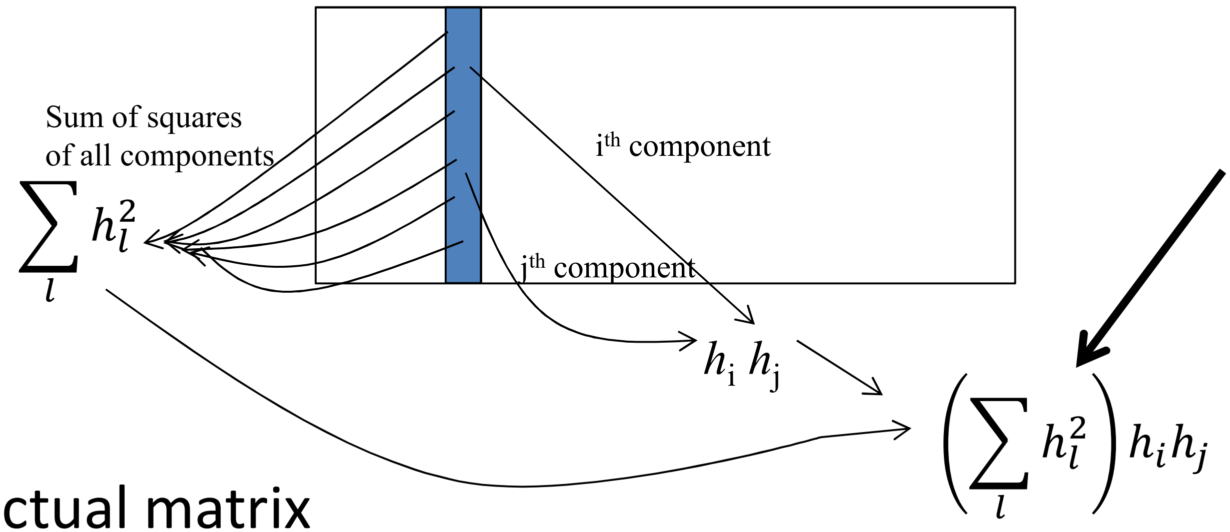
  - $\boldsymbol{h}$ are the columns of $\mathbf{H}$
  - Assuming $\boldsymbol{h}$ is real, else replace transposition with Hermitian

ICA: The D matrix

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

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

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



On the actual matrix

$$\boldsymbol{D} = \sum_{\boldsymbol{k}} \|\boldsymbol{h}_k\|^2 \boldsymbol{h}_k \boldsymbol{h}_k^{\mathrm{T}}$$

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

ICA: The D matrix

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

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

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

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

- If the h_i terms were independent and zero mean
- For $i \mathrel{!}= j$

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

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

- 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. $\mathbf{BB}^T = \mathbf{I}$

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
- $\mathbf{D} = E[\mathbf{h}^T \mathbf{h} \mathbf{h} \mathbf{h}^T] = E[\mathbf{x}^T\mathbf{BB}^T\mathbf{x} \mathbf{B}^T \mathbf{x} \mathbf{x}^T\mathbf{B}]$

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

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

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

Diagonalizing D

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

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

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

  $\mathbf{D_x} = \mathbf{U \Lambda U^\mathrm{T}}$

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

B frees the fourth moment

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

- $\mathbf{U}$ is a unitary matrix, i.e. $\mathbf{U^T U = UU^T = I}$ (identity)
- $\mathbf{H = BX = U^T X}$
- $\mathbf{h = U^T x}$
- The fourth moment matrix of $\mathbf{H}$ is

$$E[||\mathbf{h}||^2 \, \mathbf{h}\, \mathbf{h}^T] = \mathbf{U}^T E[||\mathbf{x}||^2 \, \mathbf{x}\mathbf{x}^T]\mathbf{U}$$
$$= \mathbf{U}^T \mathbf{D_x} \mathbf{U}$$
$$= \mathbf{U}^T \mathbf{U} \, \Lambda \, \mathbf{U}^T \mathbf{U} = \Lambda$$

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

Overall Solution

- Objective: Estimate A such that the rows of $\mathbf{H} = \mathbf{AM}$ are independent

- Step 1: *Whiten M*
  - $\mathbf{C}$ is the (transpose of the) matrix of Eigen vectors 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
     $\boldsymbol{R}_{\mathrm{MM}} = \mathbf{ESE}^{\mathrm{T}}, \quad \mathbf{C} = \mathbf{S}^{-1/2}\mathbf{E}^{\mathrm{T}}$
  4. Compute $\mathbf{X} = \mathbf{CM}$
  5. Compute the fourth moment matrix $\mathbf{D}' = E[||\mathbf{x}||^2\mathbf{xx}^{\mathrm{T}}]$
  6. Diagonalize $\mathbf{D}'$ via Eigen decomposition
  7. $\mathbf{D}' = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}}$
  8. Compute $\mathbf{A} = \mathbf{U}^{\mathrm{T}}\boldsymbol{C}$
- The fourth moment matrix of $\mathbf{H} = \mathbf{AM}$ 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

Enforcing Independence

- Specifically ensure that the components of $\mathbf{H}$ are independent
  - $\mathbf{H} = \mathbf{AM}$

- *Contrast function*: A non-linear function that has a minimum value when the *output components* are independent

- Define and minimize a contrast function
  - » F(**AM**)
- Contrast functions are often only *approximations* too..

A note on pre-whitening

- The mixed signal is usually “prewhitened” for all ICA methods
  - Normalize variance along all directions
  - Eliminate second-order dependence

- Eigen decomposition $\mathbf{MM}^T = \mathbf{ESE}^T$
- $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{E}^T$

- Can use *first K* columns of $\mathbf{E}$ only if only K independent sources are expected
  - In microphone array setup – only K < M sources

- $\mathbf{X} = \mathbf{CM}$
  - $E[\mathbf{x}_i\mathbf{x}_j] = \delta_{ij}$ for centered signal

The contrast function

- *Contrast function*: A non-linear function that has a minimum value when the *output components* are independent

- An explicit contrast function

$$I(\mathbf{H}) = \sum_i H(\bar{\mathbf{h}}_i) - H(\bar{\mathbf{h}})$$

- With constraint : $\mathbf{H} = \mathbf{B}\mathbf{X}$
  - $\mathbf{X}$ is “whitened” $\mathbf{M}$

Linear Functions

- $\mathbf{h} = \mathbf{Bx}, \quad \mathbf{x} = \mathbf{B}^{-1}\mathbf{h}$
  - Individual columns of the $\mathbf{H}$ and $\mathbf{X}$ matrices
  - $\mathbf{x}$ is mixed signal, $\mathbf{B}$ is the *unmixing* matrix

$$P_{\mathbf{h}}(\mathbf{h}) = P_{\mathbf{x}}(\mathbf{B}^{-1}\mathbf{h})\,|\,\mathbf{B}\,|^{-1}$$

$$H(\mathbf{x}) = -\int P(\mathbf{x}) \log P(\mathbf{x}) d\mathbf{x}$$

$$\log P(\mathbf{h}) = \log P_{\mathbf{x}}(\mathbf{B}^{-1}\mathbf{h}) - \log(|\,\mathbf{B}\,|)$$

$$H(\mathbf{h}) = H(\mathbf{x}) + \log|\,\mathbf{B}\,|$$

The contrast function

$$I(\mathbf{H}) = \sum_i H(\bar{\mathbf{h}}_i) - H(\bar{\mathbf{h}})$$

$$I(\mathbf{H}) = \sum_i H(\bar{\mathbf{h}}_i) - H(\mathbf{x}) - \log|\mathbf{B}|$$

- Ignoring $H(\mathbf{x})$ (Const)

$$J(\mathbf{H}) = \sum_i H(\bar{\mathbf{h}}_i) - \log|\mathbf{B}|$$

- Minimize the above to obtain $\mathbf{B}$

An alternate approach

- Recall PCA
- $\mathbf{M} = \mathbf{WH}$, the columns of $\mathbf{W}$ must be orthogonal
- Leads to: $\min_{\mathbf{W}} ||\mathbf{M} - \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{M}||^2 + \Lambda.\mathrm{trace}(\mathbf{W}^{\mathrm{T}}\mathbf{W})$
  - Error minimization framework to estimate $\mathbf{W}$
- Can we arrive at an error minimization framework for ICA
- Define an "Error" objective that represents independence

An alternate approach

- Definition of Independence – if x and y are independent:
  - $\mathrm{E}[f(x)g(y)] = \mathrm{E}[f(x)]\mathrm{E}[g(y)]$
  - Must hold for *every* *f*() and *g*()!!

An alternate approach

- Define $\mathbf{g(H) = g(BX)}$ (component-wise function)

$$\begin{matrix} g(h_{11}) & g(h_{21}) & \cdots \\ g(h_{12}) & g(h_{22}) & \\ \vdots & \vdots & \end{matrix}$$

- Define $\mathbf{f(H) = f(BX)}$

$$\begin{matrix} f(h_{11}) & f(h_{21}) & \cdots \\ f(h_{12}) & f(h_{22}) & \\ \vdots & \vdots & \end{matrix}$$

An alternate approach

- $\mathbf{P} = \mathbf{g(H)\ f(H)}^T = \mathbf{g(BX)\ f(BX)}^T$

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{21} & \cdots \\ P_{12} & P_{22} & \\ \vdots & \vdots & \end{bmatrix}$$

$$\mathbf{P}_{ij} = \mathbf{E}[\mathrm{g}(h_i)\mathrm{f}(h_j)]$$

This is a square matrix

- Must ideally be

$$\mathbf{Q} = \begin{bmatrix} Q_{11} & Q_{21} & \cdots \\ Q_{12} & Q_{22} & \\ \vdots & \vdots & \end{bmatrix}$$

$$Q_{ij} = E[g(h_i)]E[f(h_j)] \quad i \neq j$$

$$Q_{ii} = E[g(h_i)f(h_i)]$$

- $\text{Error} = \|\mathbf{P}\text{-}\mathbf{Q}\|_F^2$

An alternate approach

- Ideal value for $\mathbf{Q}$

$$\mathbf{Q} = \begin{bmatrix} Q_{11} & Q_{21} & \cdots \\ Q_{12} & Q_{22} & \\ \vdots & \vdots & \end{bmatrix}$$

$$Q_{ij} = E[g(h_i)]E[f(h_j)] \quad i \neq j$$

$$Q_{ii} = E[g(h_i)f(h_i)]$$

- If g() and f() are odd symmetric functions $E[g(h_i)] = 0$ for all i
  - Since = $E[h_i] = 0$ ($\mathbf{H}$ is centered)
- $\mathbf{Q}$ is a Diagonal Matrix!!!

An alternate approach

- Minimize Error

$$\mathbf{P} = \mathbf{g}(\mathbf{BX})\mathbf{f}(\mathbf{BX})^{\mathbf{T}}$$

$$\mathbf{Q} = Diagonal$$

$$error = \| \mathbf{P} - \mathbf{Q} \|_F^2$$

- Leads to trivial Widrow Hopf type iterative rule:

$$\mathbf{E} = Diag - \mathbf{g}(\mathbf{BX})\mathbf{f}(\mathbf{BX})^{\mathbf{T}}$$

$$\mathbf{B} = \mathbf{B} + \eta \mathbf{E}\mathbf{X}^{\mathbf{T}}$$

Update Rules

- Multiple solutions under different assumptions for g() and f()
- $\mathbf{H} = \mathbf{BX}$
- $\mathbf{B} = \mathbf{B} + \eta \, \Delta\mathbf{B}$
- Jutten Herraut : Online update
  - $\Delta B_{ij} = f(\mathbf{h}_i)g(\mathbf{h}_j)$; -- actually assumed a recursive neural network
- Bell Sejnowski
  - $\mathbf{\Delta B = ([B^T]^{-1} - g(H)X^T)}$

Update Rules

- Multiple solutions under different assumptions for g() and f()
- $\mathbf{H} = \mathbf{B}\mathbf{X}$
- $\mathbf{B} = \mathbf{B} + \eta\,\Delta\mathbf{B}$

- Natural gradient -- f() = identity function
  - $\Delta\mathbf{B} = (\mathbf{I} - \mathbf{g}(\mathbf{H})\mathbf{H}^T)\,\mathbf{X}^T$
- Cichoki-Unbehaeven
  - $\Delta\mathbf{B} = (\mathbf{I} - \mathbf{g}(\mathbf{H})\mathbf{f}(\mathbf{H})^T)\,\mathbf{X}^T$

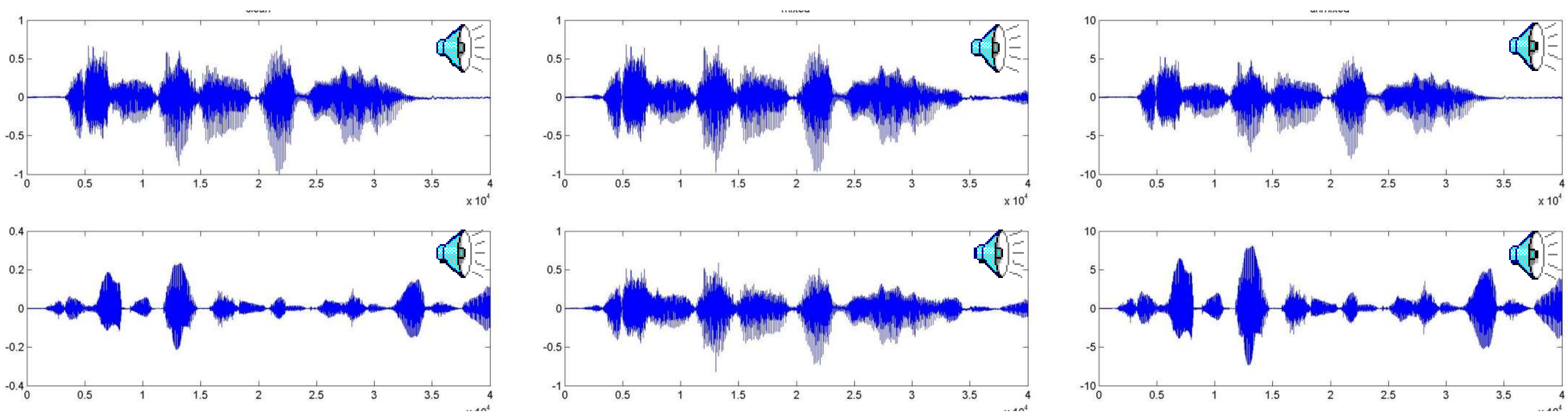
What are G() and F()

- Must be odd symmetric functions
- Multiple functions proposed

$$g(x) = \begin{cases} x + \tanh(x) & \text{x is super Gaussian} \\ x - \tanh(x) & \text{x is sub Gaussian} \end{cases}$$

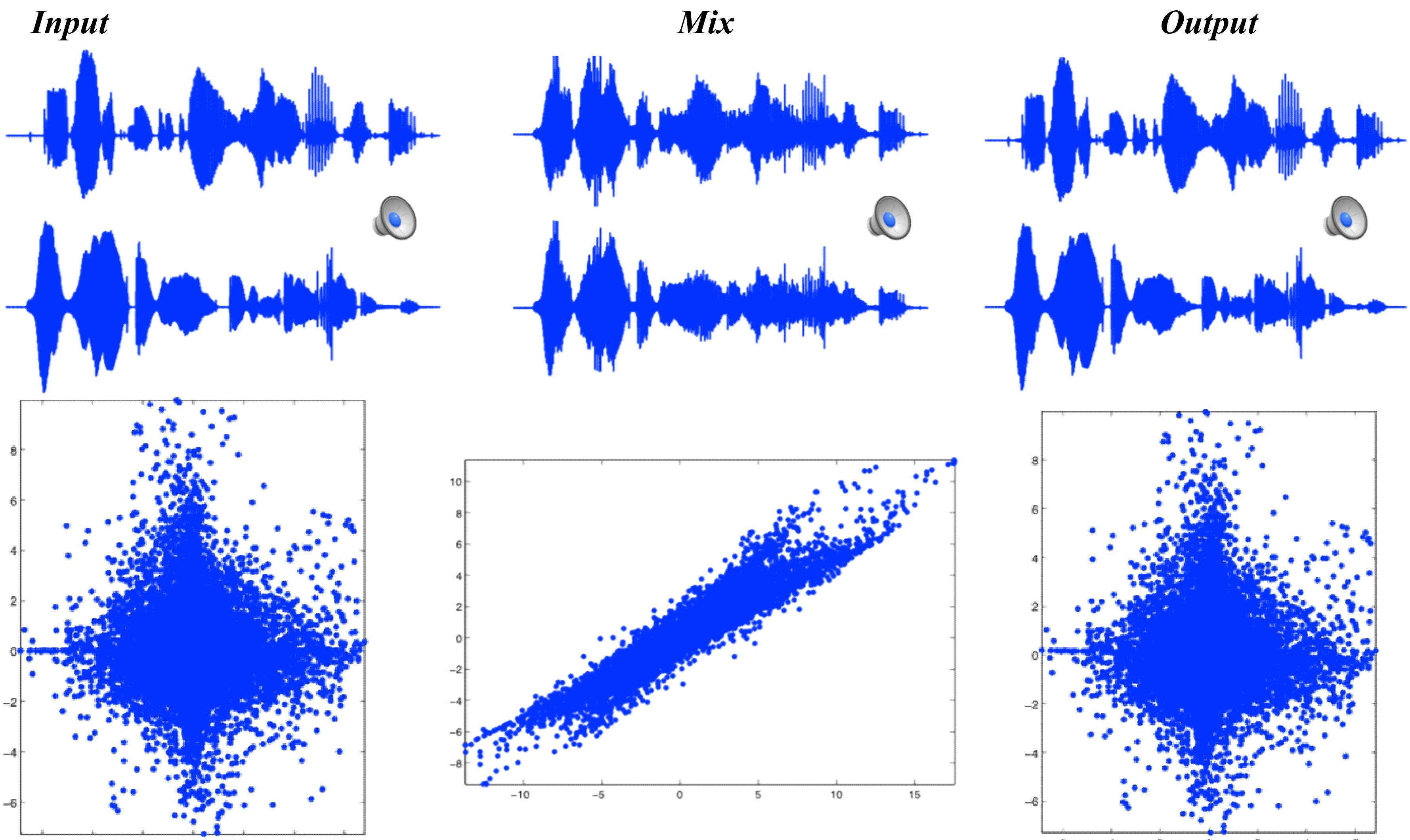
- Audio signals in general
  - $\Delta \mathbf{B} = (\mathbf{I} - \mathbf{H}\mathbf{H}^T - \mathbf{K}\tanh(\mathbf{H})\mathbf{H}^T)\, \mathbf{X}^T$
- Or simply
  - $\Delta \mathbf{B} = (\mathbf{I} - \mathbf{K}\tanh(\mathbf{H})\mathbf{H}^T)\, \mathbf{X}^T$

So how does it work?

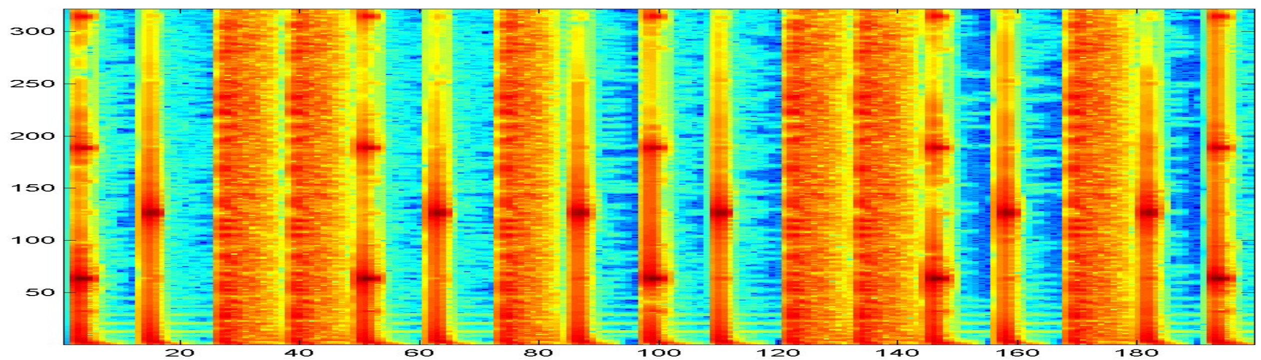


- Example with instantaneous mixture of two speakers
- Natural gradient update
- Works very well!

Another example!

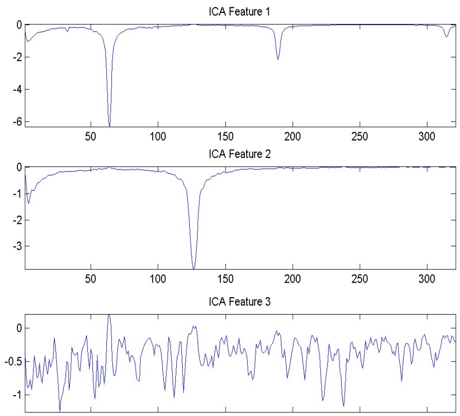


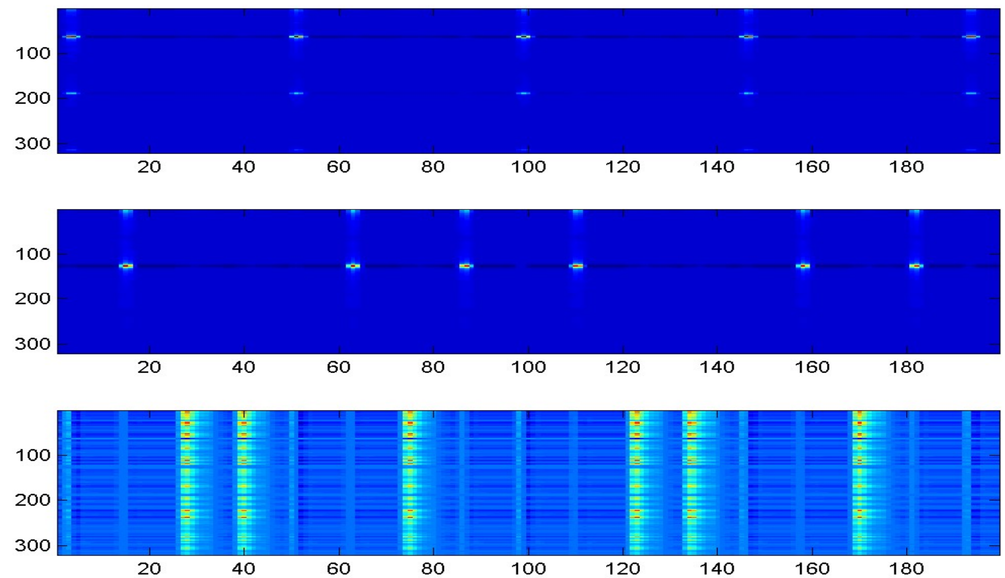
Another Example



- Three instruments..

The Notes

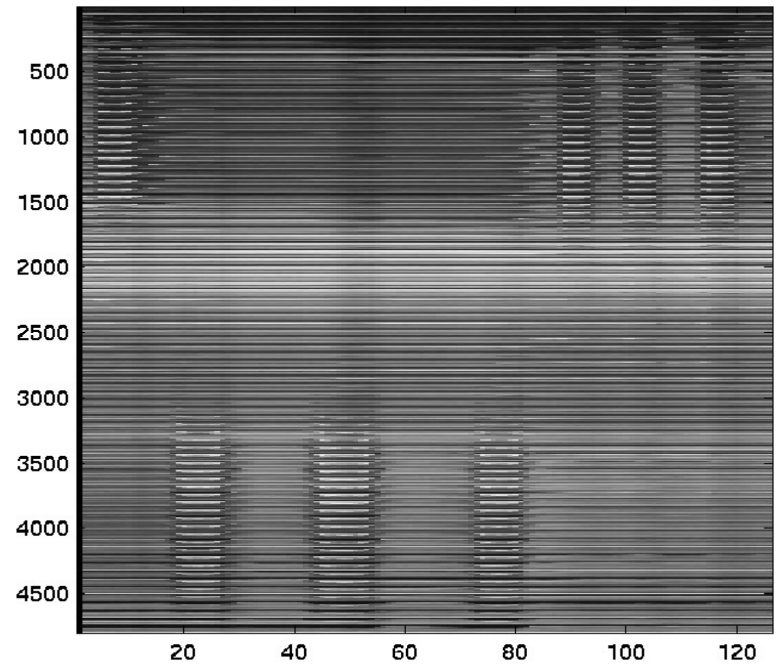




- Three instruments..

ICA for data exploration

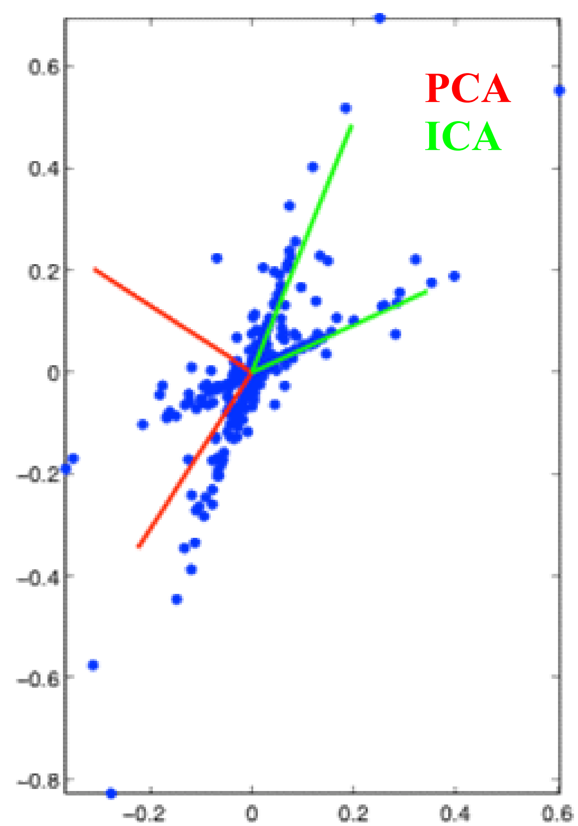
- The “bases” in PCA represent the “building blocks”
  - Ideally notes
- Very successfully used
- So can ICA be used to do the same?



ICA vs PCA bases

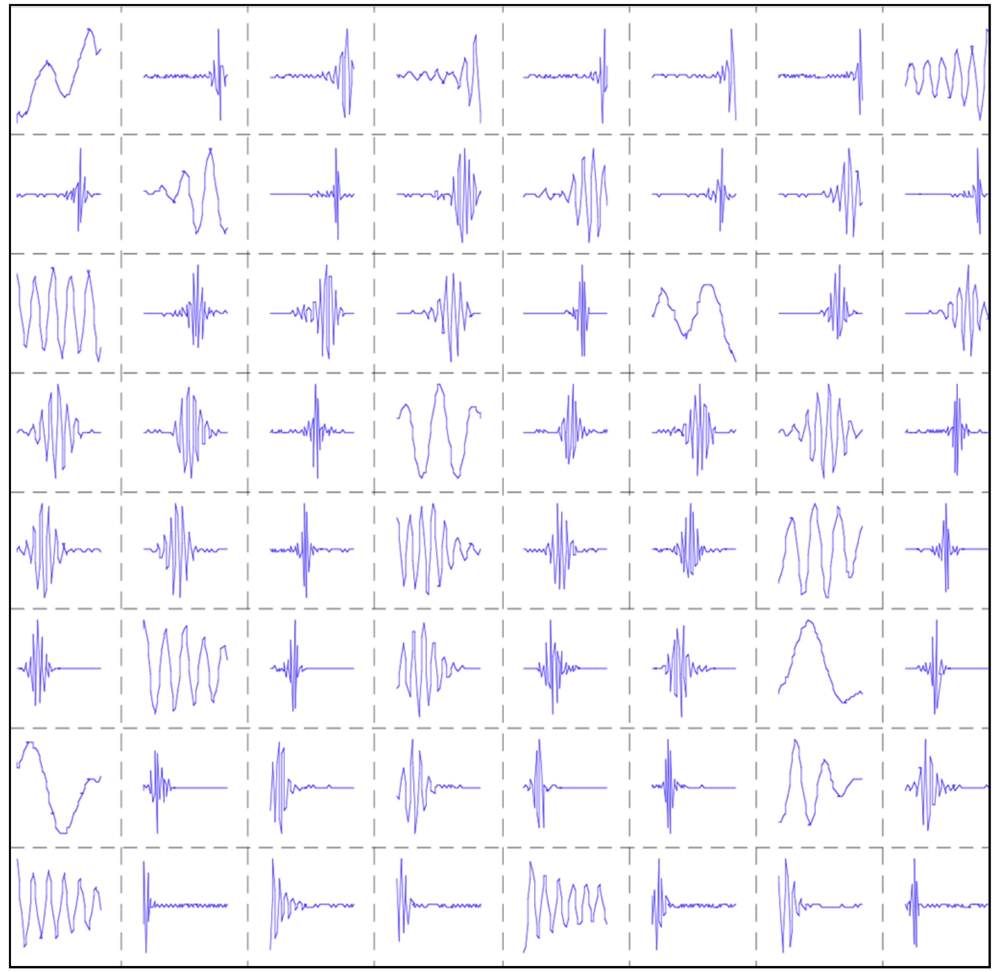
- Motivation for using ICA vs PCA
- PCA will indicate orthogonal directions of maximal variance
  - May not align with the data!
- ICA finds directions that are independent
  - More likely to "align" with the data

Non-Gaussian data

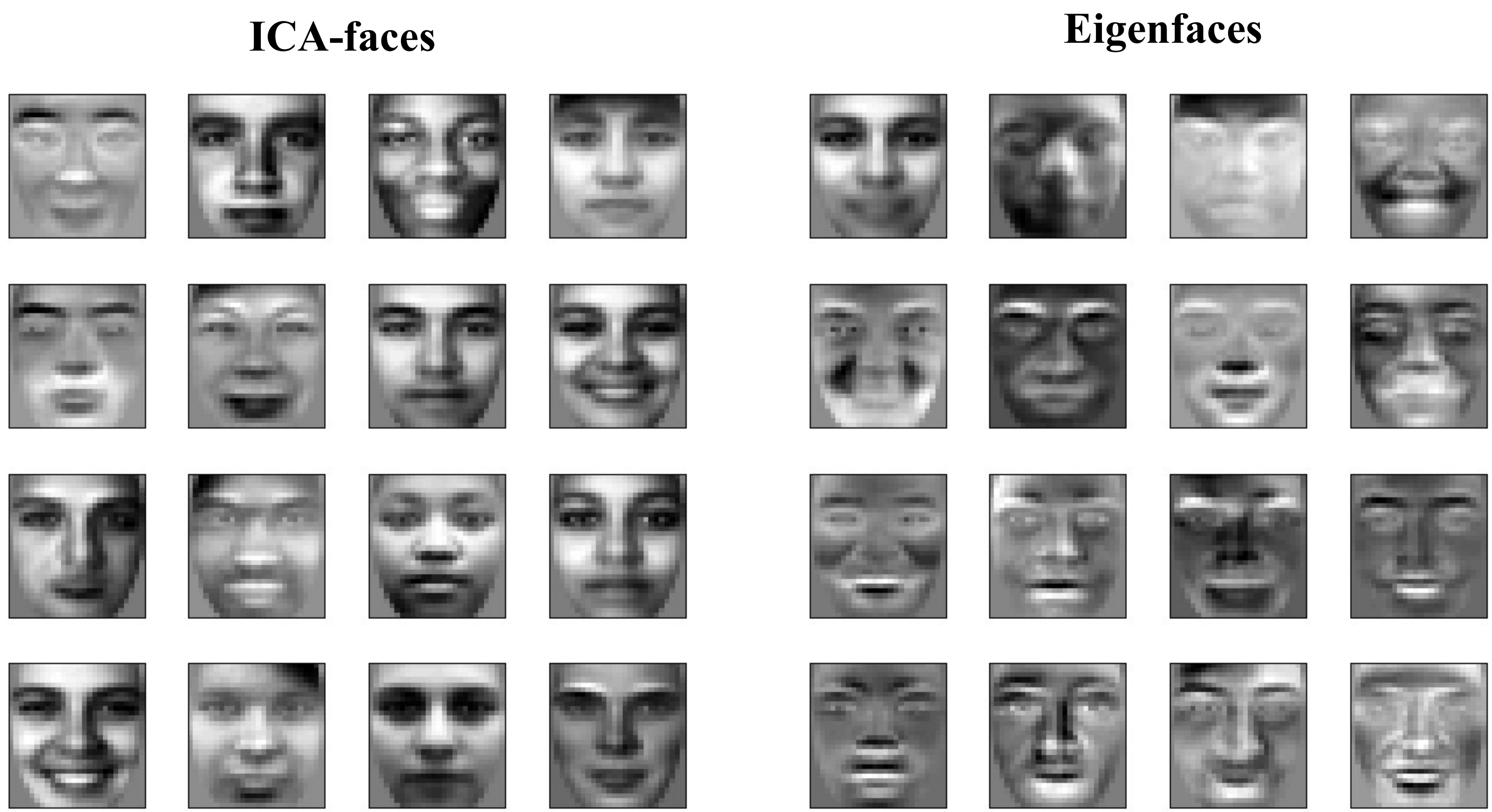


Finding useful transforms with ICA

- 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



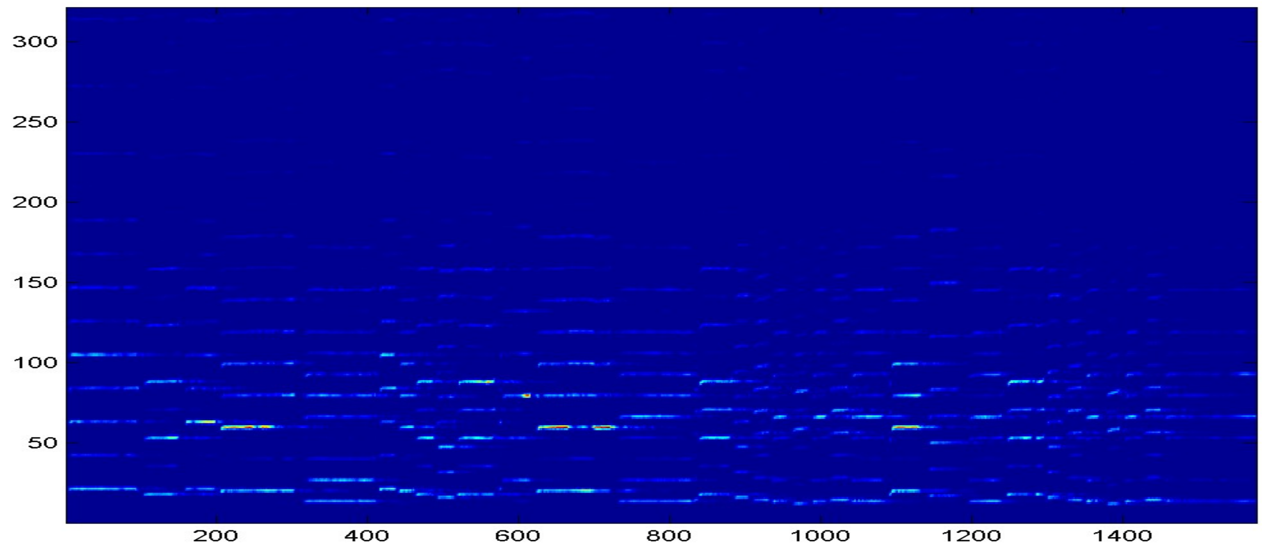
Example case: ICA-faces vs. Eigenfaces



ICA for Signal Enhncement

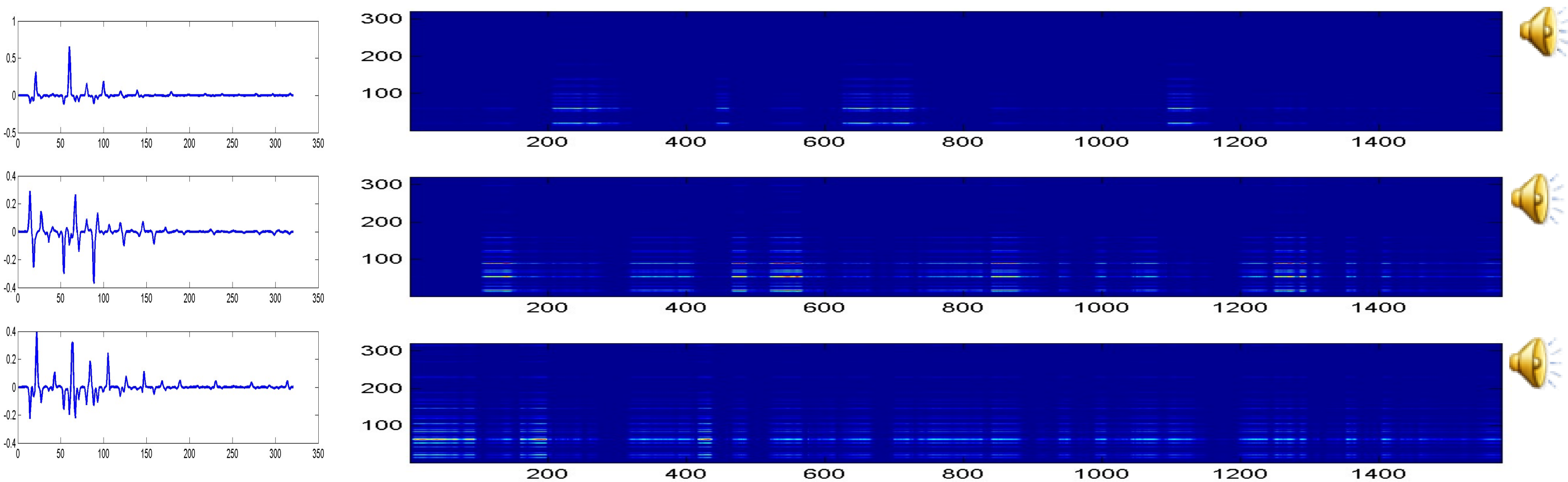
- Very commonly used to enhance EEG signals
- EEG signals are frequently corrupted by heartbeats and biorhythm signals
- ICA can be used to separate them out

So how does that work?



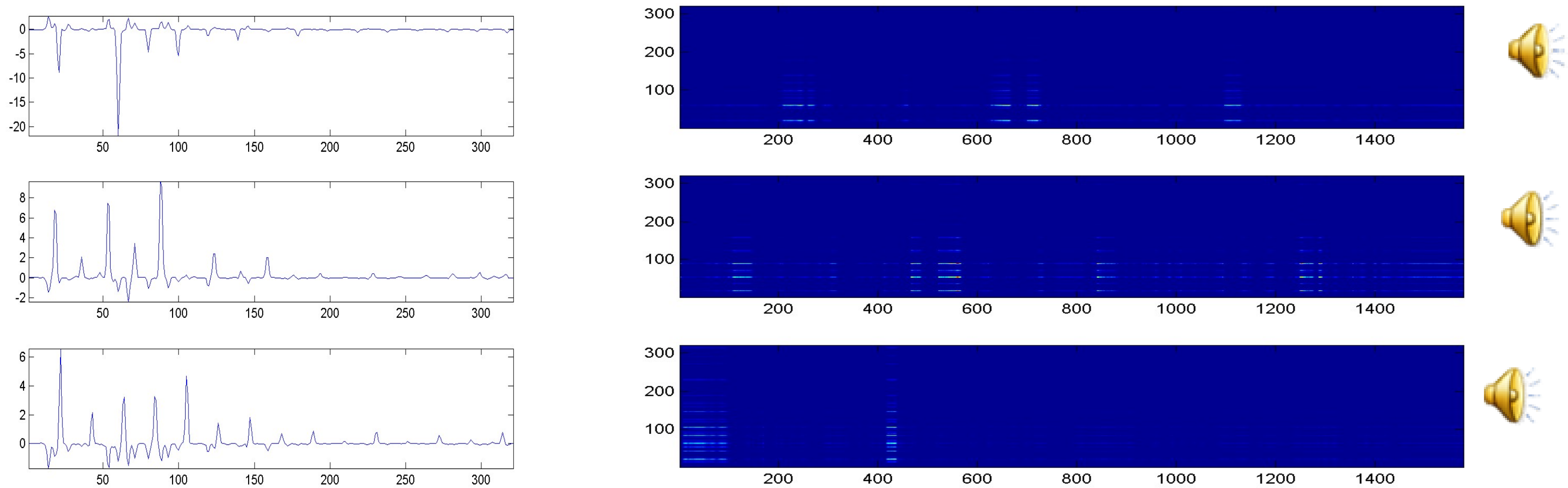
- There are 12 notes in the segment, hence we try to estimate 12 notes..

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..