Machine Learning for Signal Processing

Independent Component Analysis

Class 8. 23 Sep 2013

Instructor: Bhiksha Raj

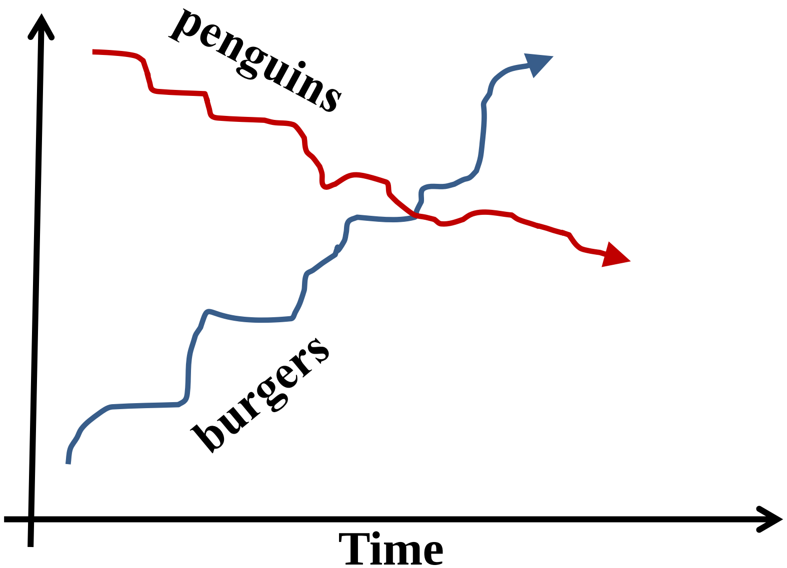
Correlation vs. Causation

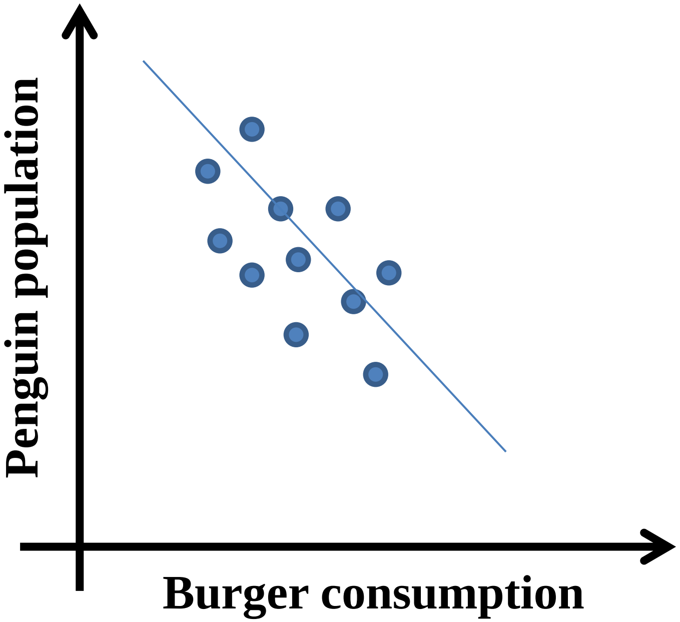
- The consumption of burgers has gone up steadily in the past decade

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

The concept of *correlation*

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





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?

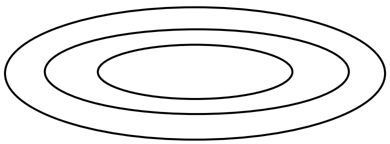
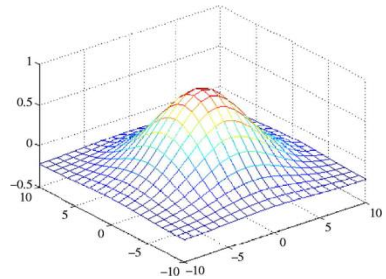
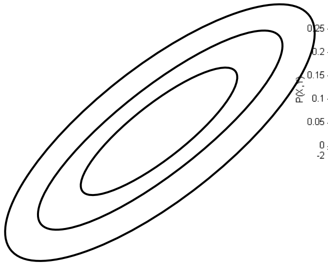
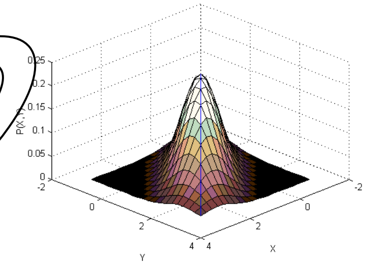
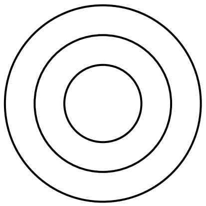
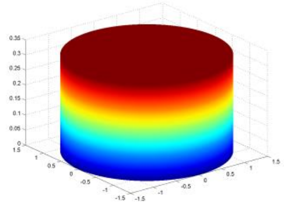
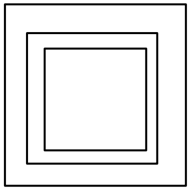
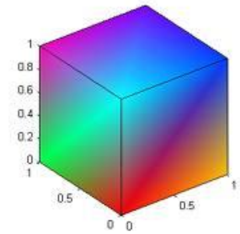
A brief review of basic probability

- *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 X is the same regardless of the value of Y

Uncorrelatedness

- Which of the above represent uncorrelated RVs?

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

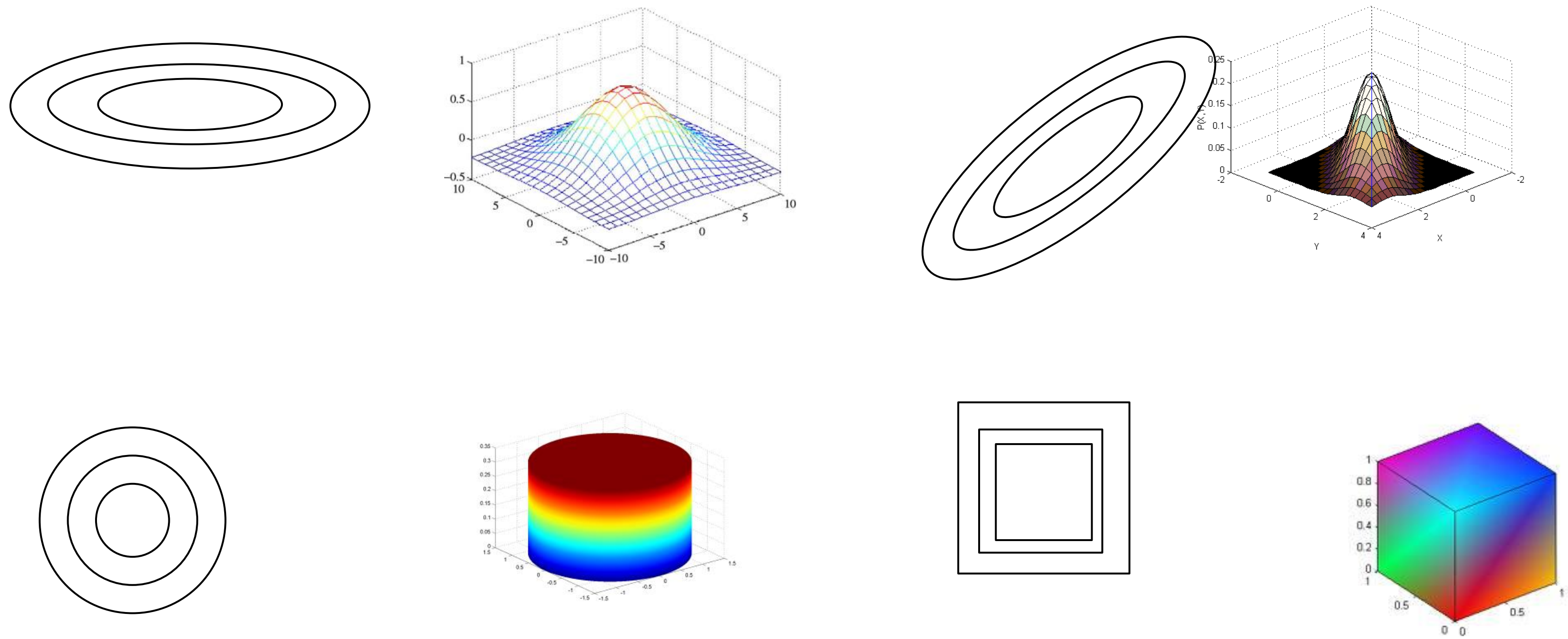
A brief review of basic probability

- ***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 not the other way

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()**

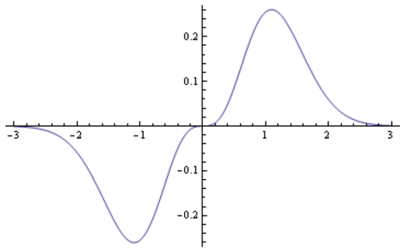
Independence

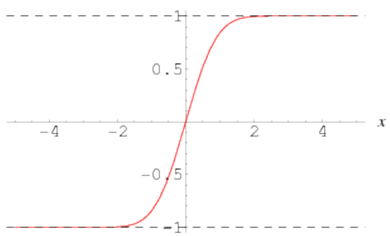


- Which of the above represent independent RVs?
- Which represent uncorrelated RVs?

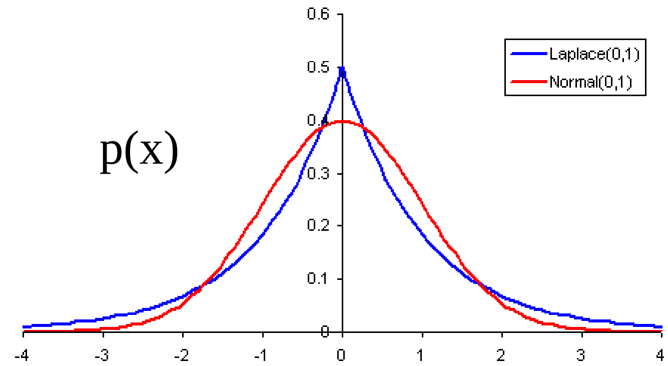
A brief review of basic probability

y = f(x)





p(x)



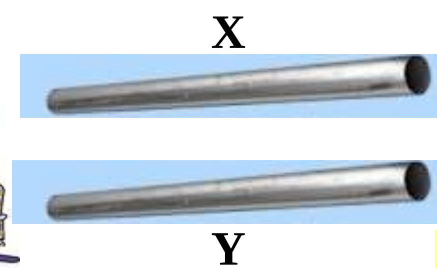
- The expected value of an odd function of an RV is 0 if
  - The RV is 0 mean
  - The PDF is of the RV is symmetric around 0
- **E[f(X)] = 0 if f(X) is odd symmetric**

A 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



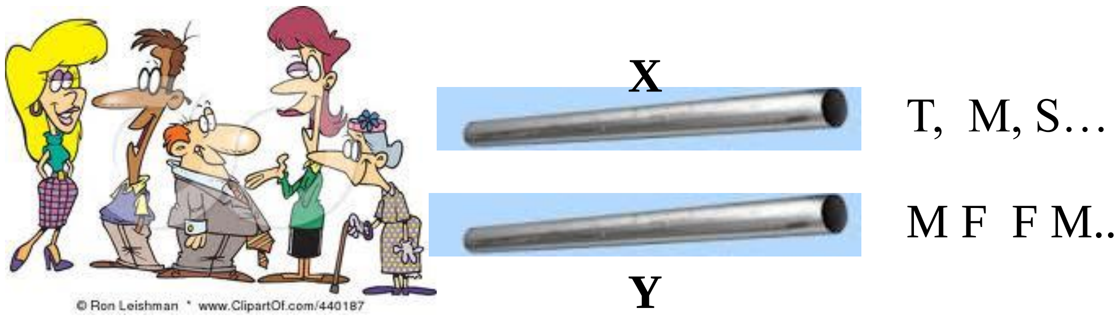
T, M, S…

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



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

- 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

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

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

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

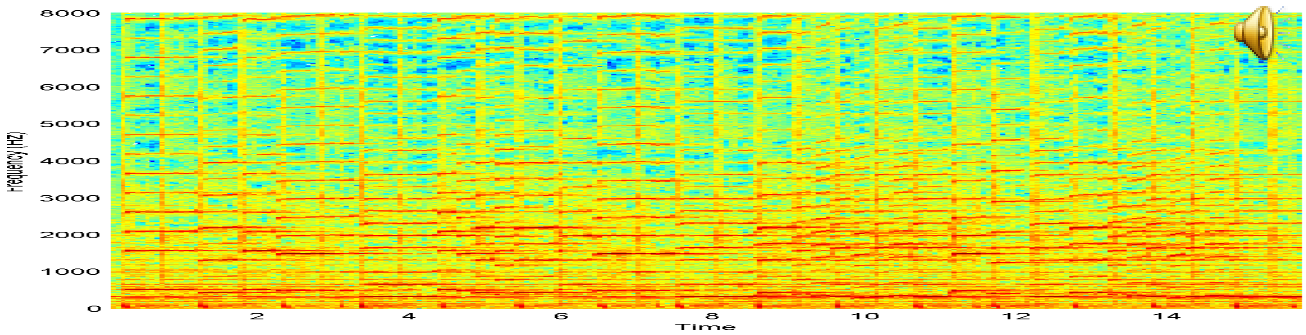
$$= -\sum_{X,Y} P(X,Y) \log P(X) - \sum_{X,Y} P(X,Y) \log P(Y) = H(X) + H(Y)$$

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

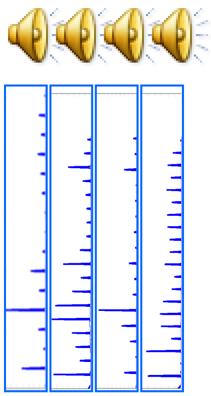
Onward..

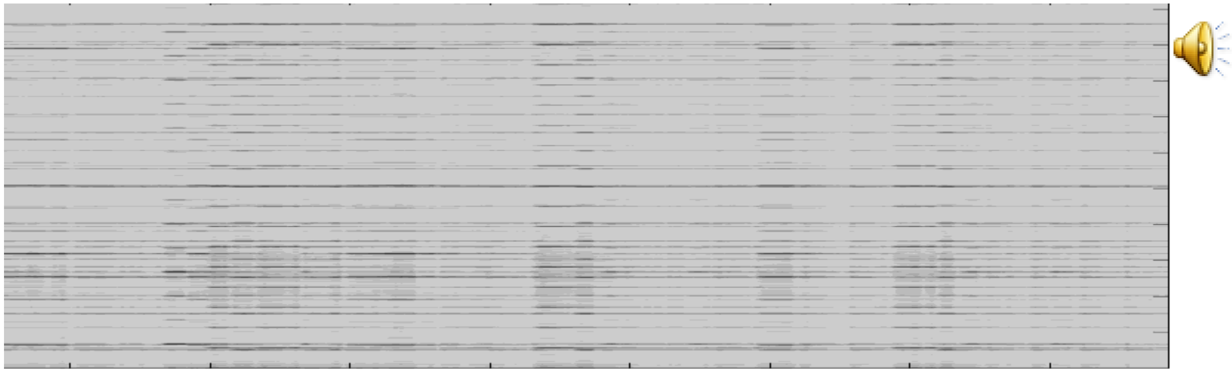
Projection: multiple notes

M =



W =

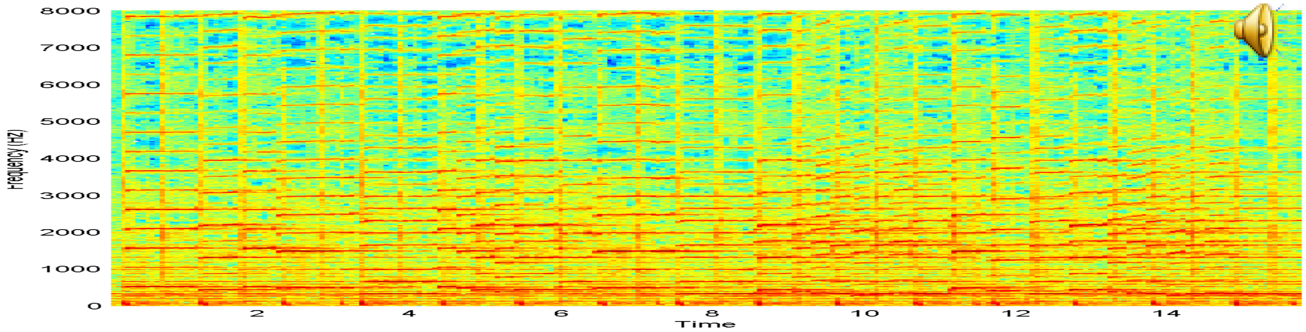




- $\mathbf{P} = \mathbf{W} (\mathbf{W}^T\mathbf{W})^{-1} \mathbf{W}^T$
- Projected Spectrogram = **PM**

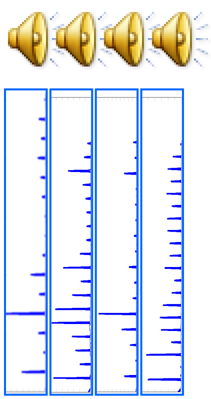
We’re actually computing a score

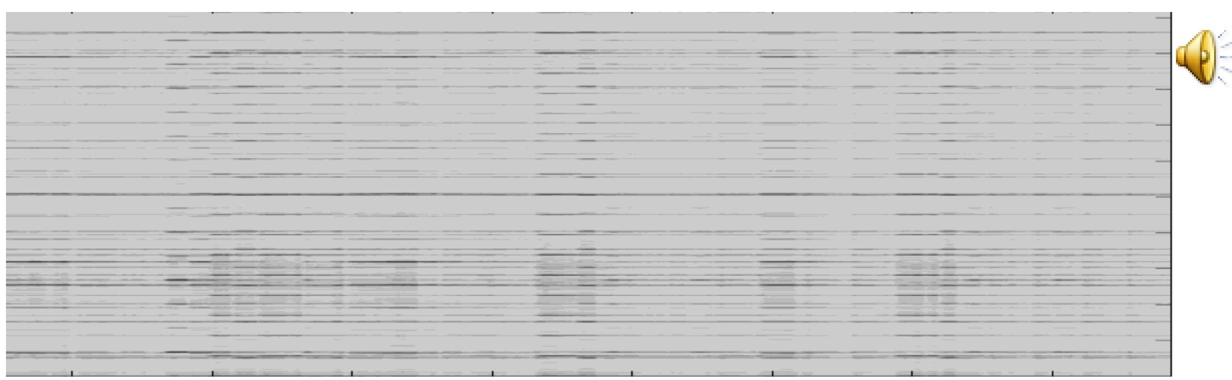
M =



H = ?

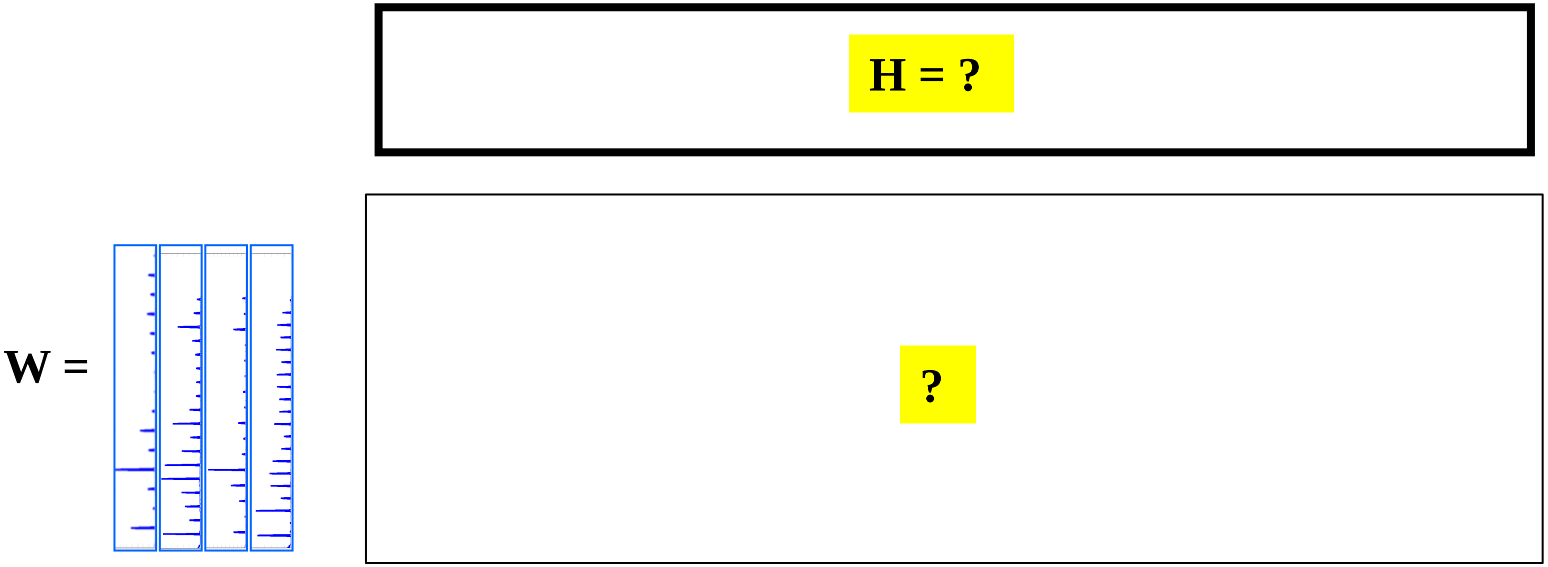
W =





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

So what are we doing here?



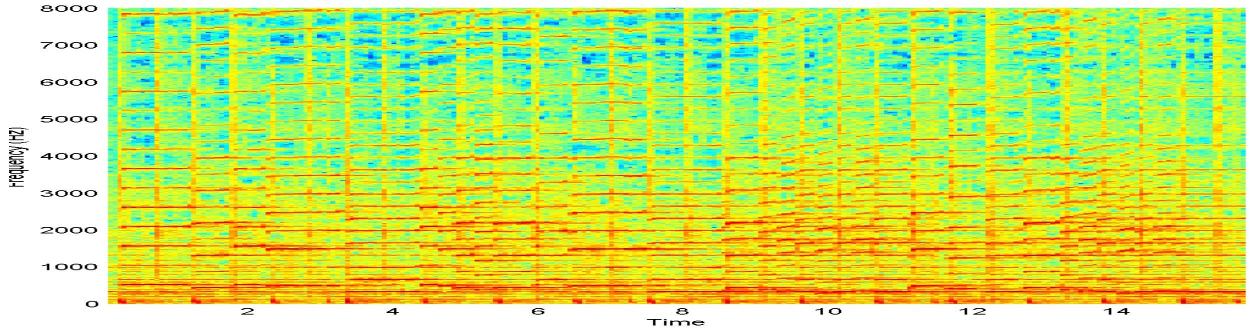
- **M** ~ **WH** is an approximation
- Given **W**, estimate **H** to minimize error

$$\mathbf{H} = \arg\min_{\overline{\mathbf{H}}} \| \mathbf{M} - \mathbf{W}\overline{\mathbf{H}} \|_F^2 = \arg\min_{\overline{\mathbf{H}}} \sum_i \sum_j \left(\mathbf{M}_{ij} - (\mathbf{W}\overline{\mathbf{H}})_{ij} \right)^2$$

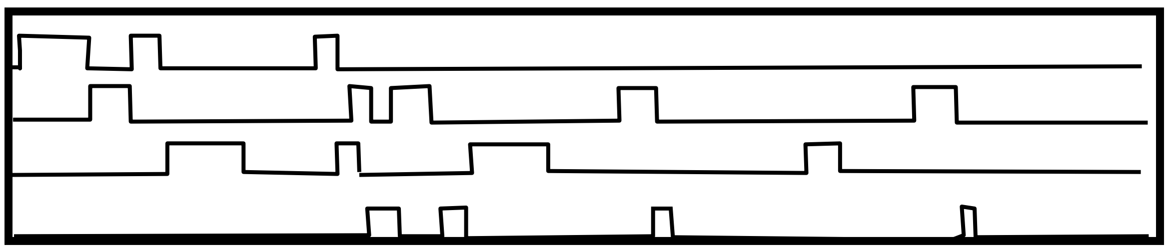
- Must ideally find *transcription* of given notes

How about the other way?

M =



H =

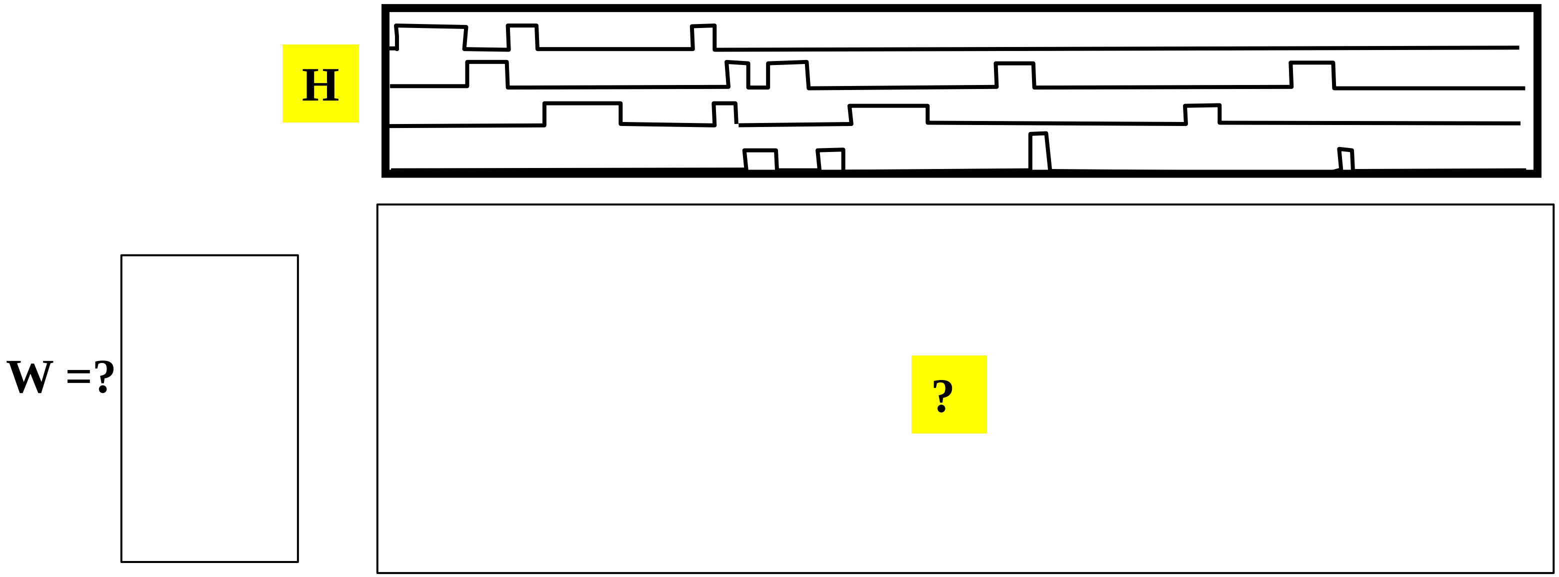


W = ?

U = ?

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

Going the other way..

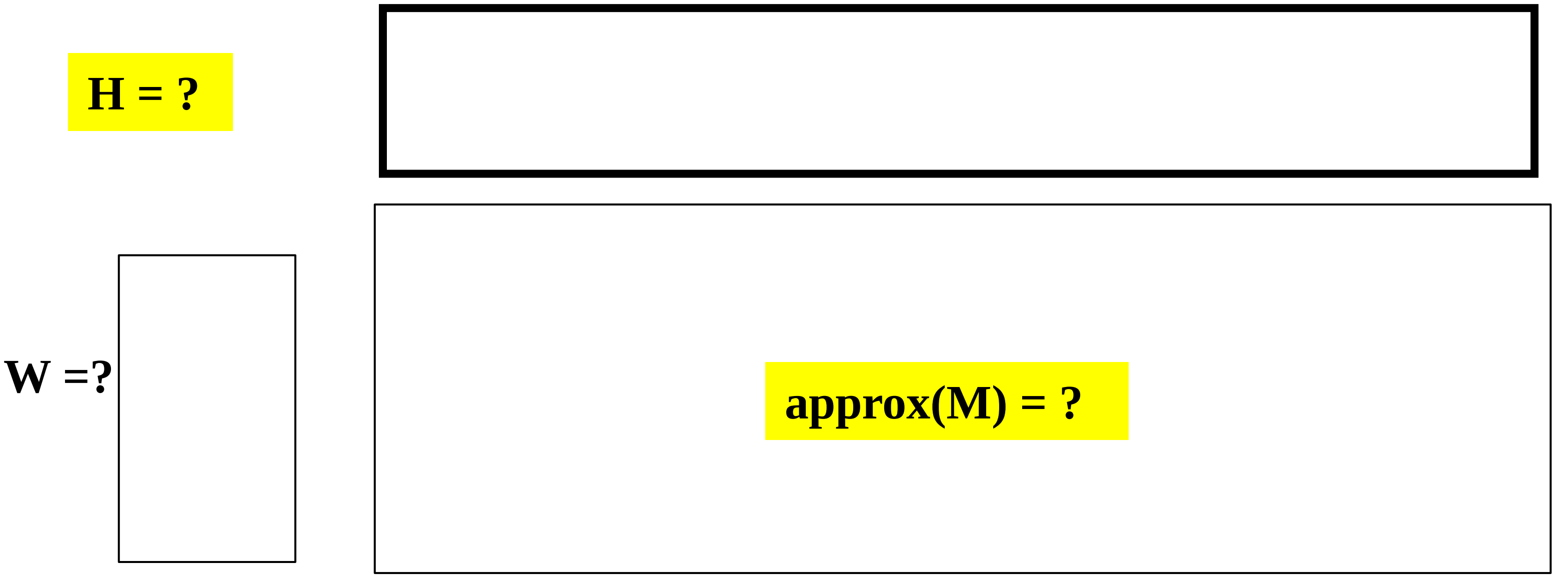


- $\mathbf{M} \sim \mathbf{WH}$ is an approximation
- Given $\mathbf{H}$, estimate $\mathbf{W}$ to minimize error

$$\mathbf{W} = \arg\min_{\overline{\mathbf{W}}} \| \mathbf{M} - \overline{\mathbf{W}}\mathbf{H} \|_F^2 = \arg\min_{\overline{\mathbf{H}}} \sum_i \sum_j \left(\mathbf{M}_{ij} - (\overline{\mathbf{W}}\mathbf{H})_{ij} \right)^2$$

- Must ideally find the *notes* corresponding to the transcription

When both parameters are unknown



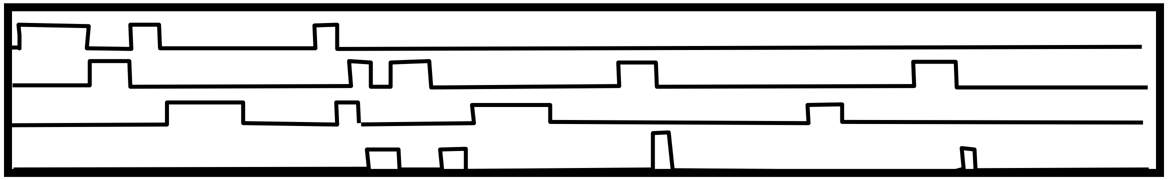
- Must estimate both **H** and **W** to best approximate **M**
- Ideally, must learn *both* the *notes* and *their* transcription!

A least squares solution

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

- Unconstrained
  - For any $\mathbf{W}$, $\mathbf{H}$ that minimizes the error, $\mathbf{W'} = \mathbf{WA}$, $\mathbf{H'} = \mathbf{A}^{-1}\mathbf{H}$ also minimizes the error for any invertible $\mathbf{A}$

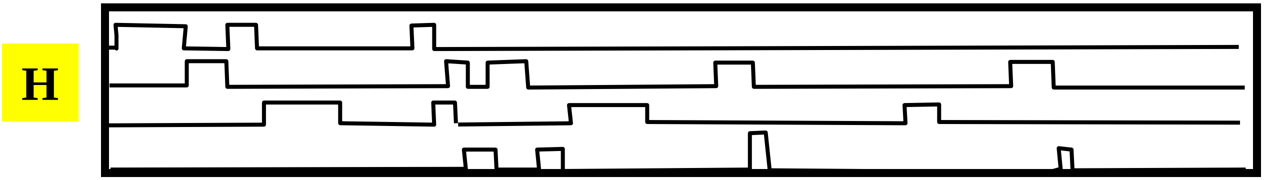
H



- Too many solutions

A *constrained* least squares solution

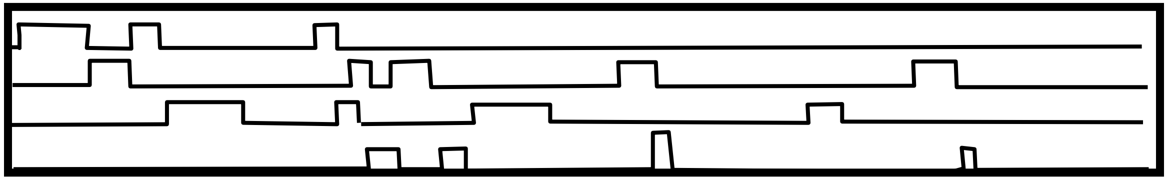
$$\mathbf{W}, \mathbf{H} = \arg\min_{\overline{\mathbf{W}}, \overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{W}}\overline{\mathbf{H}} \|_F^2$$



- For our problem, lets consider the "truth"..
- When one note occurs, the other does not
  - $\mathbf{h}_i^T \mathbf{h}_j = 0$ for all i != j
- The rows of **H** are *uncorrelated*

A least squares solution

H



- Assume: $\mathbf{H}\mathbf{H}^T = \mathbf{I}$
  - Normalizing all rows of $\mathbf{H}$ to length 1
- $\text{pinv}(\mathbf{H}) = \mathbf{H}^T$
- Projecting $\mathbf{M}$ onto $\mathbf{H}$
  - $\mathbf{W} = \mathbf{M}\,\text{pinv}(\mathbf{H}) = \mathbf{M}\mathbf{H}^T$
  - $\mathbf{W}\mathbf{H} = \mathbf{M}\,\mathbf{H}^T\mathbf{H}$

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

$$\mathbf{H} = \arg\min_{\overline{\mathbf{H}}} \| \mathbf{M} - \mathbf{M}\overline{\mathbf{H}}^T\overline{\mathbf{H}} \|_F^2$$ Constraint: Rank(H) = 4

Finding the notes

- Add the constraint: $\mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{I}$

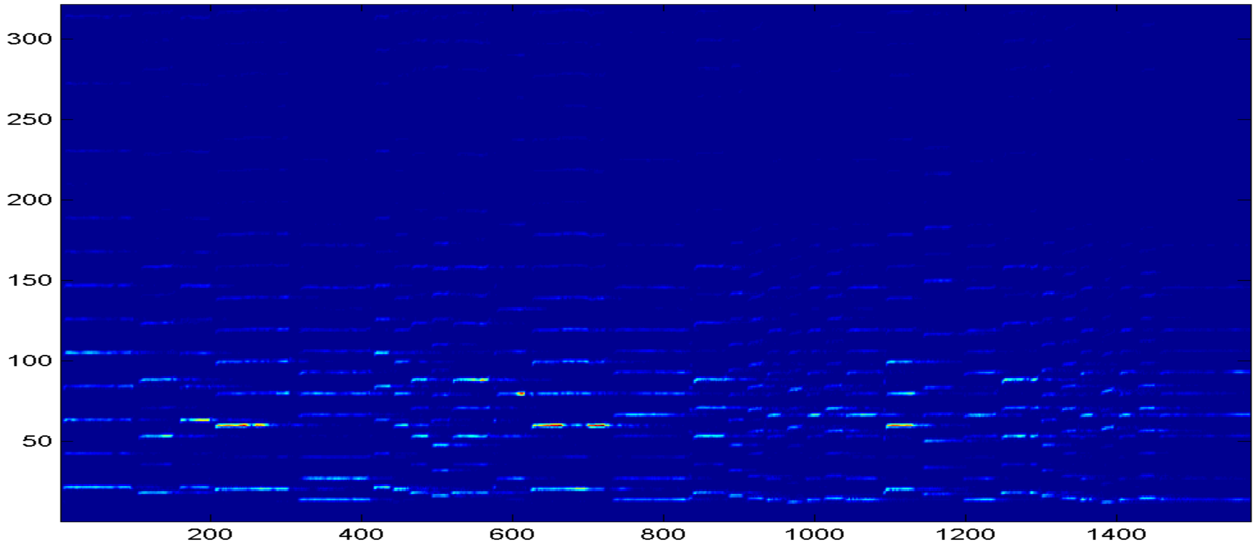
$$\mathbf{H} = \arg\min_{\overline{\mathbf{H}}} \| \mathbf{M} - \mathbf{M}\overline{\mathbf{H}}^T\overline{\mathbf{H}} \|_F^2 + \Lambda\left(\overline{\mathbf{H}}\overline{\mathbf{H}}^T\right)$$

- The solution is obtained through Eigen decomposition

$$Correlation(\mathbf{M}^T)\mathbf{H} = \mathbf{H}\Lambda$$

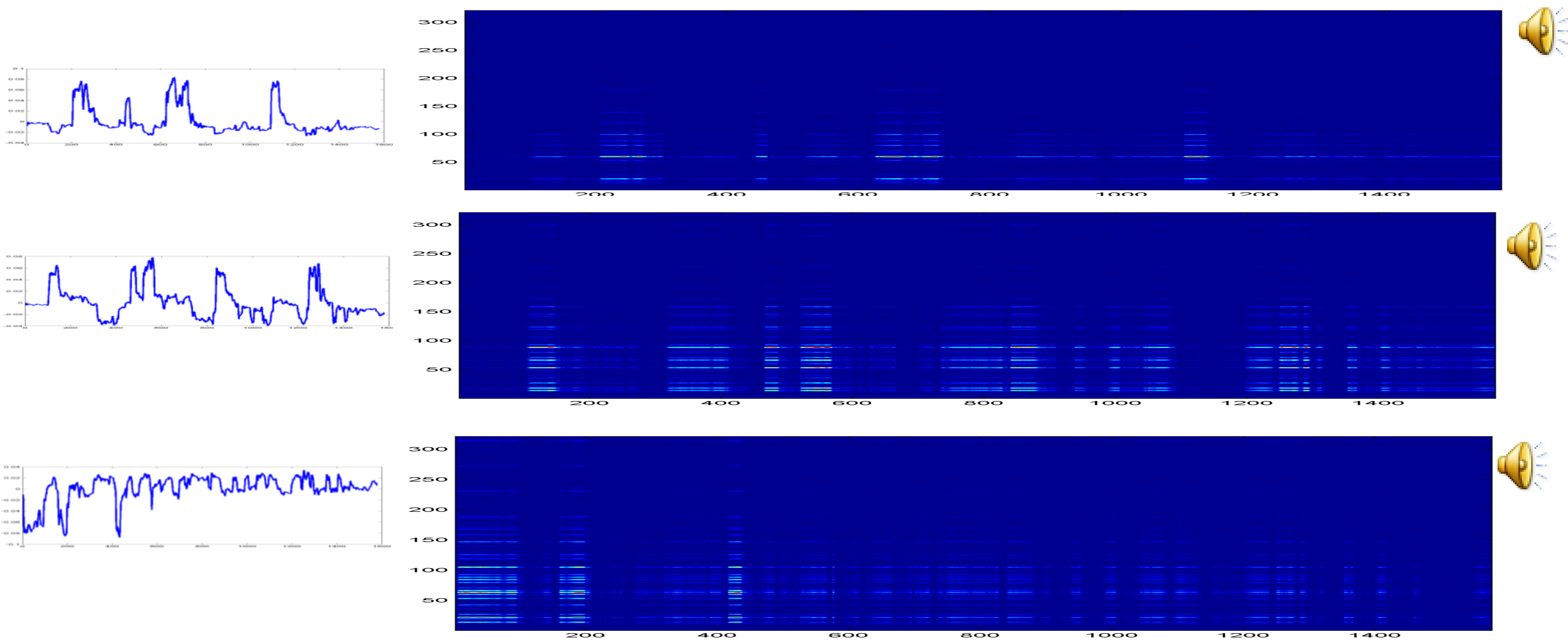
- Note: we are considering the correlation of $\mathbf{M}^{\mathrm{T}}$

So how does that work?



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

So how does that work?



- The scores of the first three “notes” and their contributions

Finding the notes

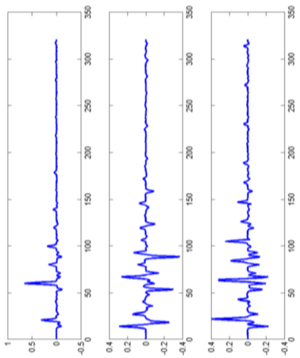
- Can find $\mathbf{W}$ instead of $\mathbf{H}$

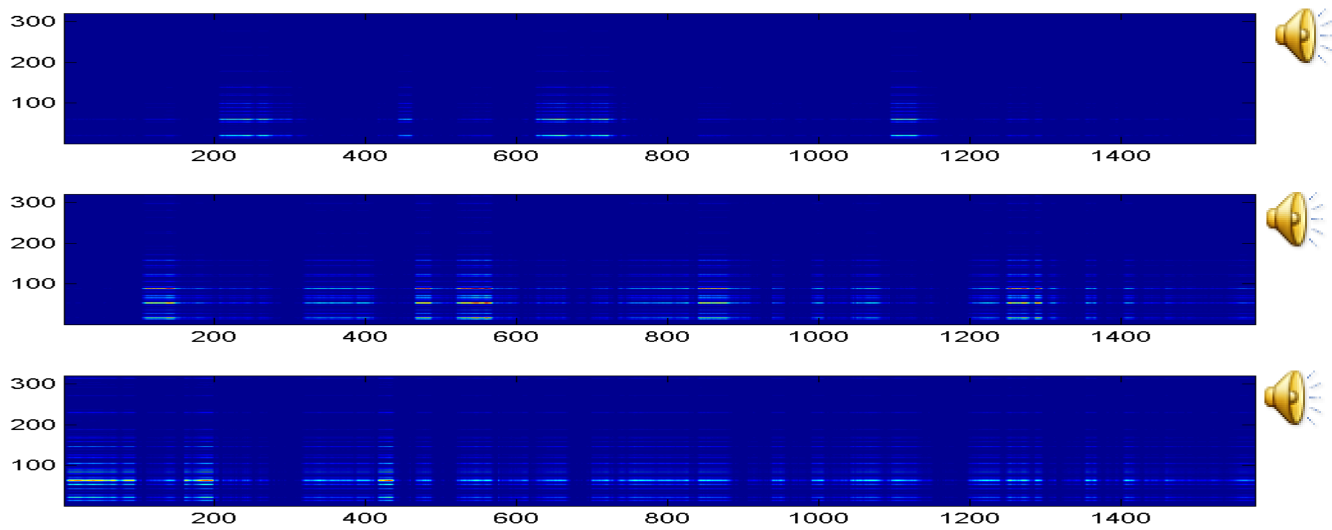
$$\mathbf{W} = \arg\min_{\overline{\mathbf{W}}} \| \mathbf{M} - \overline{\mathbf{W}}^T \overline{\mathbf{W}} \mathbf{M} \|_F^2$$

- Assume the columns of $\mathbf{W}$ are orthogonal
- This results in the more conventional Eigen decomposition

$$Correlation(\mathbf{M})\mathbf{W} = \mathbf{W}\Lambda$$

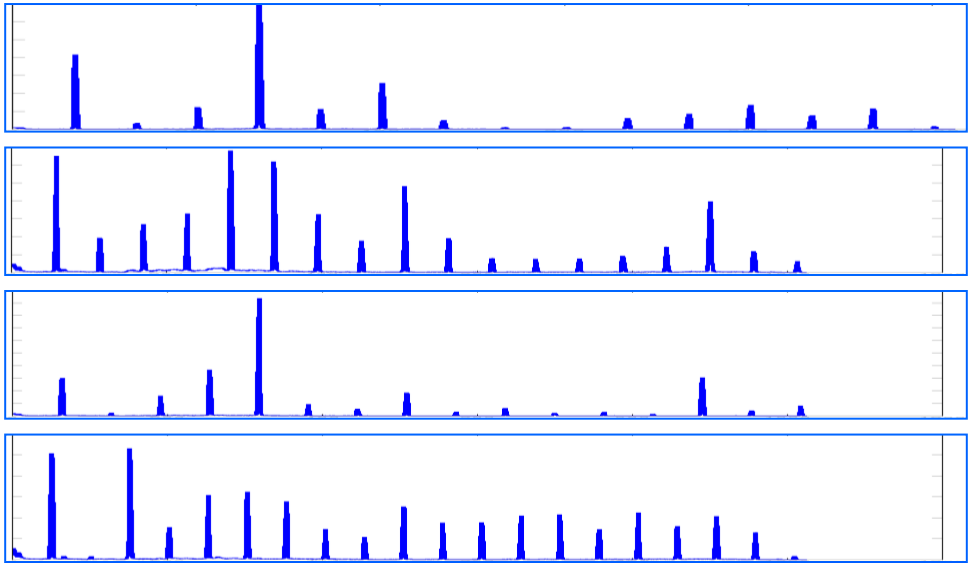
So how does that work?





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

Our notes are not orthogonal



- Overlapping frequencies
- Note occur concurrently
  - Harmonica continues to resonate to previous note
- More generally, simple orthogonality will not give us the desired solution

Eigendecomposition and SVD

$$\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^T \qquad \mathbf{M} = \mathbf{W}\mathbf{H}$$

- Matrix $\mathbf{M}$ can be decomposed as $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$
- When we assume the scores are orthogonal, we get
  $$\mathbf{H} = \mathbf{V}^{\mathrm{T}}, \ \mathbf{W} = \mathbf{U}\mathbf{S}$$
- When we assume the notes are orthogonal, we get
  $$\mathbf{W} = \mathbf{U}, \ \mathbf{H} = \mathbf{S}\mathbf{V}^{\mathrm{T}}$$
- **In either case the results are the same**
  - The notes are orthogonal and so are the scores
  - Not good in our problem

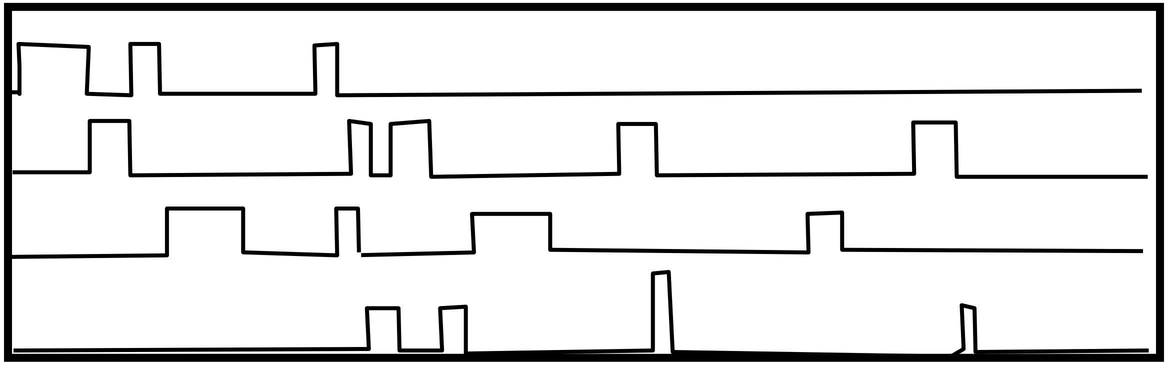
Orthogonality

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

- In any *least-squared error* decomposition $\mathbf{M}=\mathbf{WH}$, if the columns of $\mathbf{W}$ are orthogonal, the rows of $\mathbf{H}$ will also be orthogonal

- Sometimes mere orthogonality is not enough

What *else* can we look for?



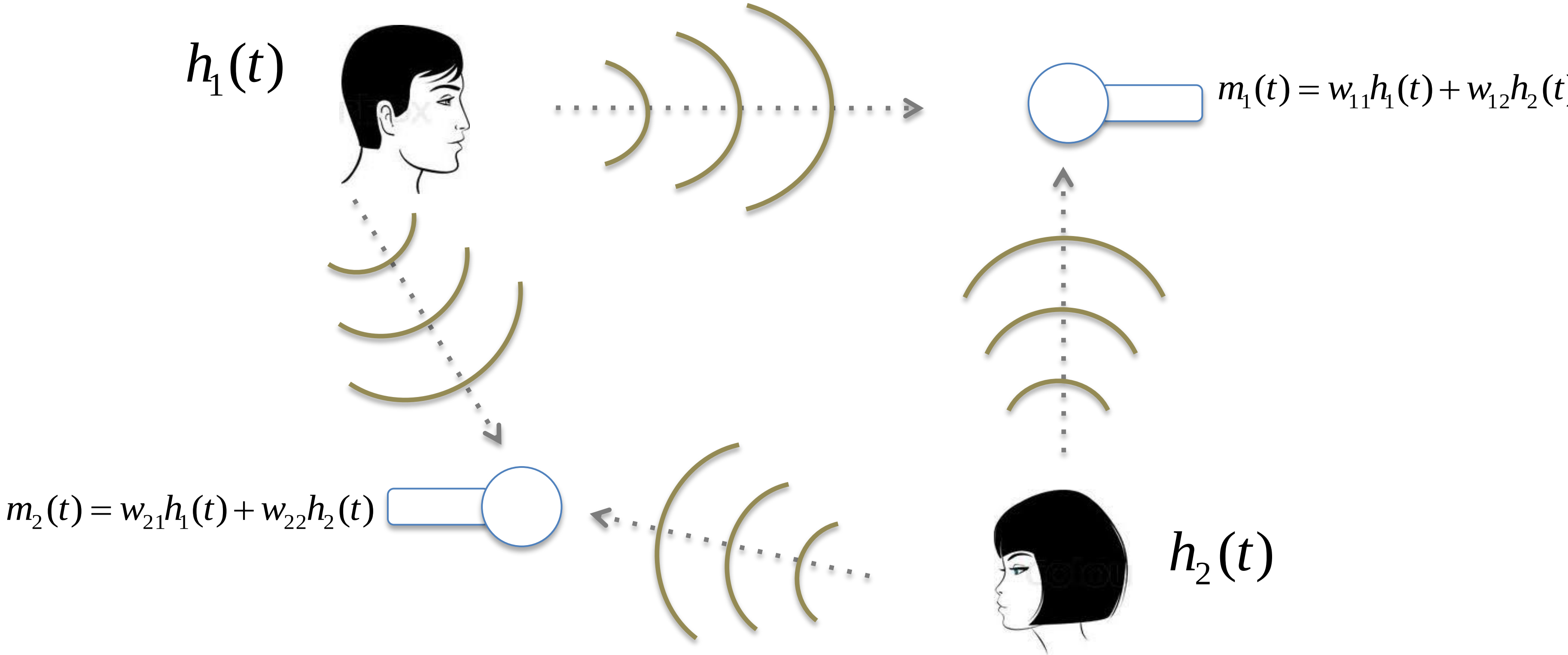
- Assume: The "transcription" of one note does not depend on what else is playing
  - Or, in a multi-instrument piece, instruments are playing independently of one another
- Not strictly true, but still..

Formulating it with Independence

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

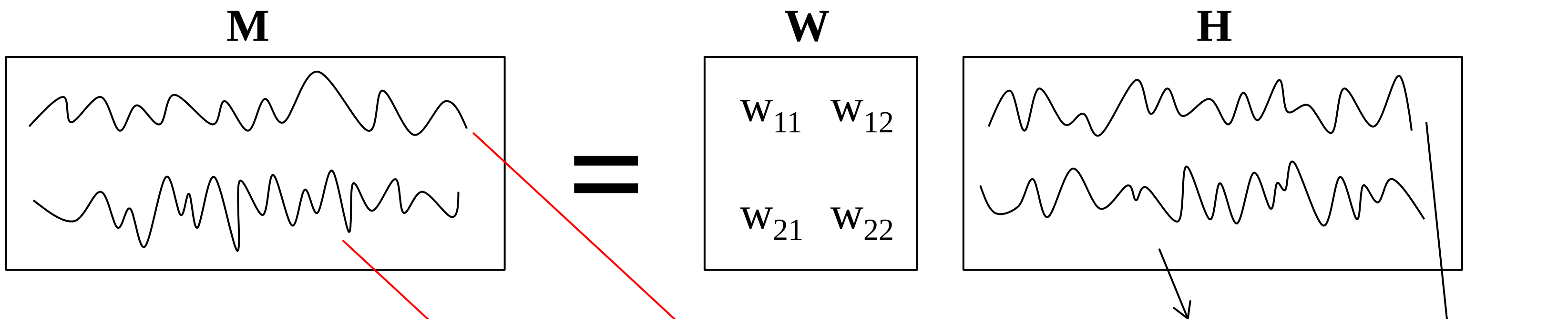
- Impose statistical independence constraints on decomposition

Changing problems for a bit



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

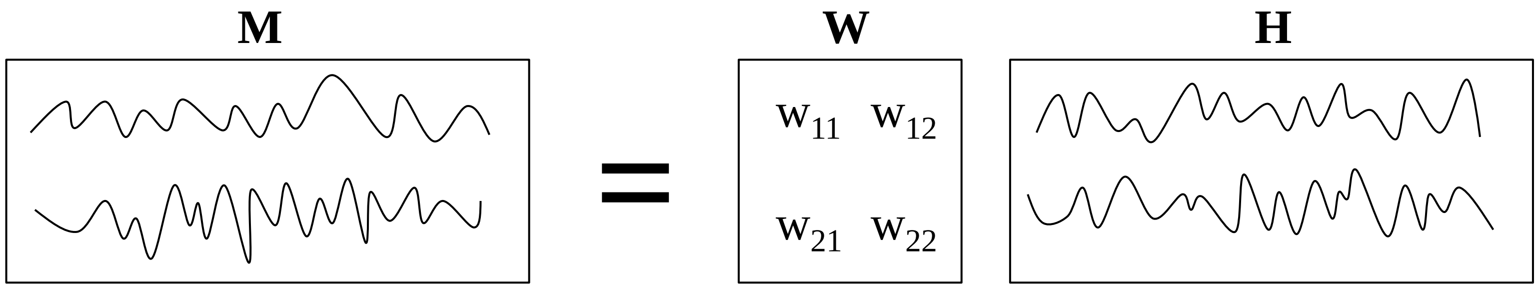
A Separation Problem



- **M** = **WH**
  - **M** = “mixed” signal
  - **W** = “notes”
  - **H** = “transcription”

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

Imposing Statistical Constraints



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

Statistical Independence

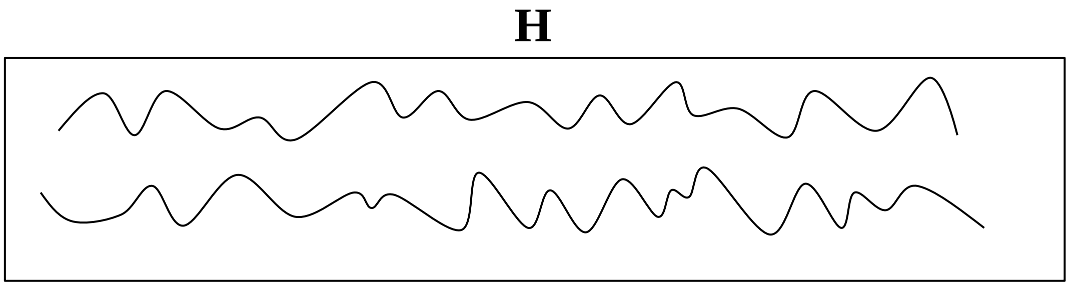
- $\mathbf{M} = \mathbf{W}\mathbf{H}$ $\quad$ $\mathbf{H} = \mathbf{A}\mathbf{M}$

Remember this form

- *Emulating independence*
  - Compute $\mathbf{W}$ (or $\mathbf{A}$) and $\mathbf{H}$ such that $\mathbf{H}$ has statistical characteristics that are observed in statistically independent variables

- *Enforcing independence*
  - Compute $\mathbf{W}$ and $\mathbf{H}$ such that the components of $\mathbf{M}$ are 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 ith and jth 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

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

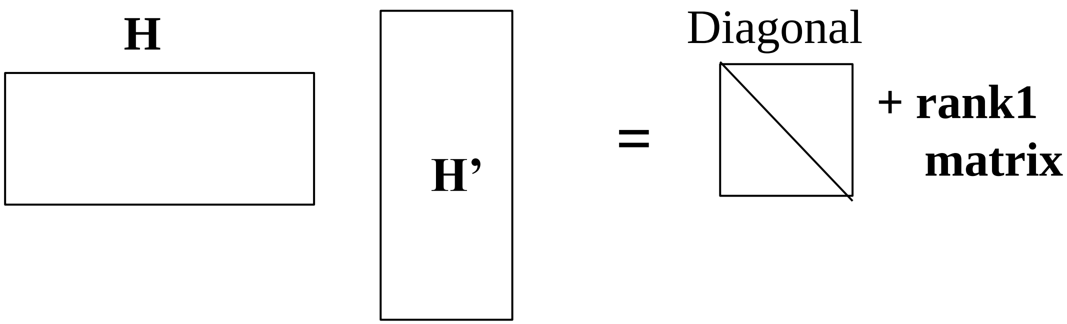
- If mean($\mathbf{M}$) = 0 => mean($\mathbf{H}$) = 0
  - $\mathrm{E}[\mathbf{H}] = \mathbf{A}.\mathrm{E}[\mathbf{M}] = \mathbf{A0} = \mathbf{0}$
  - First step of ICA: Set the mean of $\mathbf{M}$ to 0

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

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

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

Emulating Independence..



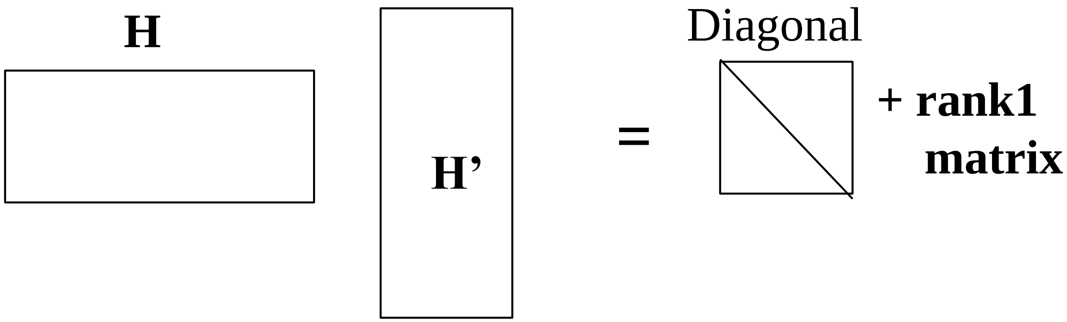
$\mathbf{H=AM}$

$\mathbf{A=BC}$

$\mathbf{H=BCM}$

- Independence → Uncorrelatedness
- Estimate a $\mathbf{C}$ such that $\mathbf{CM}$ is uncorrelated
- $\mathbf{X} = \mathbf{CM}$
  - $E[\mathbf{x}_i\mathbf{x}_j] = E[\mathbf{x}_i]E[\mathbf{x}_j] = \delta_{ij}$ [since M is now "centered"]
  - $\mathbf{XX}^T = \mathbf{I}$
    - In reality, we only want this to be a diagonal matrix, but we'll make it identity

Decorrelating



H=AM

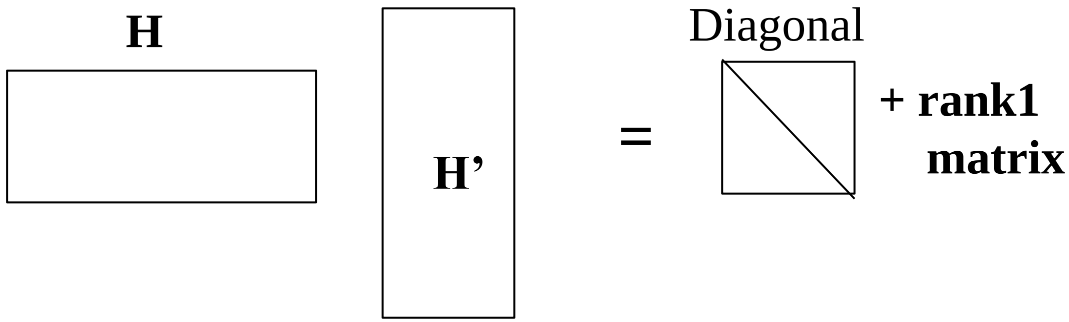
A=BC

H=BCM

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

- Eigen decomposition $\mathbf{MM}^T = \mathbf{ESE}^T$
- Let $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{E}^T$
  - $\mathbf{X} = \mathbf{S}^{-1/2}\mathbf{E}^T\mathbf{M}$
  - $\mathbf{XX}^T = \mathbf{WMM}^T\mathbf{W}^T = \mathbf{S}^{-1/2}\mathbf{E}^T\mathbf{ESE}^T\mathbf{ES}^{-1/2} = \mathbf{I}$

Decorrelating



$\mathbf{H=AM}$

$\mathbf{A=BC}$

$\mathbf{H=BCM}$

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

- Eigen decomposition $\mathbf{MM}^T = \mathbf{ESE}^T$
- Let $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{E}^T$
  - $\mathbf{X} = \mathbf{S}^{-1/2}\mathbf{E}^T\mathbf{M}$
  - $\mathbf{WMM}^T\mathbf{W}^T = \mathbf{S}^{-1/2}\mathbf{E}^T\,\mathbf{ESE}^T\mathbf{ES}^{-1/2} = \mathbf{I}$

- $\mathbf{X}$ is called the ***whitened*** version of $\mathbf{M}$
  - The process of decorrelating $\mathbf{M}$ is called *whitening*
  - $\mathbf{C}$ is the *whitening matrix*

Uncorrelated != Independent

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

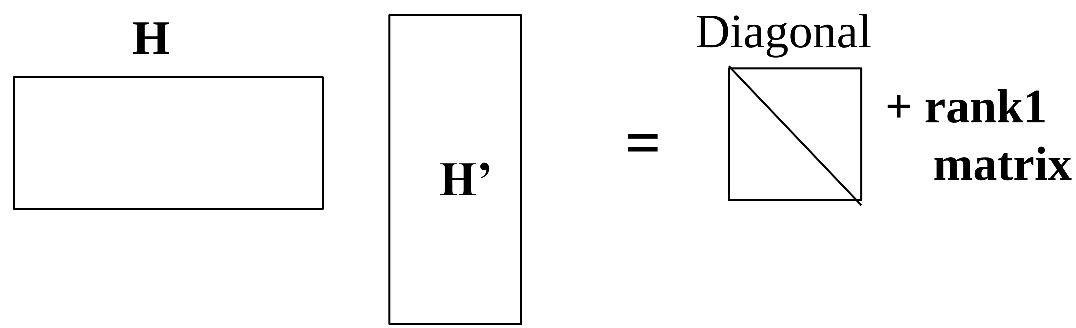
$$E[\mathbf{x}_i \mathbf{x}_j] = 0 \text{ if } i \mathrel{!=} j$$

- 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



H=AM

A=BC

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

ICA: Freeing Fourth Moments

- We have $E[\mathbf{x}_i\ \mathbf{x}_j] = 0$ if i != j
  - Already been decorrelated

- $\mathbf{A}=\mathbf{BC}$, $\mathbf{H} = \mathbf{BCM}$, $\mathbf{X} = \mathbf{CM}$, → $\mathbf{H} = \mathbf{BX}$

- 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

ICA: Freeing Fourth Moments

- Create a matrix of fourth moment terms that would be diagonal were the rows of $\mathbf{H}$ independent and diagonalize it
- A good candidate
  - Good because it incorporates the energy in all rows of $\mathbf{H}$

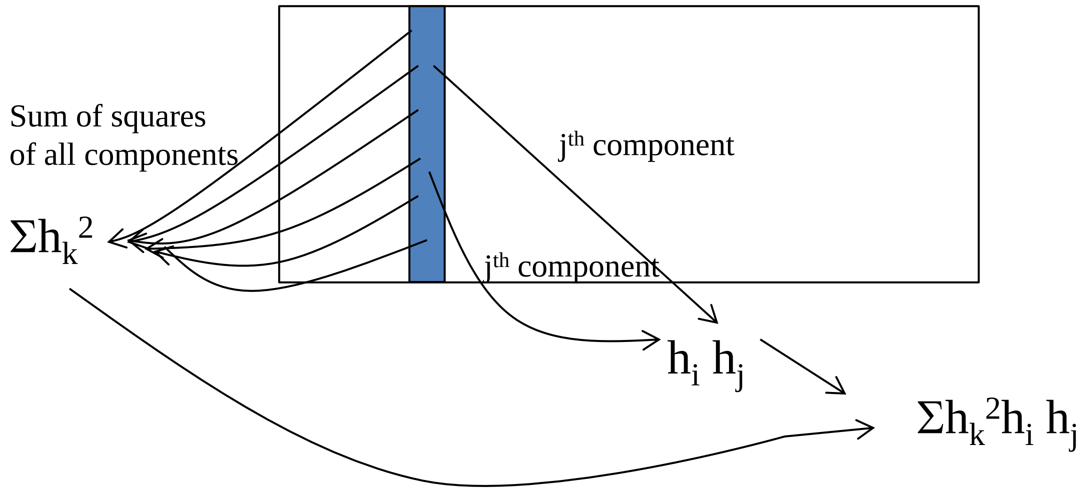
$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & .. \\ d_{21} & d_{22} & d_{23} & .. \\ .. & .. & .. & .. \end{bmatrix}$$

  - Where
    $d_{ij} = \mathrm{E}[\ \Sigma_k\ \mathbf{h}_k^2\ \mathbf{h}_i\ \mathbf{h}_j]$
  - i.e.
    $D = \mathrm{E}[\mathbf{h}^T\mathbf{h}\ \mathbf{h}\ \mathbf{h}^T]$
    - $\mathbf{h}$ are the columns of $\mathbf{H}$
    - Assuming $\mathbf{h}$ is real, else replace transposition with Hermition

ICA: The D matrix

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

$$d_{ij} = E[\Sigma_k h_k^2 h_i h_j] = \frac{1}{cols(\mathbf{H})} \sum_m \sum_k h_{mk}^2 h_{mi} h_{mj}$$



- Average above term across all columns of **H**

ICA: The D matrix

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

$$d_{ij} = E[\Sigma_k \mathbf{h}_k^2 \mathbf{h}_i \mathbf{h}_j] = \frac{1}{cols(\mathbf{H})} \sum_m \sum_k h_{mk}^2 h_{mi} h_{mj}$$

- If the $\mathbf{h}_i$ terms were independent
  - For i != j

$$E\left[\sum_k \mathbf{h}_k^2 \mathbf{h}_i \mathbf{h}_j\right] = E[\mathbf{h}_i^3]E[\mathbf{h}_j] + E[\mathbf{h}_j^3]E[\mathbf{h}_i] + \sum_{k \neq i, k \neq j} E[\mathbf{h}_k^2]E[\mathbf{h}_i]E[\mathbf{h}_j]$$

  - Centered: $E[\mathbf{h}_j] = 0$ ➔ $E[\Sigma_k \mathbf{h}_k^2 \mathbf{h}_i \mathbf{h}_j] = 0$ for i != j
  - For i = j

$$E\left[\sum_k \mathbf{h}_k^2 \mathbf{h}_i \mathbf{h}_j\right] = E[\mathbf{h}_i^4] + E[\mathbf{h}_i^2] \sum_{k \neq i} E[\mathbf{h}_k^2] \neq 0$$

- Thus, if the $\mathbf{h}_i$ terms were independent, $d_{ij} = 0$ if i != j
- i.e., if $\mathbf{h}_i$ were independent, D would be a diagonal matrix
  - **Let us diagonalize *D***

Diagonalizing D

- Compose a fourth order matrix from **X**
  - Recall: $\mathbf{X} = \mathbf{CM}$, $\mathbf{H} = \mathbf{BX} = \mathbf{BCM}$
    - **B** is what we're trying to learn to make **H** independent

  - Compose $\mathbf{D}' = E[\mathbf{x}^T \mathbf{x}\, \mathbf{x}\, \mathbf{x}^T]$

- Diagonalize **D'** via Eigen decomposition

  $\mathbf{D}' = \mathbf{U\Lambda U}^T$

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

B frees the fourth moment

$$\mathbf{D'} = \mathbf{U\Lambda U}^T \;;\quad \mathbf{B} = \mathbf{U}^T$$

- $\mathbf{U}$ is a unitary matrix, i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$ (identity)
- $\mathbf{H} = \mathbf{BX} = \mathbf{U}^T\mathbf{X}$
- $\mathbf{h} = \mathbf{U}^T\mathbf{x}$
- The fourth moment matrix of H is
  $E[\mathbf{h}^T\,\mathbf{h}\;\mathbf{h}\;\mathbf{h}^T] = E[\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x}\;\mathbf{U}^T\,\mathbf{x}\;\mathbf{x}^T\mathbf{U}]$
  $= E[\mathbf{x}^T\mathbf{x}\;\mathbf{U}^T\,\mathbf{x}\;\mathbf{x}^T\mathbf{U}]$
  $= \mathbf{U}^T\,E[\mathbf{x}^T\mathbf{x}\;\mathbf{x}\mathbf{x}^T]\mathbf{U}$
  $= \mathbf{U}^T\,\mathbf{D'}\,\mathbf{U}$
  $= \mathbf{U}^T\,\mathbf{U}\,\Lambda\,\mathbf{U}^T\,\mathbf{U} = \Lambda$
- The fourth moment matrix of $\mathbf{H} = \mathbf{U}^T\mathbf{X}$ is Diagonal!!

Overall Solution

- $\mathbf{H} = \mathbf{AM} = \mathbf{BCM}$
  - $\mathbf{C}$ is the (transpose of the) matrix of Eigen vectors of $\mathbf{MM}^T$
- $\mathbf{X} = \mathbf{CM}$
- $\mathbf{A} = \mathbf{BC} = \mathbf{U}^T\mathbf{C}$
  - $\mathbf{B}$ is the (transpose of the) matrix of Eigenvectors of $\mathbf{X}.\mathit{diag}(\mathbf{X}^T\mathbf{X}).\mathbf{X}^T$

Independent Component Analysis

- Goal: to derive a matrix **A** such that the rows of **AM** are independent
- Procedure:
  1. "Center" **M**
  2. Compute the autocorrelation matrix $\boldsymbol{R}_{MM}$ of **M**
  3. Compute whitening matrix **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 $\mathbf{D}' = E[\mathbf{x}^{\mathrm{T}}\mathbf{xxx}^{\mathrm{T}}]$
  6. Diagonalize **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 **H=AM** is diagonal
  - **Note that the autocorrelation matrix of H will also be diagonal**

ICA by diagonalizing moment matrices

- The procedure just outlined, while fully functional, has shortcomings
  - Only a subset of fourth order moments are considered
  - There are many other ways of constructing fourth-order moment matrices that would ideally be diagonal
    - 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 several fourth-order moment matrices
  - More effective than the procedure shown, but computationally more expensive

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”
  - Normalize variance along all directions
  - Eliminate second-order dependence
- Eigen decomposition $\mathbf{M}\mathbf{M}^T = \mathbf{E}\mathbf{S}\mathbf{E}^T$
- $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{E}^T$
- Can use *first K* columns of $\mathbf{E}$ only if only K independent sources are expected
  - In microphone array setup – only K < M sources
- $\mathbf{X} = \mathbf{C}\mathbf{M}$
  - $E[\mathbf{x}_i\mathbf{x}_j] = E[\mathbf{x}_i]E[\mathbf{x}_j] = \delta_{ij}$ 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 : $\mathbf{H} = \mathbf{BX}$
  - $\mathbf{X}$ is "whitened" $\mathbf{M}$

Linear Functions

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

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

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

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

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 uncorrelated
- Leads to: $\min_{\mathbf{W}} ||\mathbf{M} - \mathbf{W}^T\mathbf{WM}||^2 + \Lambda \mathrm{trace}(\mathbf{WW}^T)$
  - 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:
  - $E[f(x)g(y)] = E[f(x)]E[g(y)]$
  - Must hold for *every* *f*() and *g*()!!

An alternate approach

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

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

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

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

An alternate approach

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

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

$$P_{ij} = \sum_k g(h_{ik}) f(h_{jk})$$

This is a square matrix

- Must ideally be

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

$$Q_{ij} = \sum_k g(h_{ik}) \sum_l f(h_{jl}) \qquad i \neq j$$

$$Q_{ii} = \sum_k g(h_{ik}) f(h_{il})$$

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

An alternate approach

- Ideal value for **Q**

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

$$Q_{ij} = \sum_k g(h_{ik}) \sum_l f(h_{jl}) \qquad i \neq j$$

$$Q_{ii} = \sum_k g(h_{ik}) f(h_{il})$$

- If g() and h() are odd symmetric functions
  $\Sigma_j g(h_{ij}) = 0$ for all i
  - Since = $\Sigma_j h_{ij} = 0$ (**H** is centered)
  - **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:

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

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

Update Rules

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

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
  - $\Delta\mathbf{B} = (\mathbf{I} - \mathbf{g}(\mathbf{H})\mathbf{H}^T)\mathbf{W}$
- Cichoki-Unbehaeven
  - $\Delta\mathbf{B} = (\mathbf{I} - \mathbf{g}(\mathbf{H})\mathbf{f}(\mathbf{H})^T)\mathbf{W}$

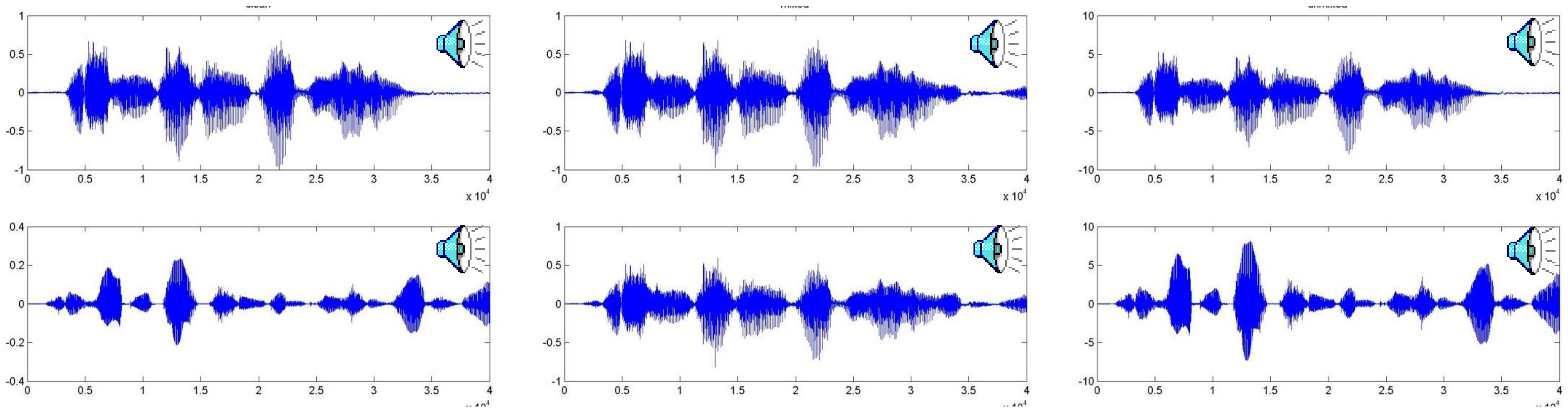
What are G() and H()

- 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
  - $\Delta \mathbf{B} = (\mathbf{I} - \mathbf{H}\mathbf{H}^T - \mathbf{K}\tanh(\mathbf{H})\mathbf{H}^T)\mathbf{W}$
- Or simply
  - $\Delta \mathbf{B} = (\mathbf{I} - \mathbf{K}\tanh(\mathbf{H})\mathbf{H}^T)\mathbf{W}$

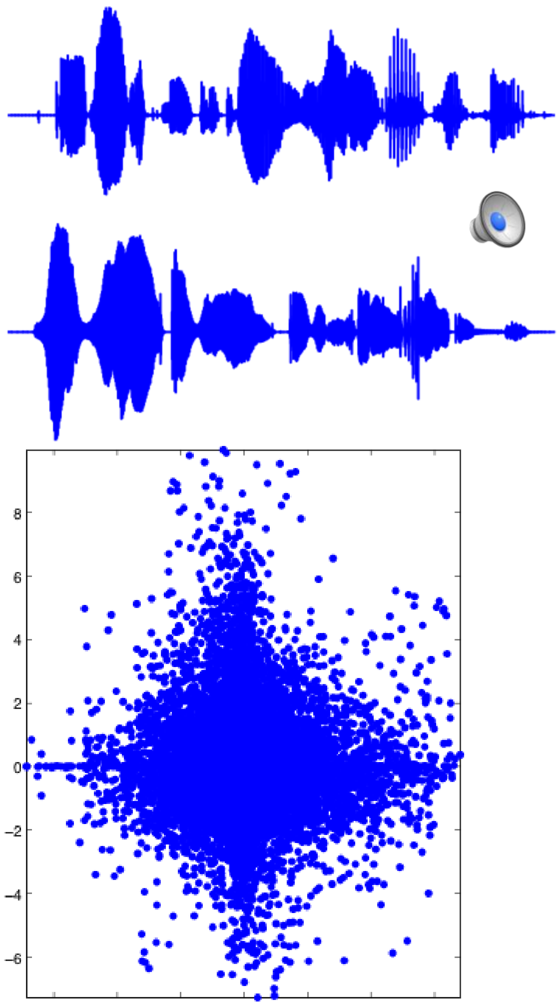
So how does it work?



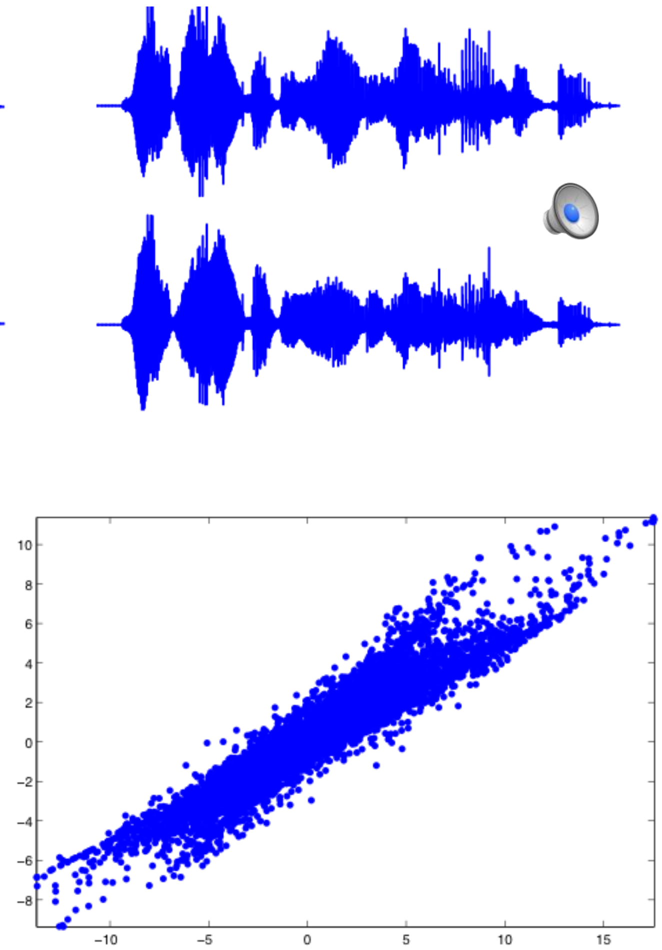
- Example with instantaneous mixture of two speakers
- Natural gradient update
- Works very well!

Another example!

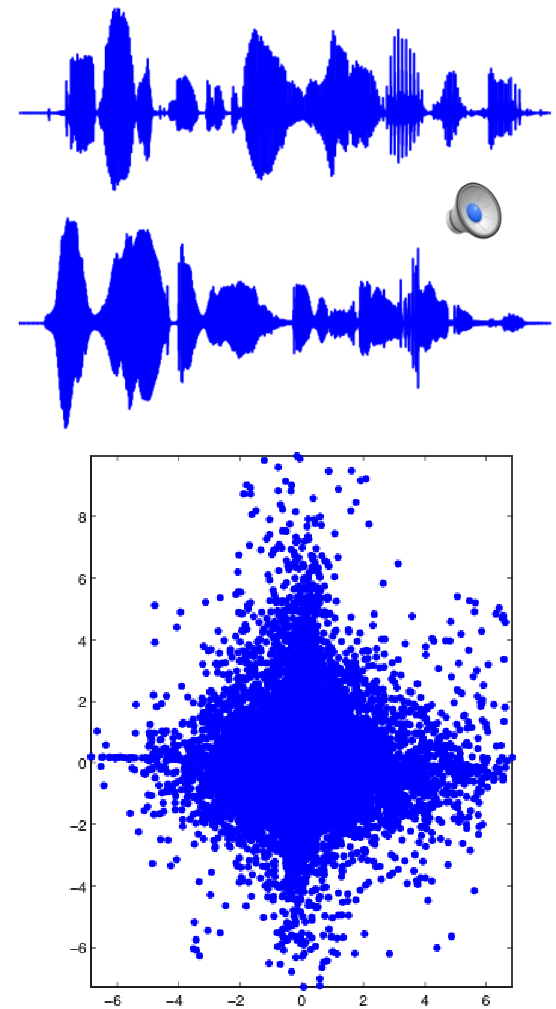
Input



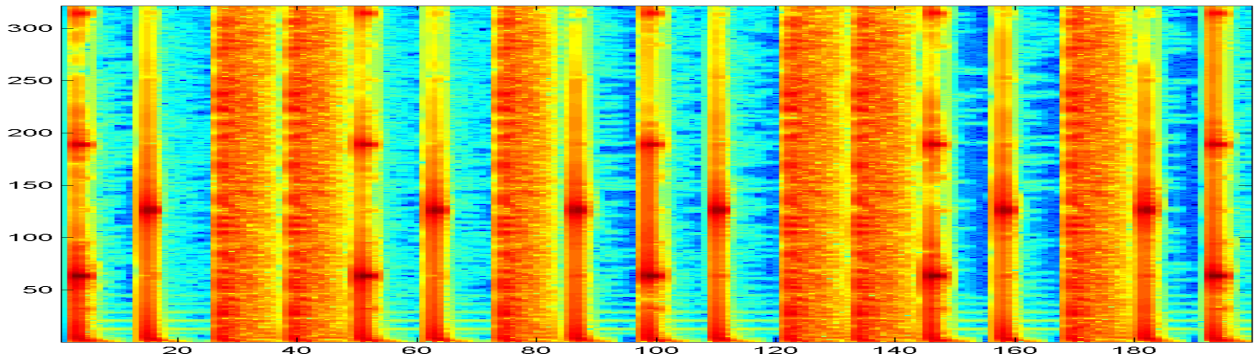
Mix



Output

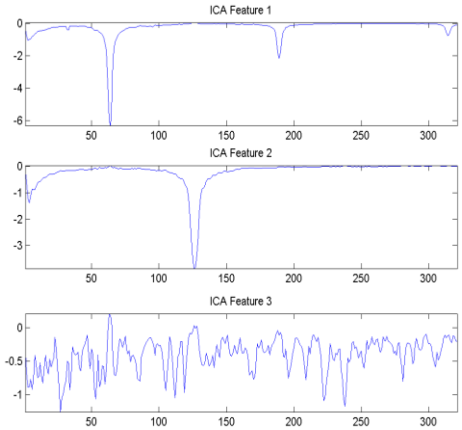


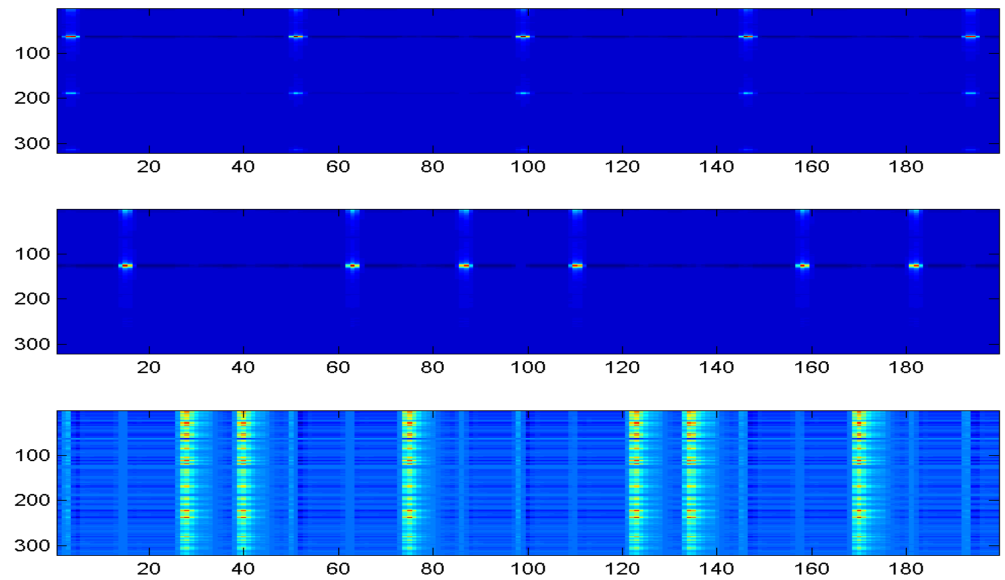
Another Example



- Three instruments..

The Notes

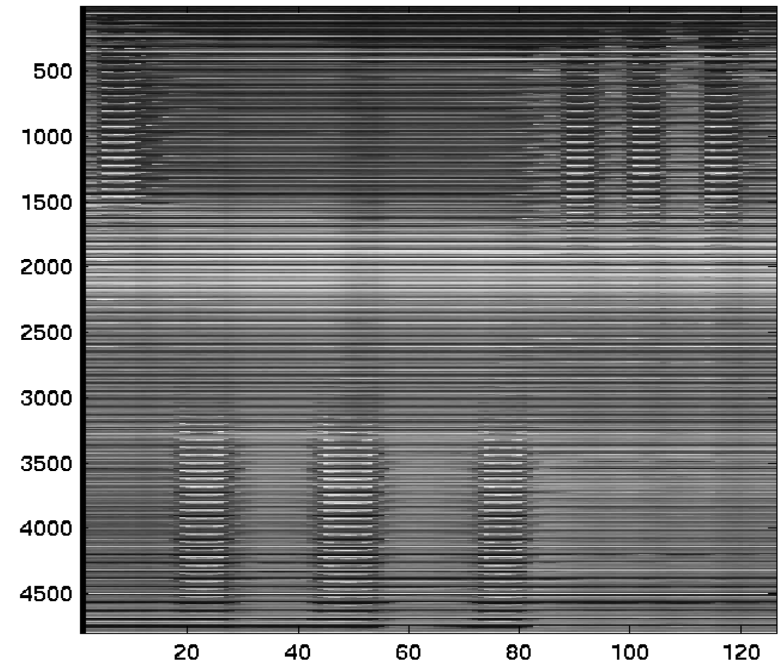




- Three instruments..

ICA for data exploration

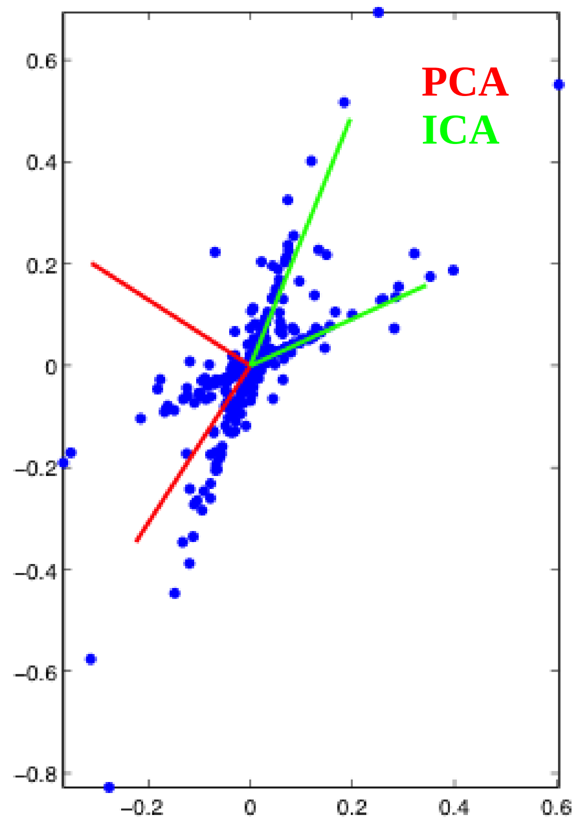
- The “bases” in PCA represent the “building blocks”
  - Ideally notes
- Very successfully used
- So can ICA be used to do the same?



ICA vs PCA bases

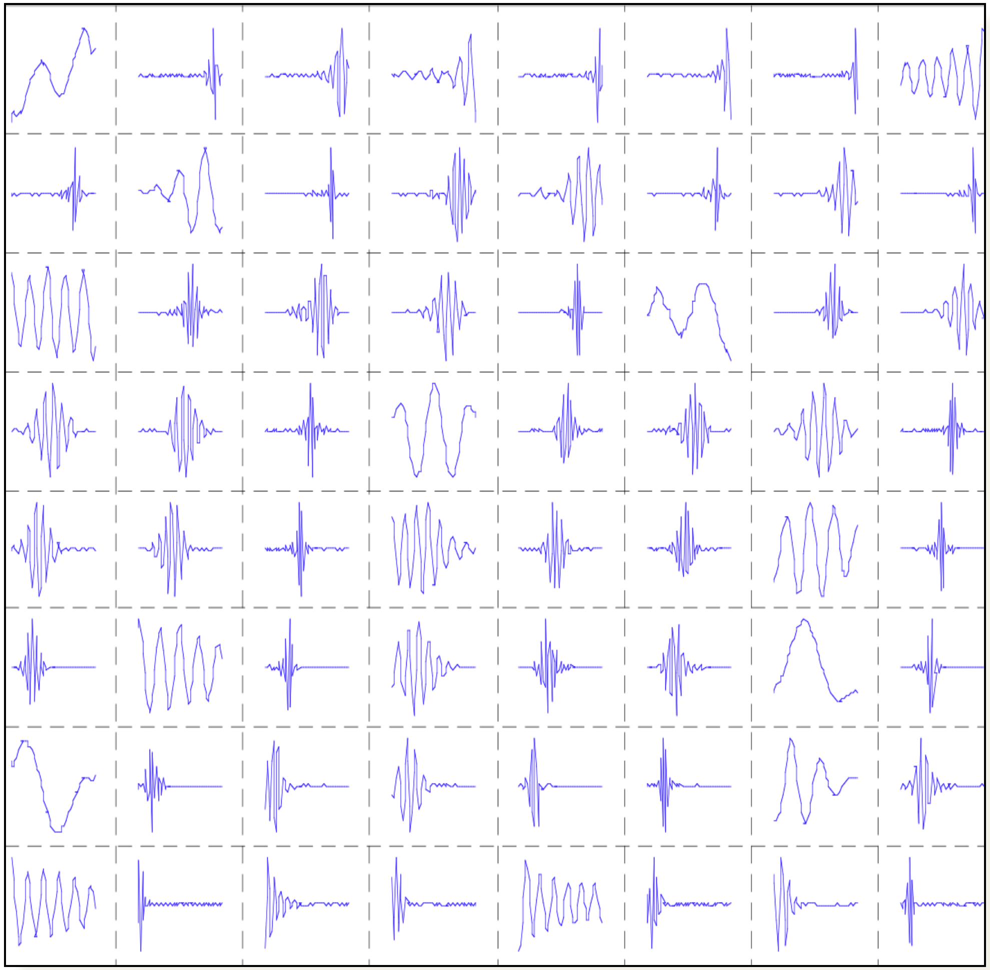
- Motivation for using ICA vs PCA
- PCA will indicate orthogonal directions of maximal variance
  - May not align with the data!
- ICA finds directions that are independent
  - More likely to "align" with the data

Non-Gaussian data



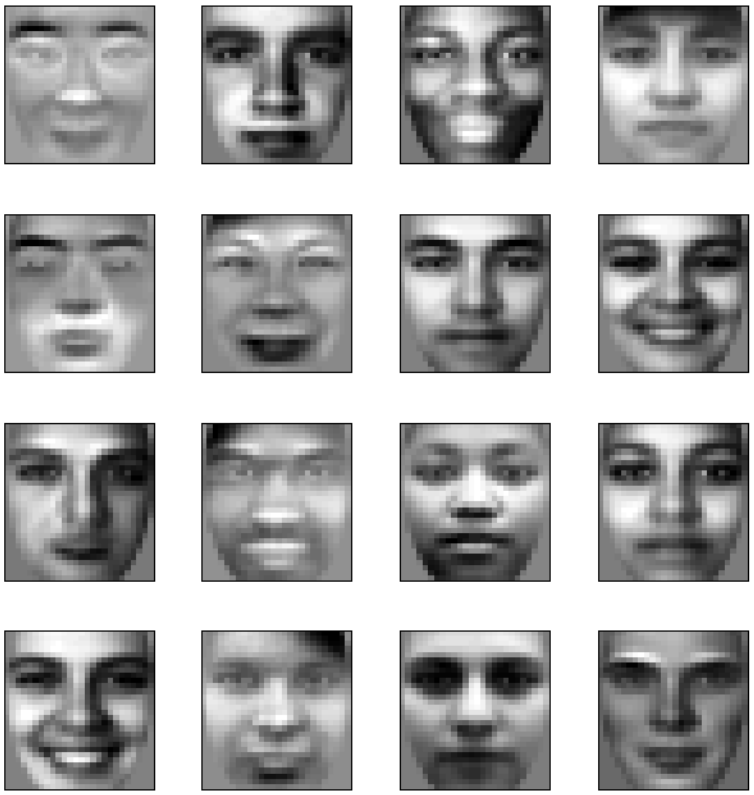
Finding useful transforms with ICA

- 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

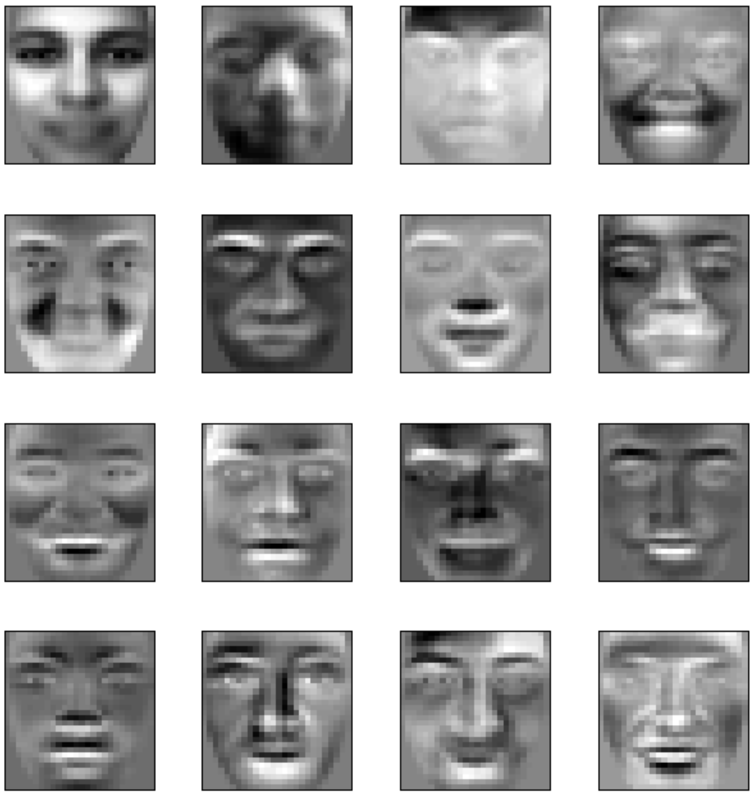


Example case: ICA-faces vs. Eigenfaces

ICA-faces



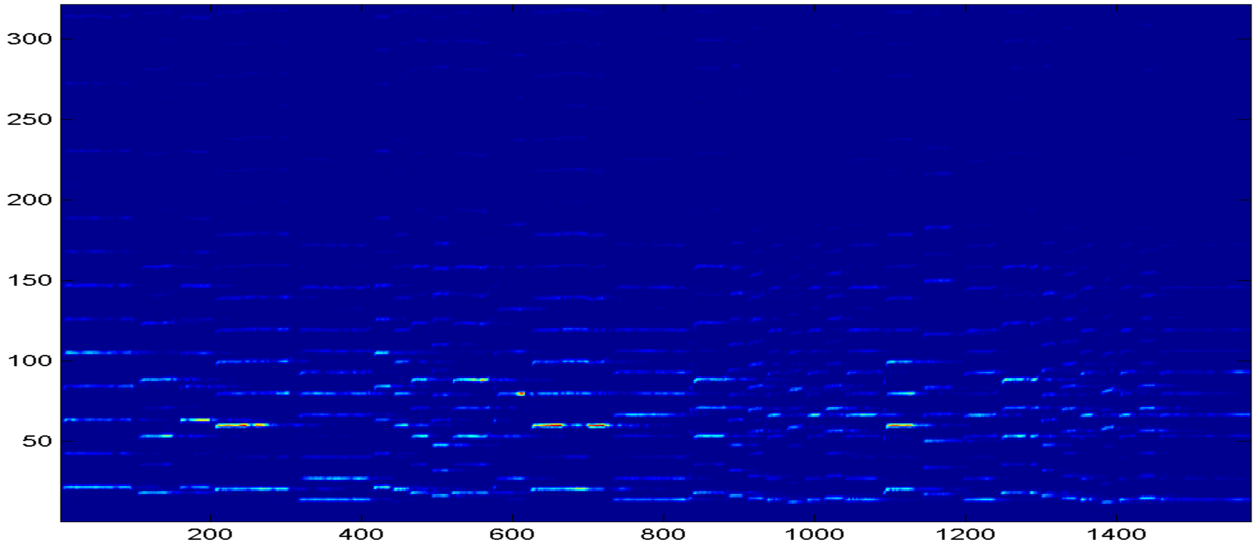
Eigenfaces



ICA for Signal Enhncement

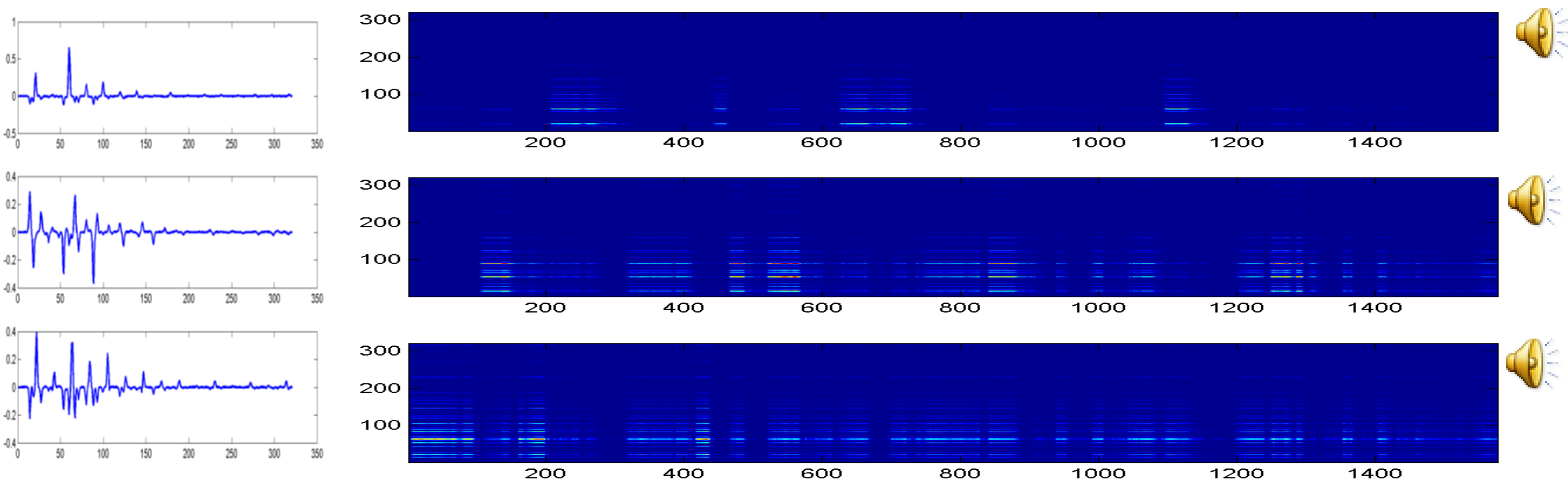
- Very commonly used to enhance EEG signals
- EEG signals are frequently corrupted by heartbeats and biorhythm signals
- ICA can be used to separate them out

So how does that work?



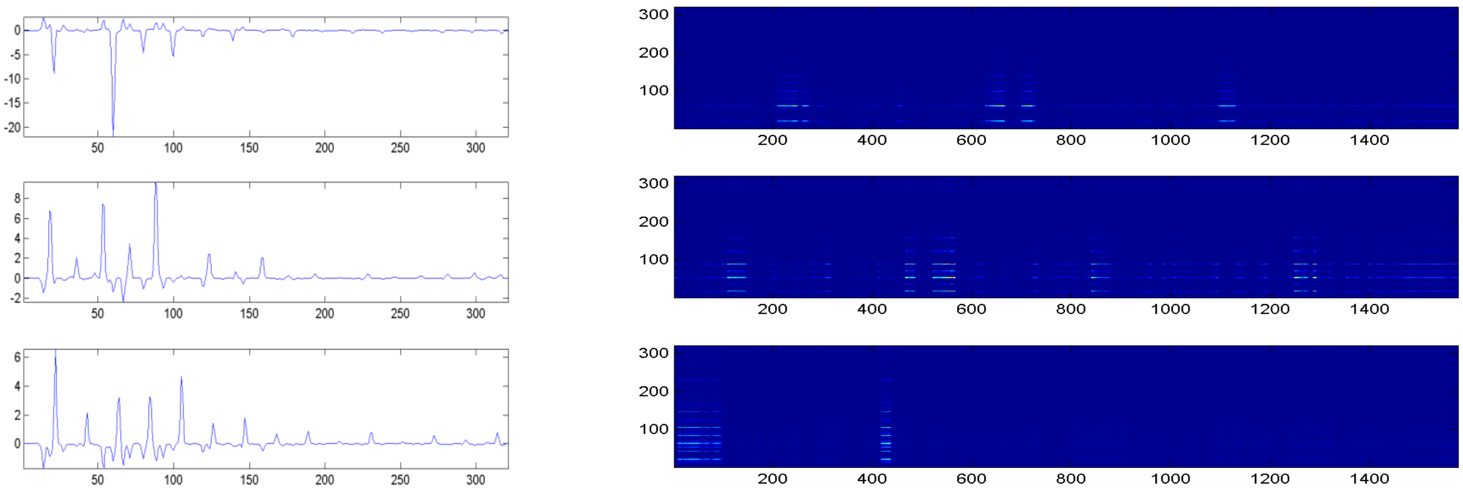
- There are 12 notes in the segment, hence we try to estimate 12 notes..

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..