Independent Component Analysis

Class 20. 8 Nov 2012

Instructor: Bhiksha Raj

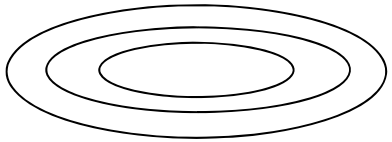
A brief review of basic probability

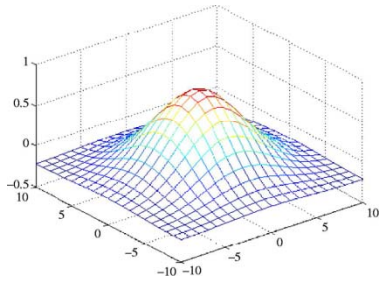
- *Uncorrelated:* Two random variables X and Y are uncorrelated iff:
  - The *average* value of the product of the variables equals the product of their individual averages

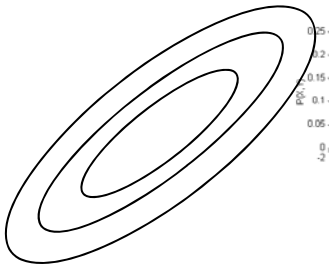
- Setup: Each draw produces one instance of X and one instance of Y
  - I.e one instance of (X,Y)
- E[XY] = E[X]E[Y]

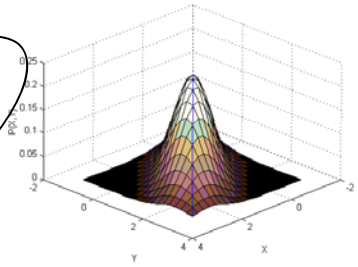
- The average value of X is the same regardless of the value of Y

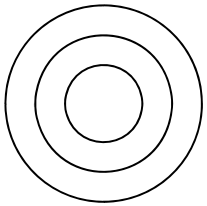
Uncorrelatedness

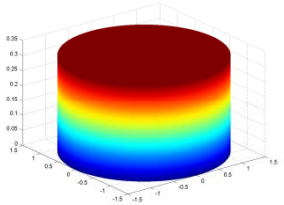


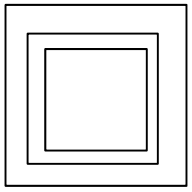


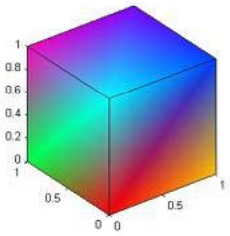












- Which of the above represent uncorrelated RVs?

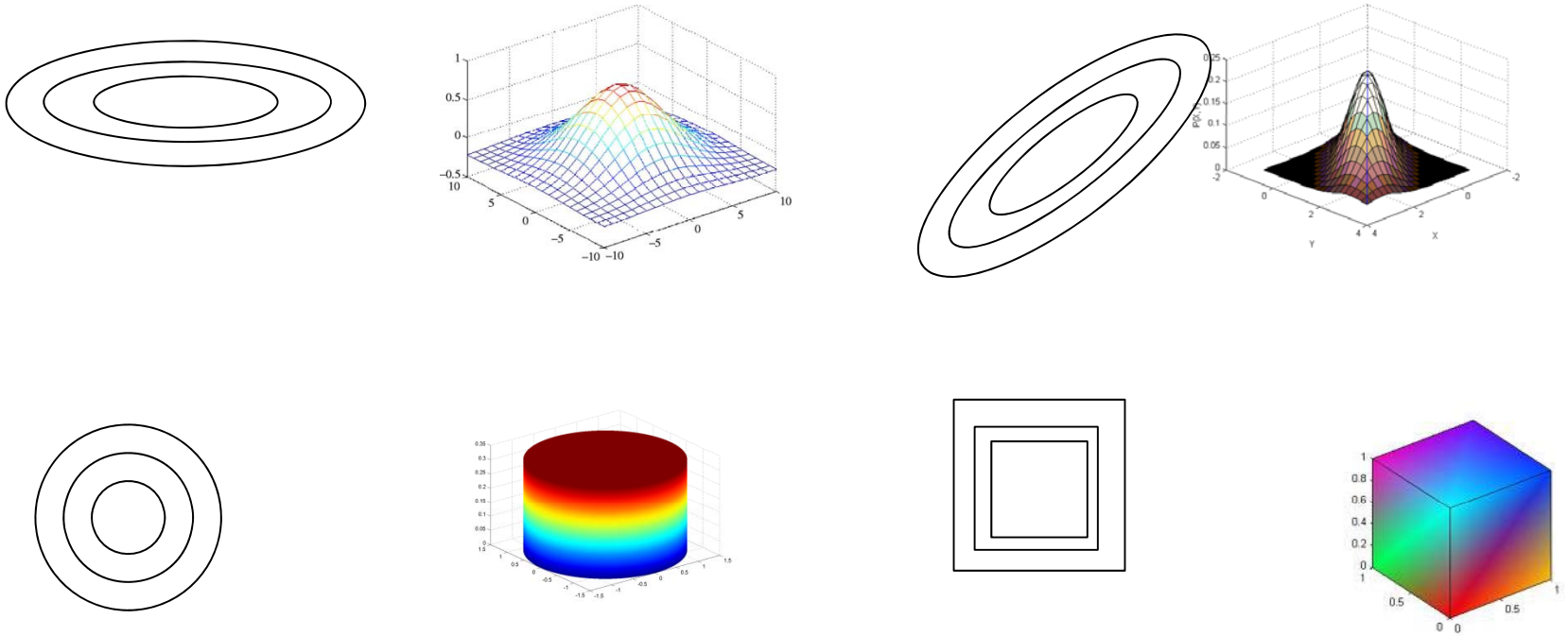
A brief review of basic probability

- *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)
- ➔ The average value of X is the same regardless of the value of Y
  - E[X|Y] = E[X]

A brief review of basic probability

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

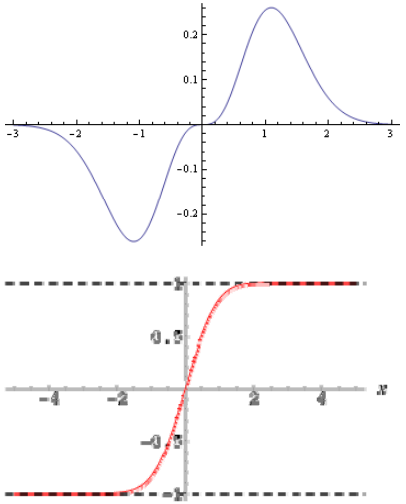
Independence



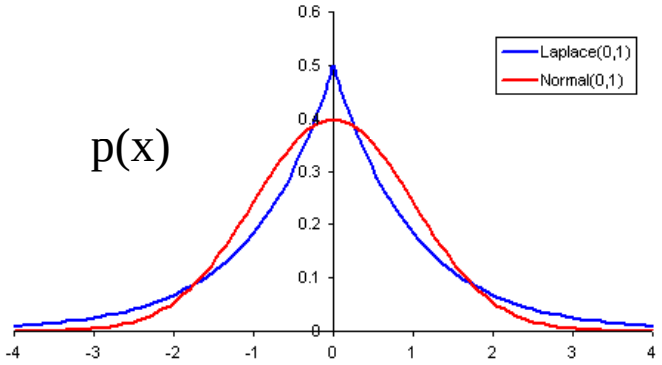
- Which of the above represent independent RVs?
- Which represent uncorrelated RVs?

A brief review of basic probability

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

X

T, M, S…

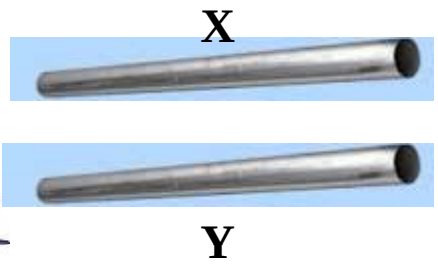
M F F M..

Y

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

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

A brief review of basic info. theory



T, M, S…

M F F M..

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

- 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

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

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

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

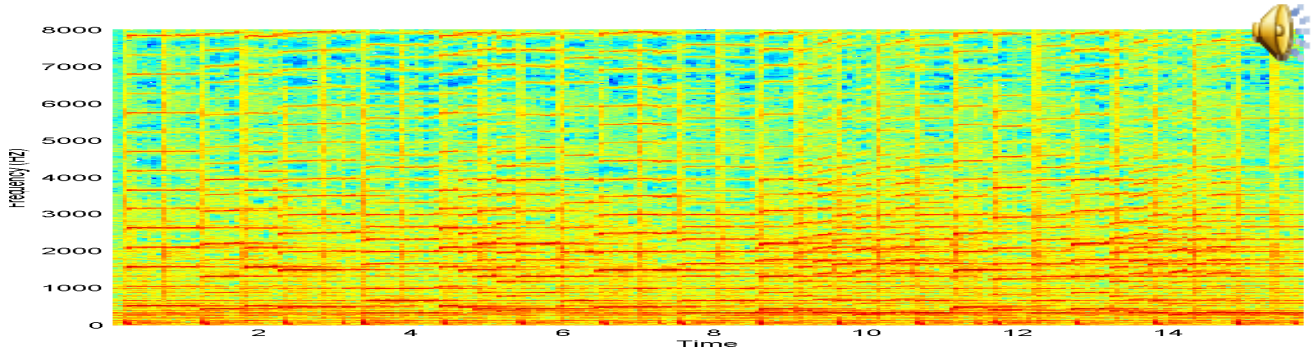
$$= -\sum_{X,Y} P(X,Y) \log P(X) - \sum_{X,Y} P(X,Y) \log P(Y) = H(X) + H(Y)$$

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

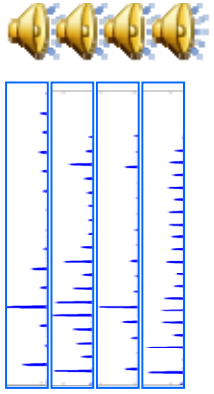
Onward..

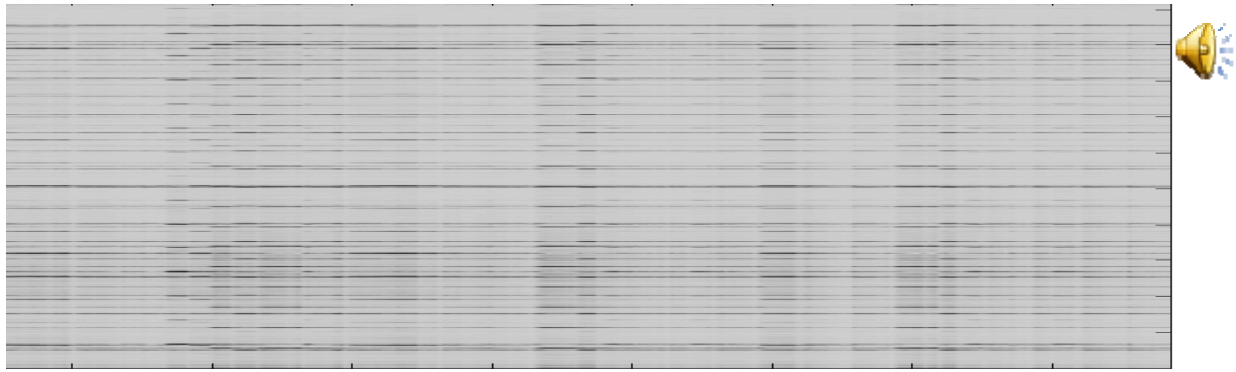
Projection: multiple notes

M =



W =

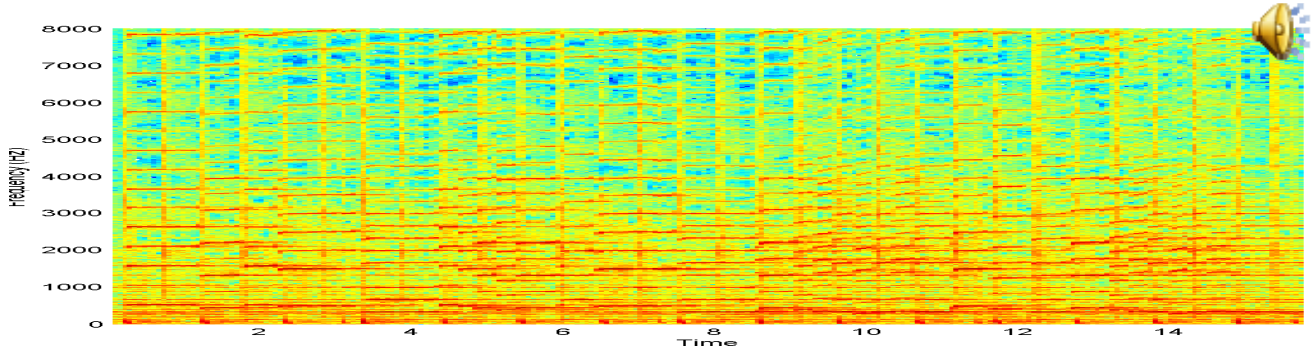




- $P = W (W^T W)^{-1} W^T$
- Projected Spectrogram = P * M

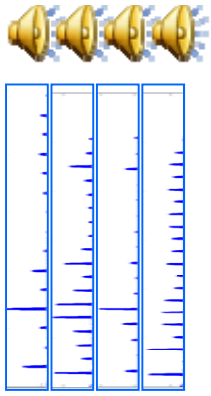
We're actually computing a score

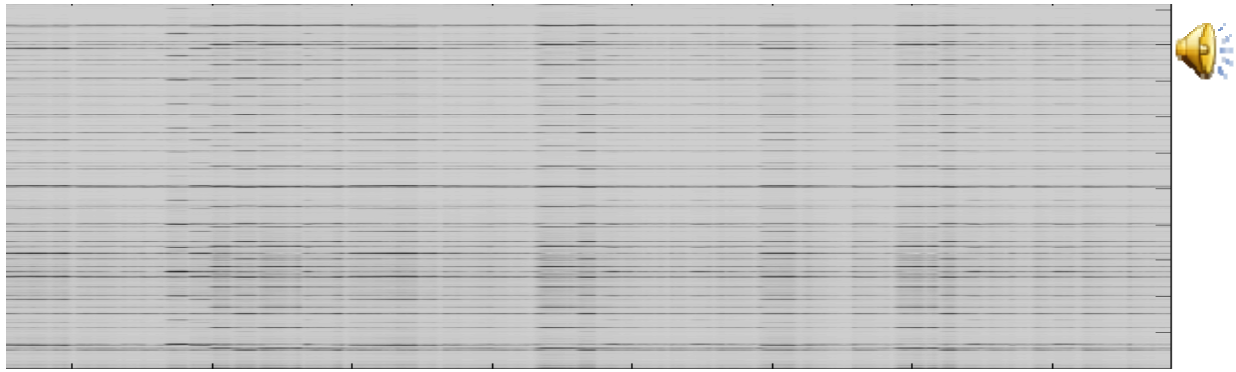
M =



H = ?

W =

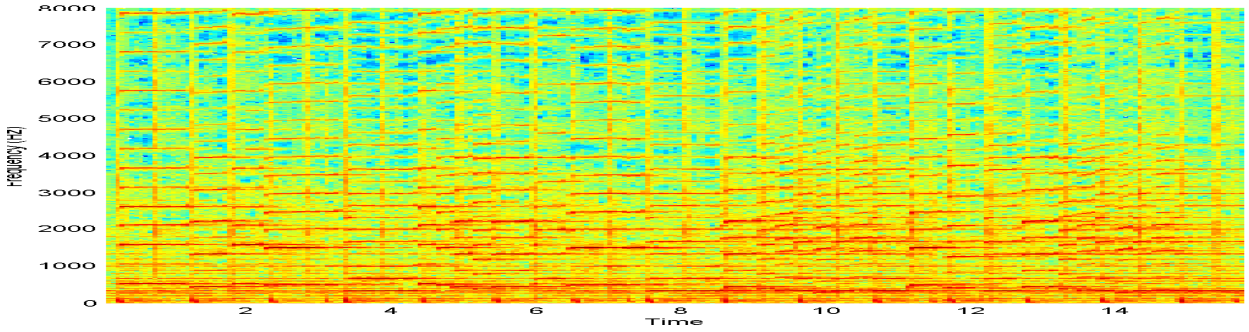




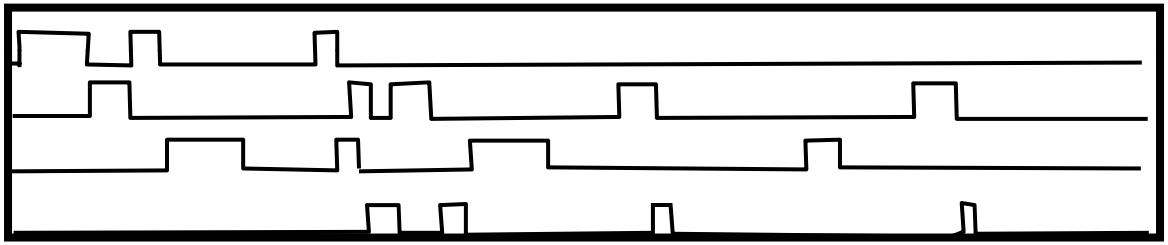
- **M ~ WH**
- **H = pinv(W)M**

How about the other way?

M =



H =

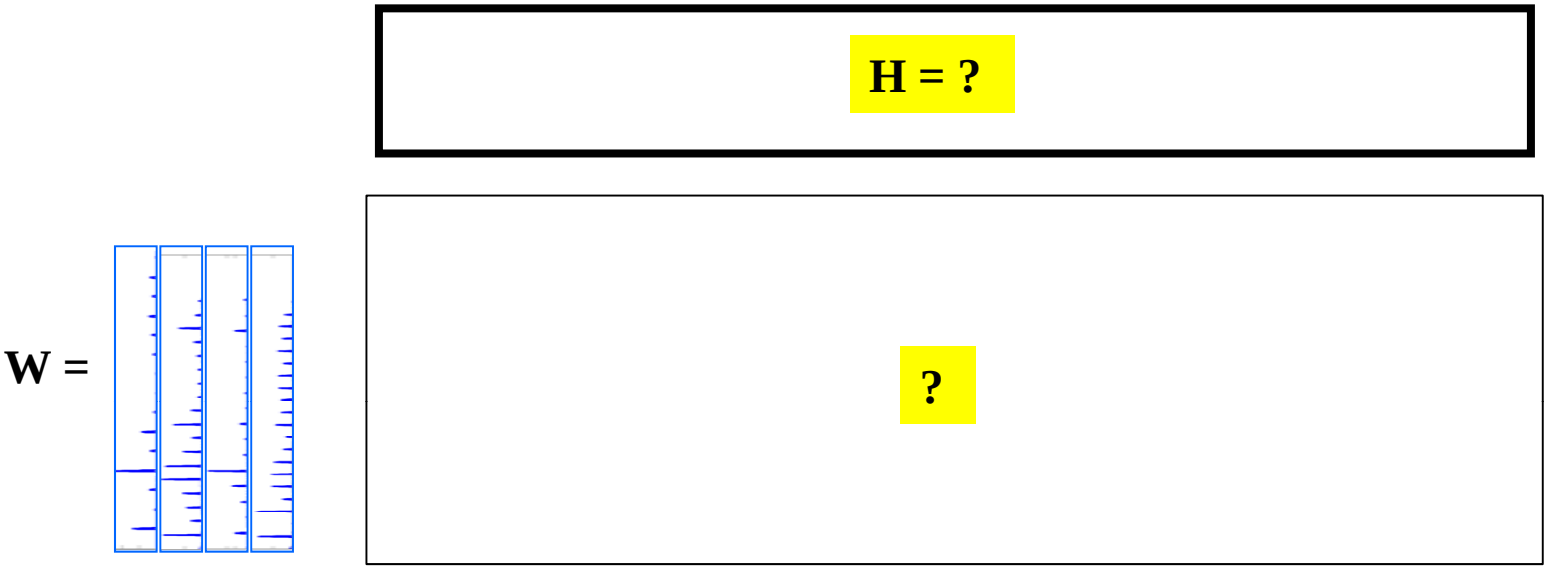


W = **?**

U = **?**

- **M ~ WH** **W = Mpinv(V)** **U = WH**

So what are we doing here?

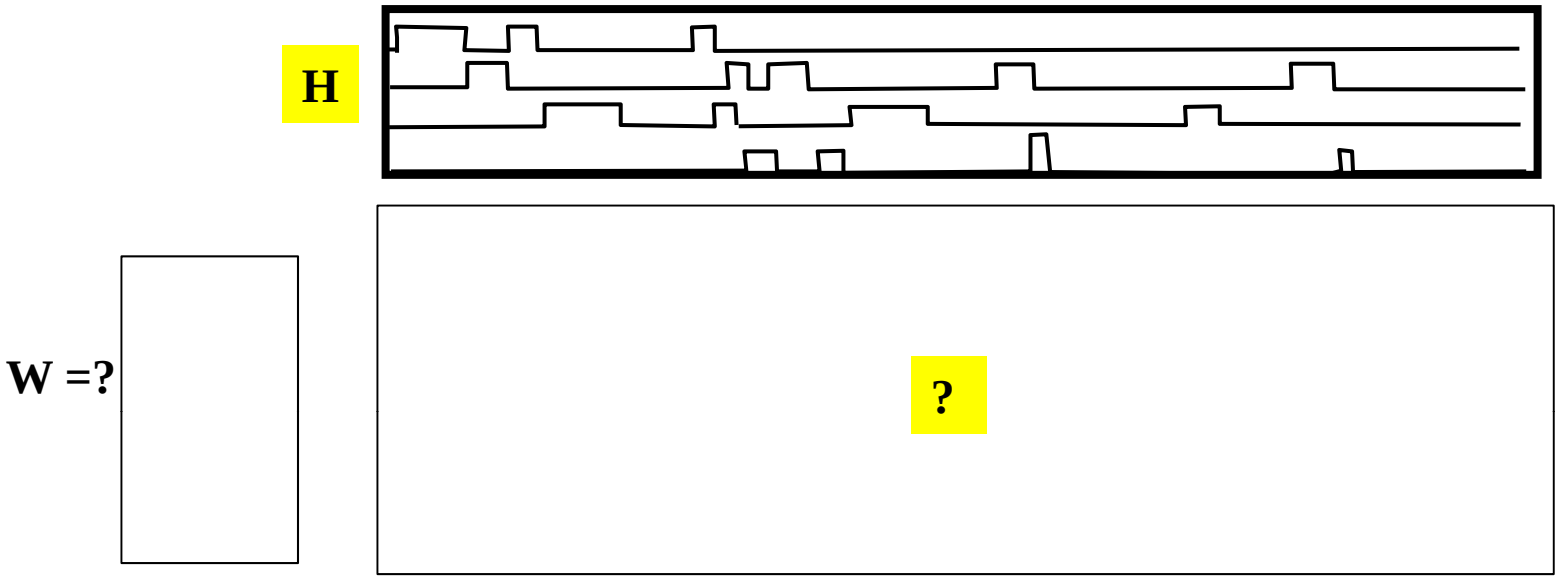


- **M ~ WH** is an approximation
- Given **W**, estimate **H** to minimize error

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

- Must ideally find *transcription* of given notes

Going the other way..

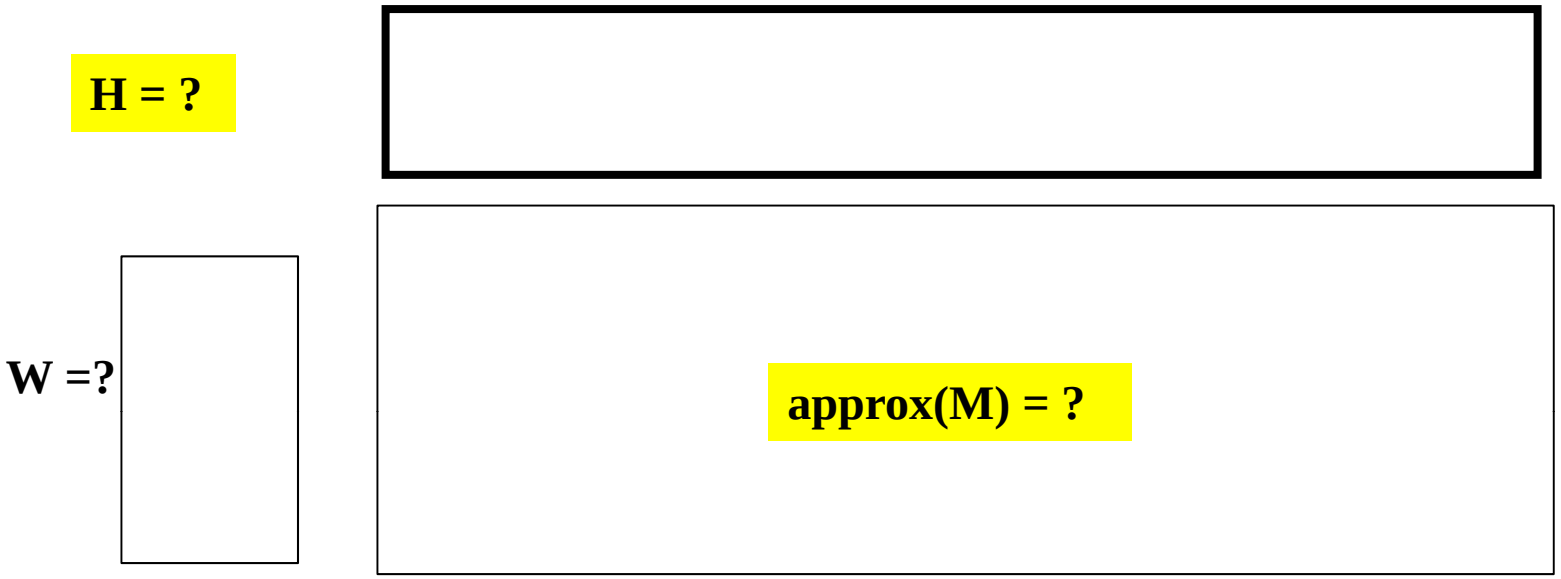


- $\mathbf{M} \sim \mathbf{WH}$ is an approximation
- Given $\mathbf{H}$, estimate $\mathbf{W}$ to minimize error

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

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

When both parameters are unknown

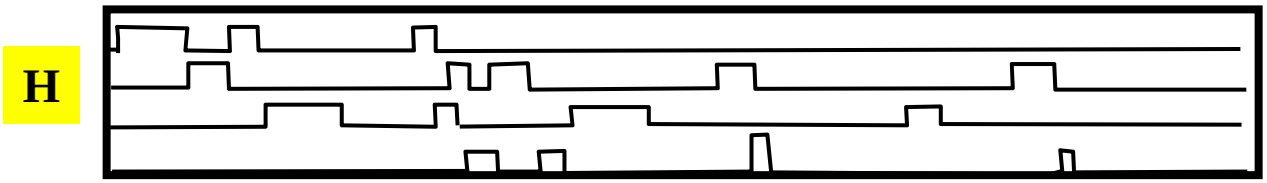


- Must estimate both **H** and **W** to best approximate **M**
- Ideally, must learn *both* the *notes* and *their* transcription!

A least squares solution

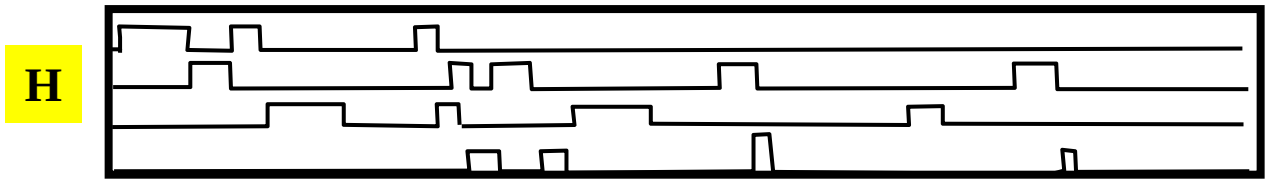
$$\mathbf{W},\mathbf{H} = \arg\min_{\overline{\mathbf{W}},\overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{W}}\overline{\mathbf{H}} \|_F^2$$

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



- 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 $\mathbf{H}$ are *uncorrelated*

A least squares solution



- 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

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

- Note $\mathbf{H}^T\mathbf{H} \; != \; \mathbf{I}$
  - Only $\mathbf{H}\mathbf{H}^T = \mathbf{I}$
  - Could also be rewritten as

$$\mathbf{H} = \arg\min_{\overline{\mathbf{H}}} trace\left(\mathbf{M}(\mathbf{I} - \overline{\mathbf{H}}^T\overline{\mathbf{H}})\mathbf{M}^T\right)$$

$$\mathbf{H} = \arg\min_{\overline{\mathbf{H}}} trace\left(\mathbf{M}^T\mathbf{M}(\mathbf{I} - \overline{\mathbf{H}}^T\overline{\mathbf{H}})\right)$$

$$\mathbf{H} = \arg\min_{\overline{\mathbf{H}}} trace\left(Correlation(\mathbf{M}^T)(\mathbf{I} - \overline{\mathbf{H}}^T\overline{\mathbf{H}})\right)$$

$$\mathbf{H} = \arg\max_{\overline{\mathbf{H}}} trace\left(Correlation(\mathbf{M}^T)\overline{\mathbf{H}}^T\overline{\mathbf{H}}\right)$$

Finding the notes

- Constraint: every row of $\mathbf{H}$ has length 1

$$\mathbf{H} = \arg\max_{\overline{\mathbf{H}}} trace\left(Correlation(\mathbf{M}^T)\overline{\mathbf{H}}^T\overline{\mathbf{H}}\right) - trace\left(\Lambda\overline{\mathbf{H}}^T\overline{\mathbf{H}}\right)$$

- Differentiating and equating to 0

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

- Simply requiring the rows of H to be orthonormal gives us that $\mathbf{H}$ is the set of Eigenvectors of the data in $\mathbf{M}^T$

Equivalences

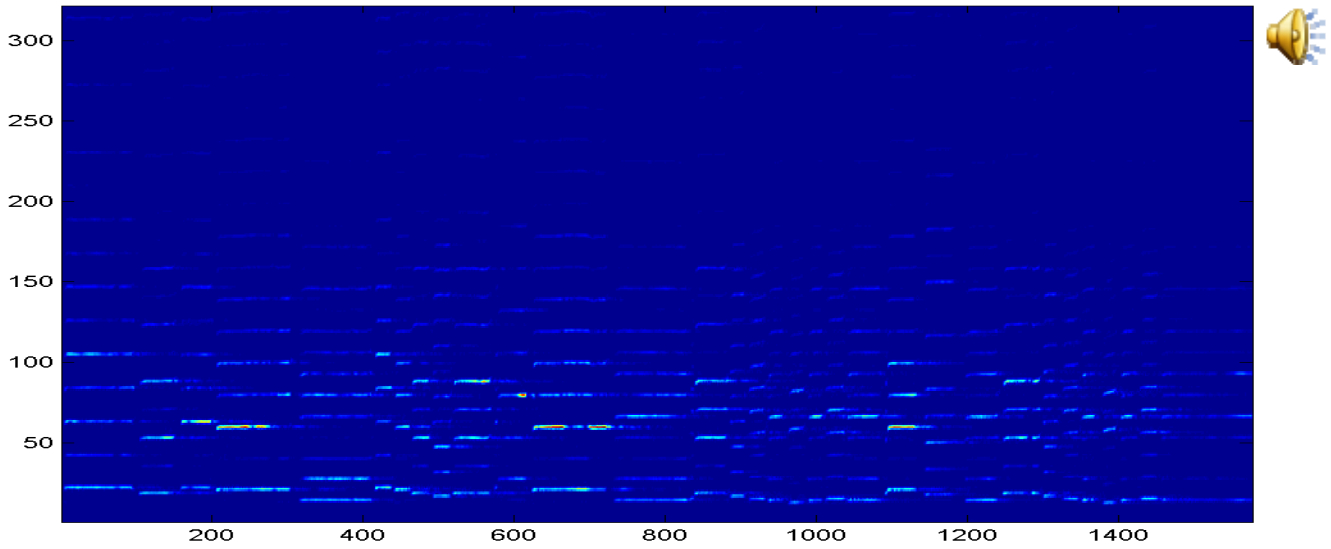
$$\mathbf{H} = \arg\max_{\overline{\mathbf{H}}} trace\left(Correlation(\mathbf{M}^T)\overline{\mathbf{H}}^T\overline{\mathbf{H}}\right) - trace\left(\Lambda\overline{\mathbf{H}}^T\overline{\mathbf{H}}\right)$$

- is identical to

$$\mathbf{W},\mathbf{H} = \arg\min_{\overline{\mathbf{W}},\overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{W}}\overline{\mathbf{H}} \|_F^2 + \sum_i \lambda_i \| \overline{\mathbf{h}}_i \|^2 + \sum_{i \neq j} \lambda_{ij} \overline{\mathbf{h}}_i^T \overline{\mathbf{h}}_j$$

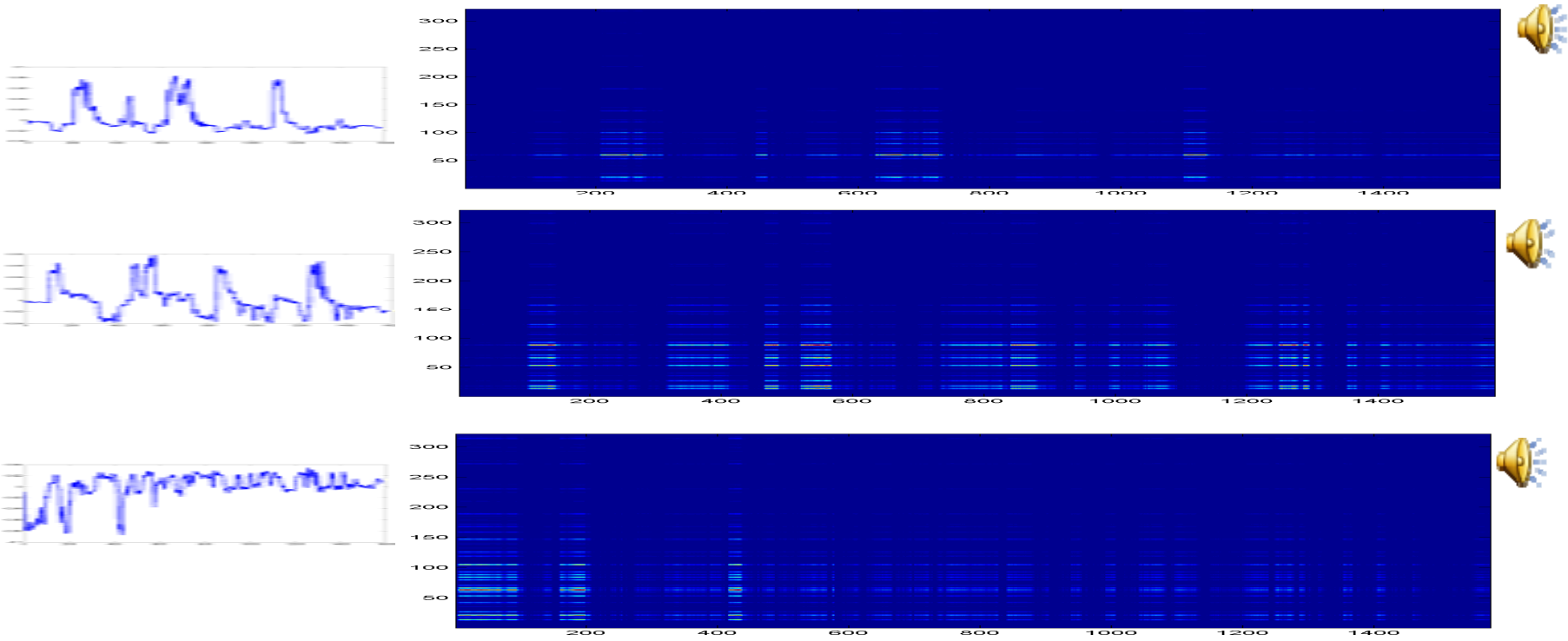
- Minimize least squares error with the constraint that the rows of $\mathbf{H}$ are length 1 and orthogonal to one another

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 first three "notes" and their contributions
  - The spectrograms of the notes are statistically uncorrelated

Finding the notes

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

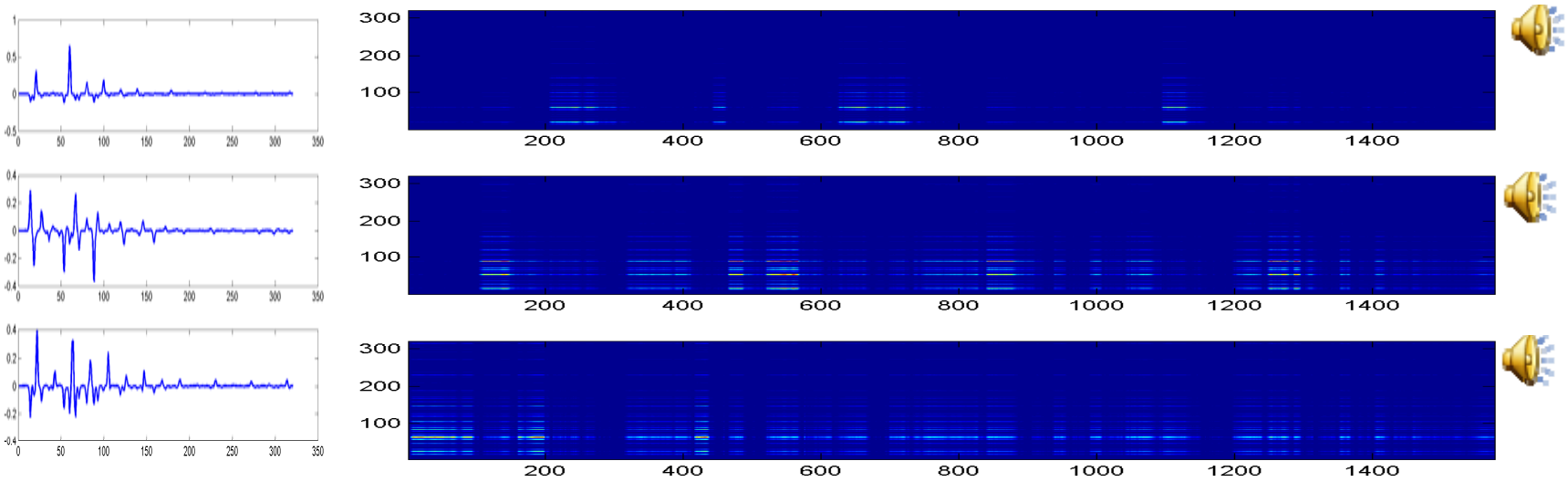
$$\mathbf{W} = \arg\min_{\overline{\mathbf{W}}} \| \mathbf{M} - \overline{\mathbf{W}}^T \overline{\mathbf{W}} \mathbf{M} \|_F^2$$

- Solving the above, with the constraints that the columns of $\mathbf{W}$ are orthonormal gives you the eigen vectors of the data in $\mathbf{M}$

$$\mathbf{W} = \arg\max_{\overline{\mathbf{W}}} trace\left(\overline{\mathbf{W}}^T \overline{\mathbf{W}} Correlation(\mathbf{M})\right) - trace\left(\Lambda \overline{\mathbf{W}}^T \overline{\mathbf{W}}\right)$$

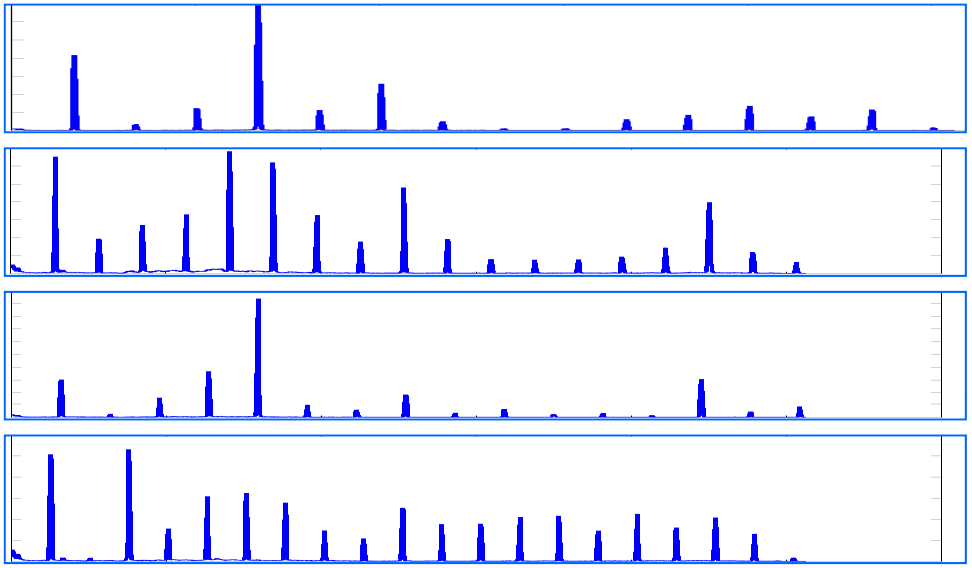
$$Correlation(\mathbf{M})\mathbf{W} = \Lambda \mathbf{W}$$

So how does that work?



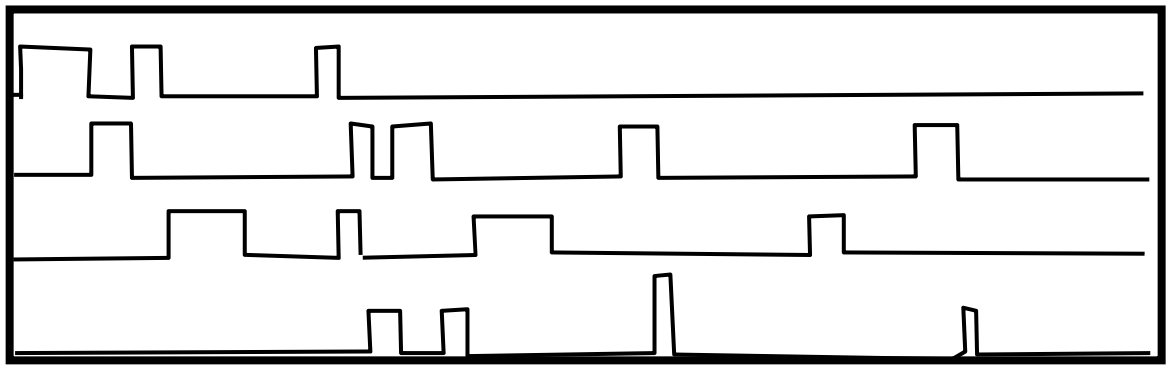
- There are 12 notes in the segment, hence we try to estimate 12 notes..

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

What *else* can we look for?



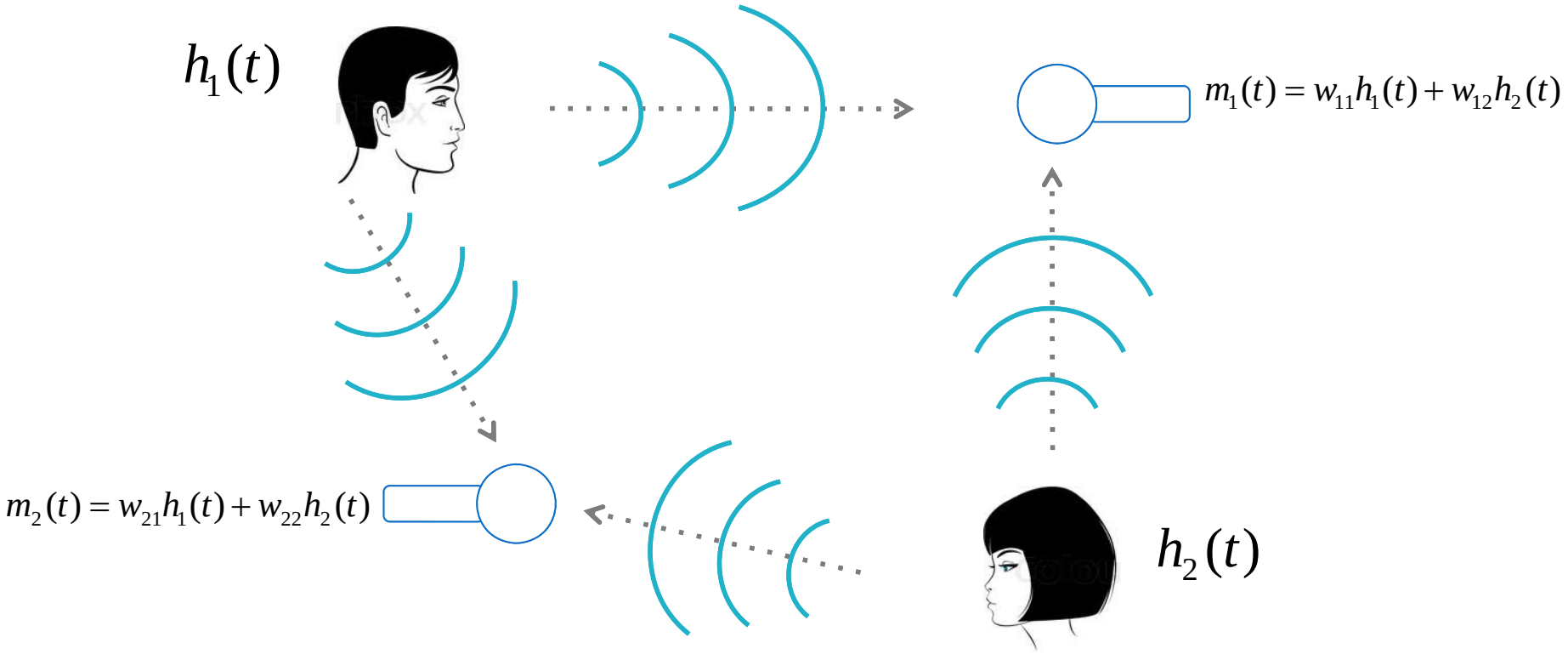
- Assume: The "transcription" of one note does not depend on what else is playing
  - Or, in a multi-instrument piece, instruments are playing independently of one another
- Not strictly true, but still..

Formulating it with Independence

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

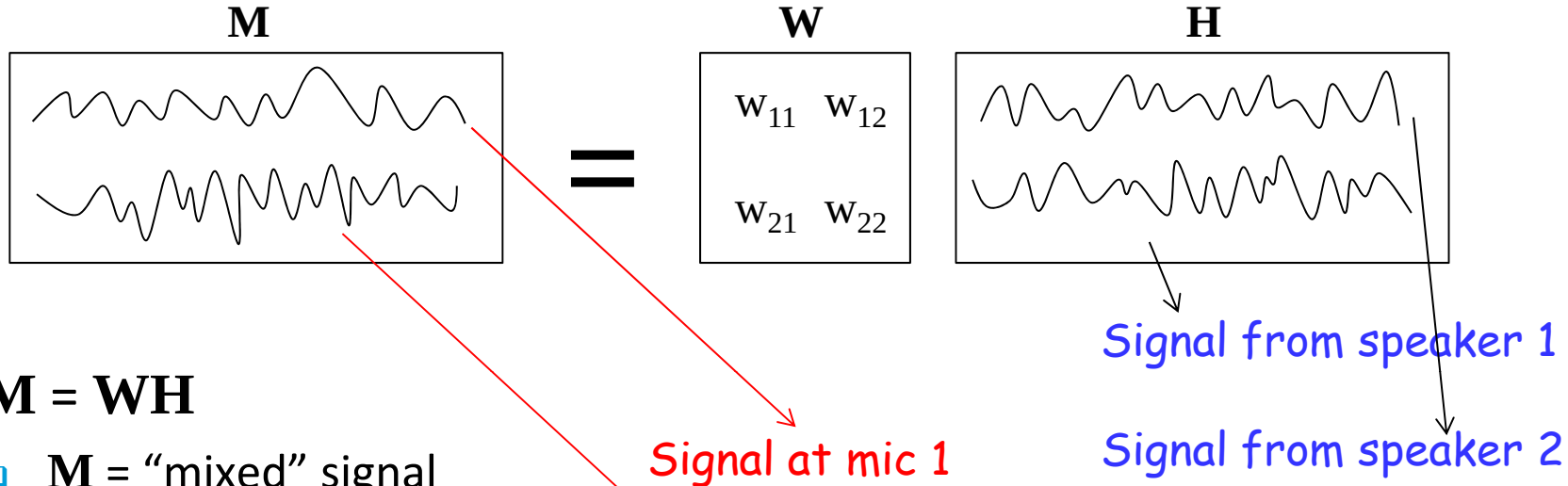
- Impose statistical independence constraints on decomposition

Changing problems for a bit



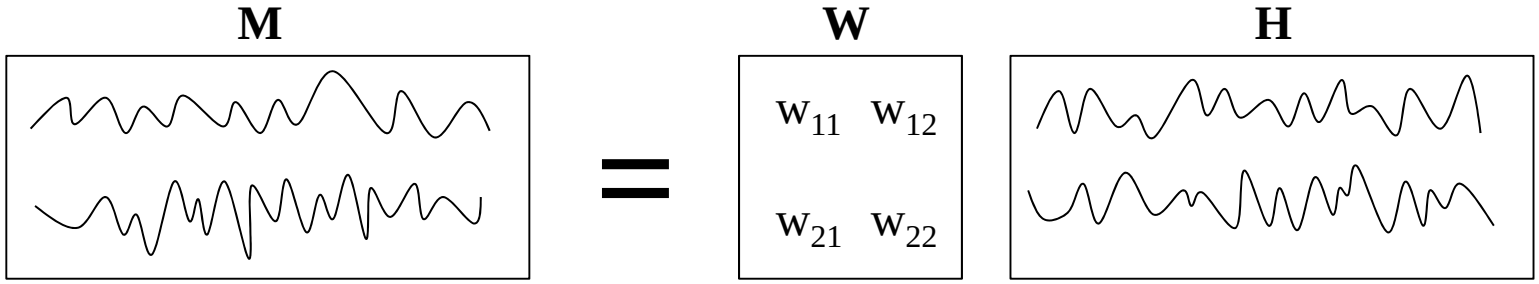
- Two people speak simultaneously
- Recorded by two microphones
- Each recorded signal is a mixture of both signals

Imposing Statistical Constraints



- **M** = **WH**
  - **M** = "mixed" signal
  - **W** = "notes"
  - **H** = "transcription"
- Given only **M** estimate **H**
- Ensure that the components of the vectors in the estimated **H** are statistically independent
- Multiple approaches..

Imposing Statistical Constraints

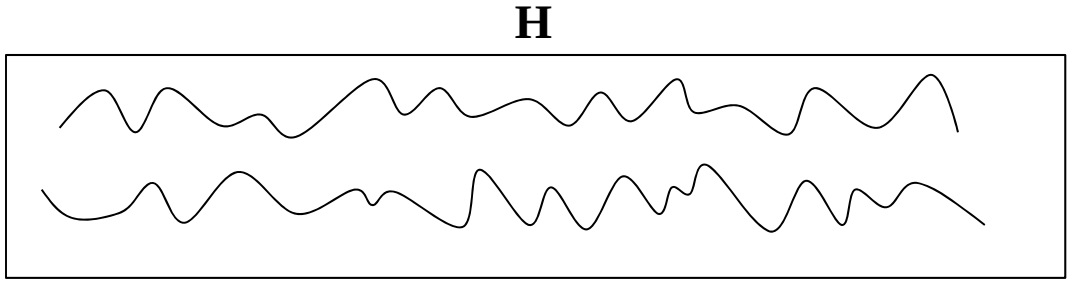


- $\mathbf{M} = \mathbf{WH}$
- Given only $\mathbf{M}$ estimate $\mathbf{H}$
- $\mathbf{H} = \mathbf{W}^{-1}\mathbf{M} = \mathbf{AM}$
- Estimate $\mathbf{A}$ such that the components of $\mathbf{AM}$ are statistically independent
  - $\mathbf{A}$ is the *unmixing* matrix
- Multiple approaches..

Statistical Independence

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

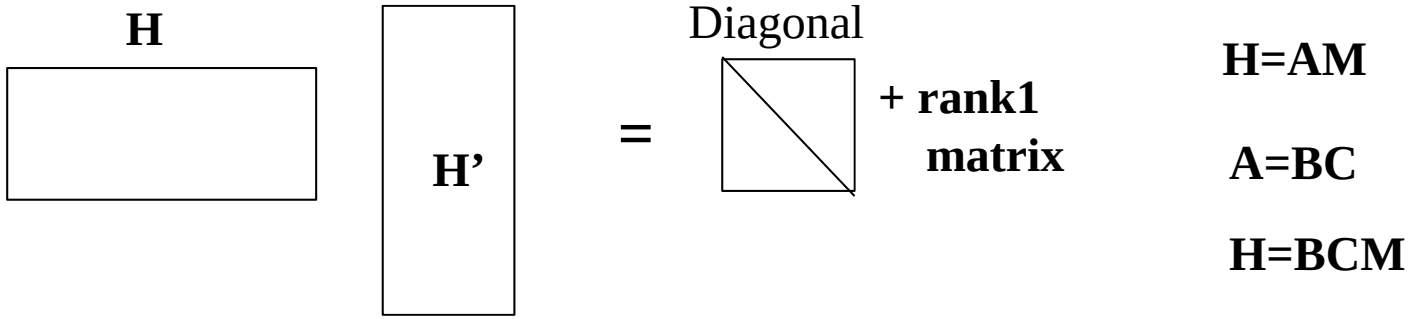
- Usual to assume *zero mean* processes
  - Otherwise, some of the math doesn't work well
- $\mathbf{M} = \mathbf{WH}$ $\quad\quad$ $\mathbf{H} = \mathbf{AM}$
- If mean($\mathbf{M}$) = 0 => mean($\mathbf{H}$) = 0
  - $\mathbf{E[H] = AE[M] = A0 = 0}$
  - First step of ICA: Set the mean of $\mathbf{M}$ to 0

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

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

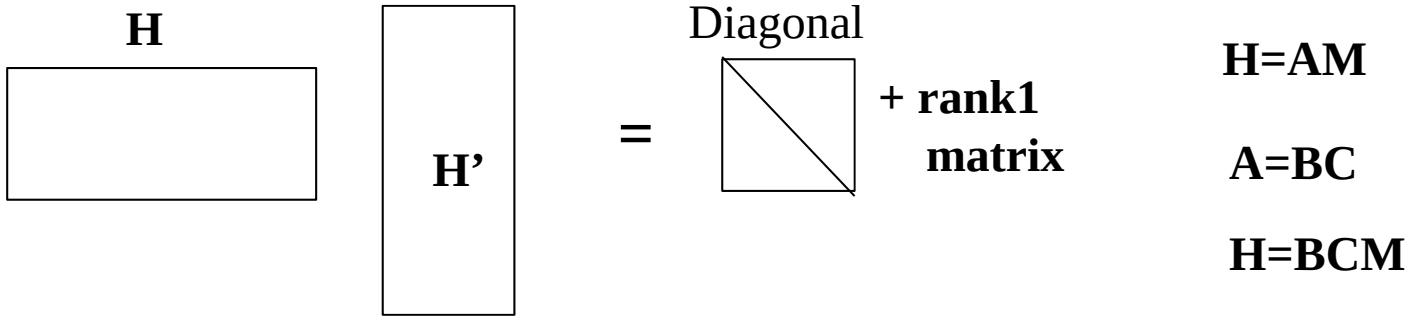
  - $\mathbf{m}_i$ are the columns of $\mathbf{M}$

Emulating Independence..



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

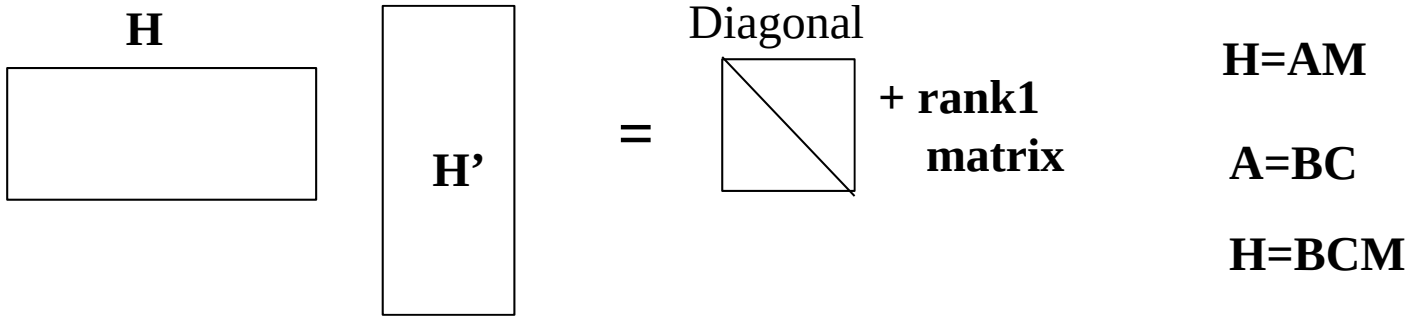
Decorrelating



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

- Eigen decomposition $\mathbf{M}\mathbf{M}^T = \mathbf{U}\mathbf{S}\mathbf{U}^T$
- Let $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{U}^T$
  - $\mathbf{W}\mathbf{M}\mathbf{M}^T\mathbf{W}^T = \mathbf{S}^{-1/2}\mathbf{U}^T\,\mathbf{U}\mathbf{S}\mathbf{U}^T\mathbf{U}\mathbf{S}^{-1/2} = \mathbf{I}$

Decorrelating



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

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

- $\mathbf{X}$ 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.

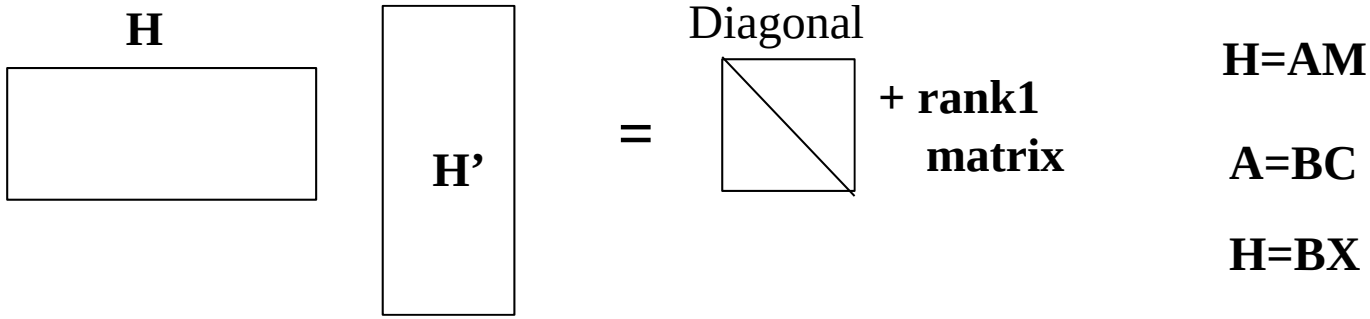
$$E[\mathbf{x}_i \mathbf{x}_j] = 0 \text{ if } i \mathrel{!}= j$$

- 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



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

- Will multiplying $\mathbf{X}$ by $\mathbf{B}$ *re-correlate* the components?
- Not if B is *unitary*
  - $\mathbf{B}\mathbf{B}^T = \mathbf{B}^T\mathbf{B} = \mathbf{I}$
- So we want to find a *unitary* matrix
  - Since the rows of $\mathbf{H}$ 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
- **A=BC, H = BCM, X = CM, → H = BX**
- The fourth moments of **H** have the form:
  $E[\mathbf{h}_i\ \mathbf{h}_j\ \mathbf{h}_k\ \mathbf{h}_l]$
- If the rows of **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 **B** such that the fourth moments of **H = BX** are decoupled
  - While ensuring that **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 $\mathbf{H}$

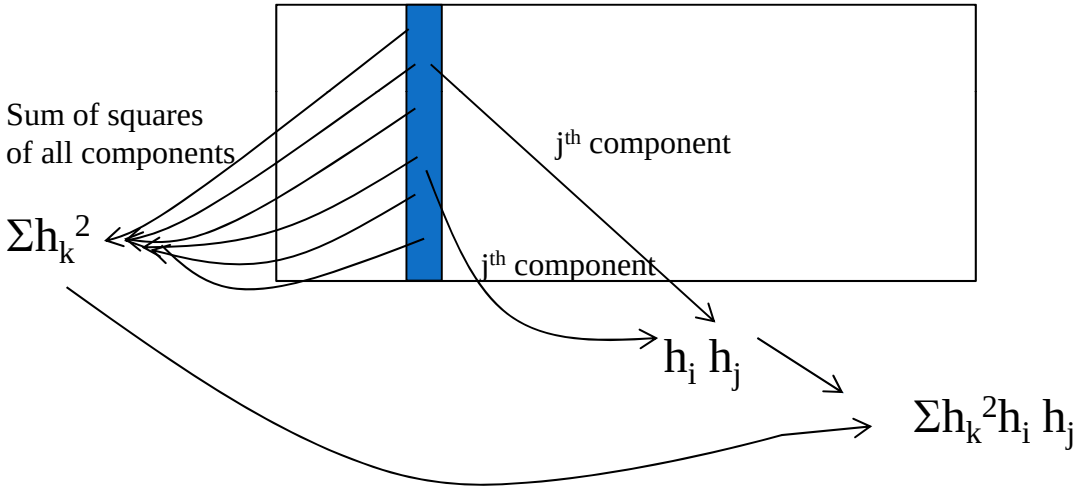
$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & .. \\ d_{21} & d_{22} & d_{23} & .. \\ .. & .. & .. & .. \end{bmatrix}$$

  - 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

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

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



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

ICA: The D matrix

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

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

- 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 $\mathbf{X}$
  - Recall: $\mathbf{X = CM}$, $\mathbf{H = BX = BCM}$
    - $\mathbf{B}$ is what we're trying to learn to make $\mathbf{H}$ independent
  - Compose $\mathbf{D'} = E[\mathbf{x}^T\ \mathbf{x}\ \mathbf{x}\ \mathbf{x}^T]$
- Diagonalize $\mathbf{D'}$ via Eigen decmpositin

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

B frees the fourth moment

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

- $\mathbf{U}$ 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 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

- **H = AM = BCM = BX**
- **A = BC = U^TC**

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 R_{MM} of **M**
  3. Compute whitening matrix **C** via Eigen decomposition
     $R_{XX} = \mathbf{E}\mathbf{S}\mathbf{E}^T, \quad \mathbf{C} = \mathbf{S}^{-1/2}\mathbf{E}^T$
  4. Compute $\mathbf{X} = \mathbf{C}\mathbf{M}$
  5. Compute the fourth moment matrix $D' = E[\mathbf{x}^T\mathbf{x}\mathbf{x}\mathbf{x}^T]$
  6. Diagonalize D' via Eigen decomposition
  7. $D' = \mathbf{U}\Lambda\mathbf{U}^T$
  8. Compute $\mathbf{A} = \mathbf{U}^T\mathbf{C}$
- 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 more computationally 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
- $\mathbf{X} = \mathbf{CM}$
  - $E[\mathbf{x}_i\mathbf{x}_j] = E[\mathbf{x}_i]E[\mathbf{x}_j] = \delta_{ij}$ for centered signals
- 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

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

- **h = Bx**
  - Individual columns of the **H** and **X** matrices
  - **x** is mixed signal, **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)

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

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

An alternate approach

- Recall PCA
- $\mathbf{M} = \mathbf{WH}$, the columns of $\mathbf{W}$ must be statistically independent
- Leads to: $\min_{\mathbf{W}} ||\mathbf{M} - \mathbf{W}^T\mathbf{WM}||^2$
  - 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)

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

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

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

An alternate approach

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

$\mathbf{P} =$

$$\begin{matrix} P_{11} & P_{21} & \cdots \\ P_{12} & P_{22} & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{matrix}$$

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

This is a square matrix

- Must ideally be

$\mathbf{Q} =$

$$\begin{matrix} Q_{11} & Q_{21} & \cdots \\ Q_{12} & Q_{22} & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{matrix}$$

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

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

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

An alternate approach

- Ideal value for **Q**

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

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

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

- 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

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

$$\mathbf{Q} = \mathit{Diagonal}$$

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

- Leads to trivial Widrow Hopf type iterative rule:

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

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

Update Rules

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

Update Rules

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

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

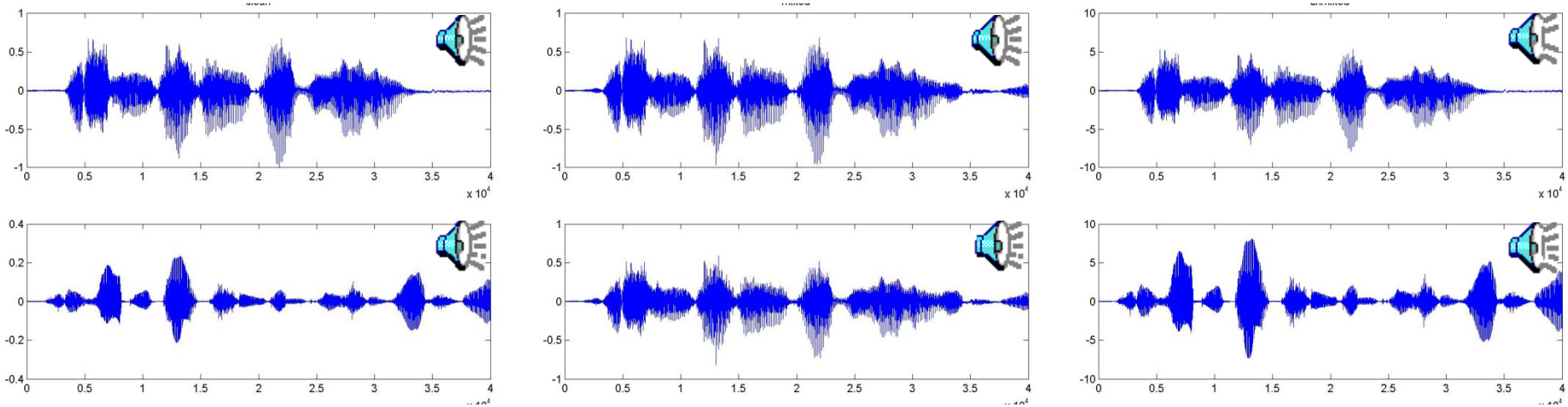
What are G() and H()

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

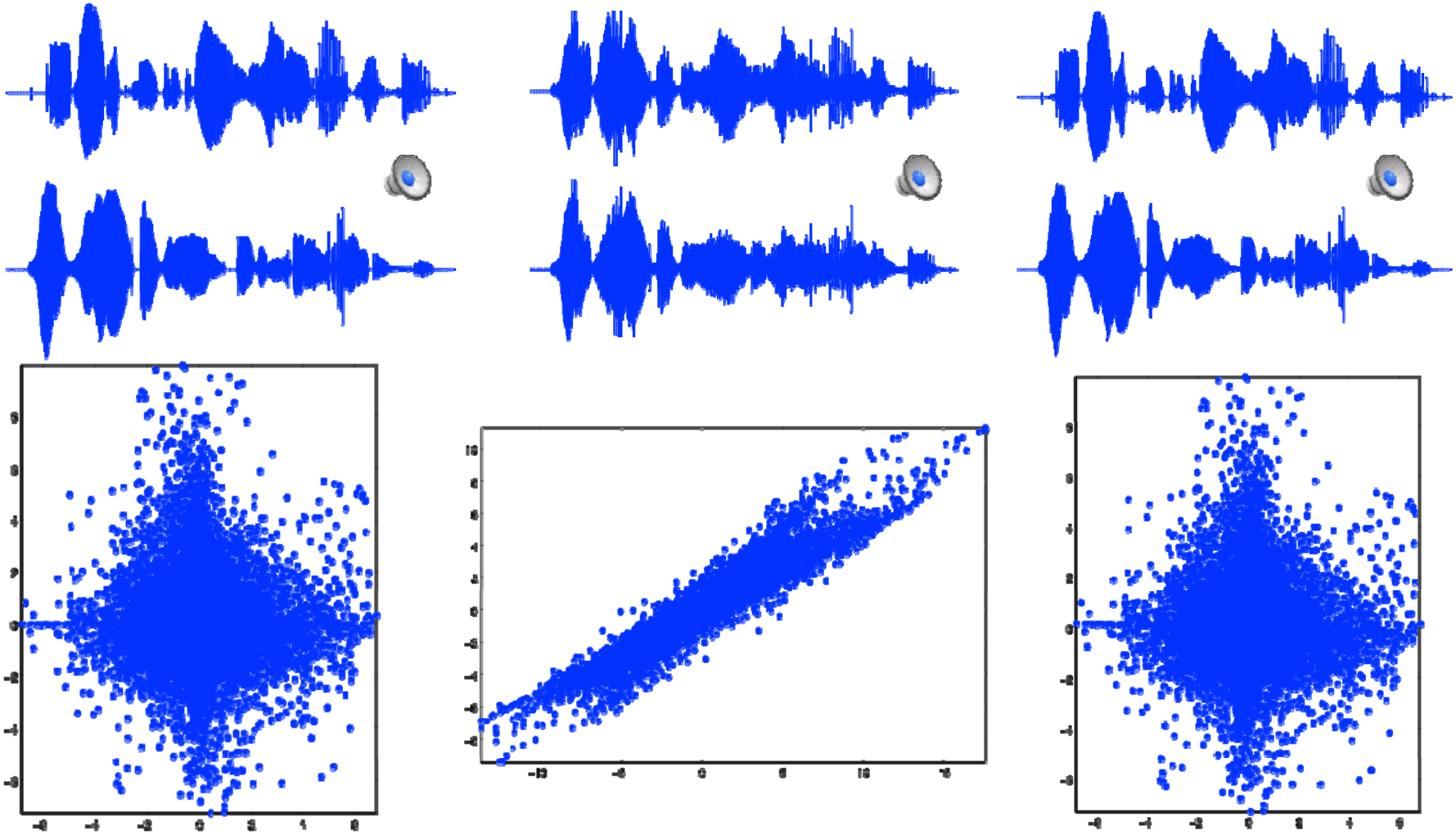
So how does it work?



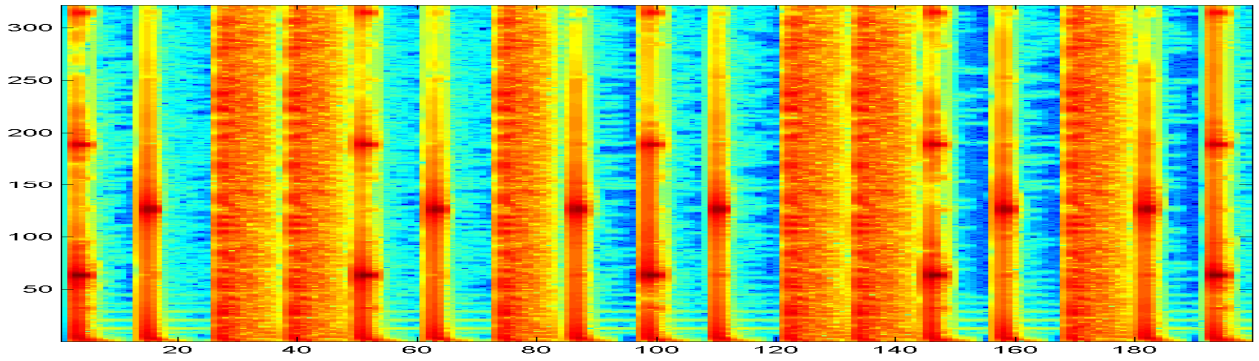
- Example with instantaneous mixture of two speakers
- Natural gradient update
- Works very well!

Another example!

Input ***Mix*** ***Output***

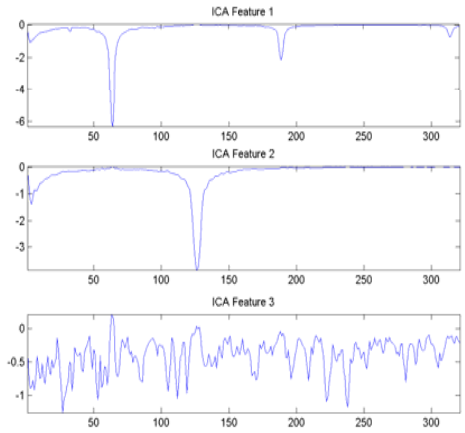


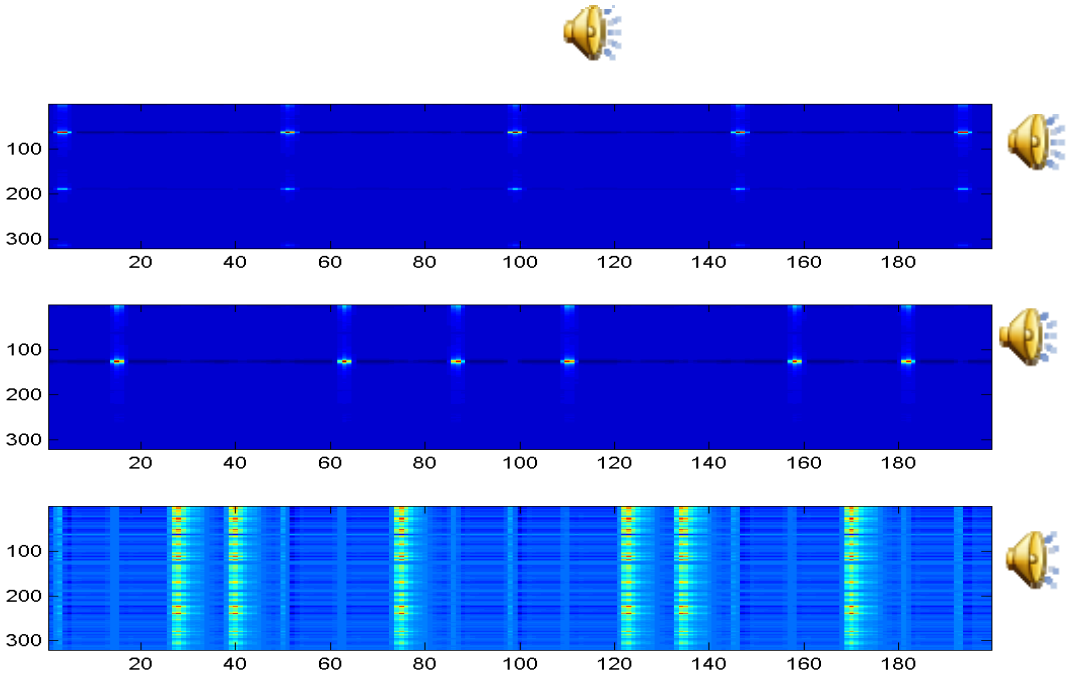
Another Example



- Three instruments..

The Notes

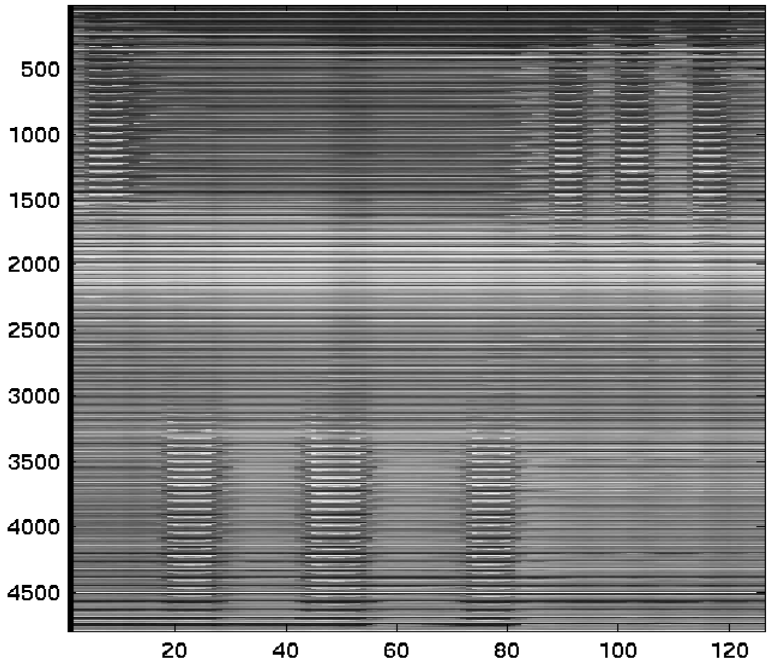




- Three instruments..

ICA for data exploration

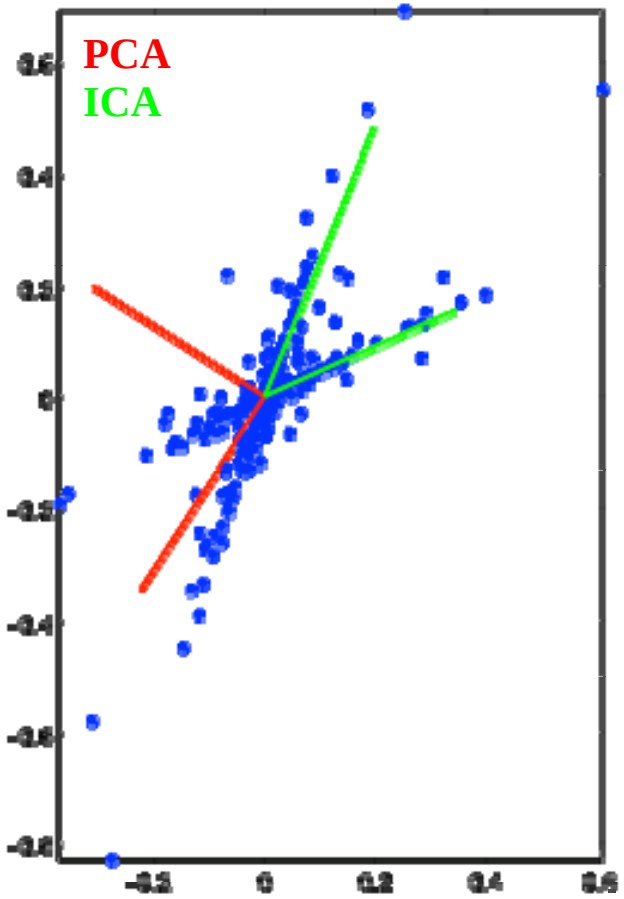
- The "bases" in PCA represent the "building blocks"
  - Ideally notes
- Very successfully used
- So can ICA be used to do the same?



ICA vs PCA bases

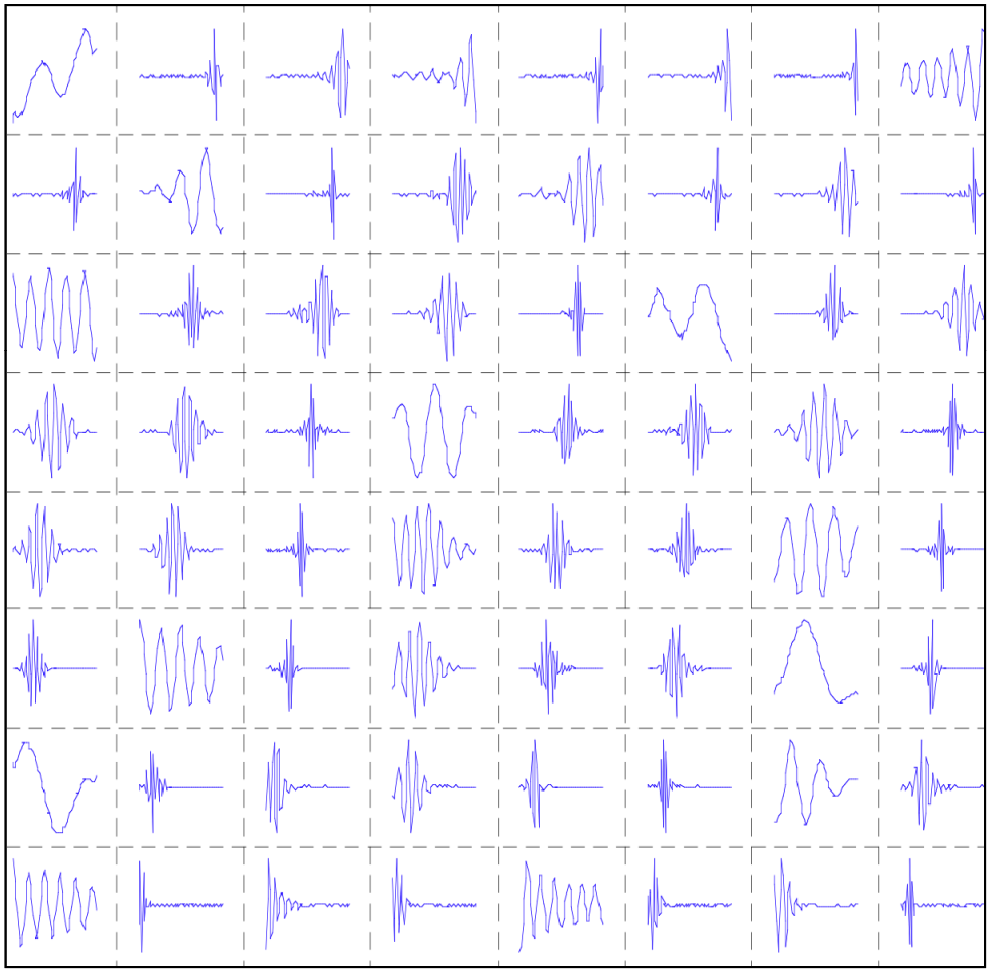
- Motivation for using ICA vs PCA
- PCA will indicate orthogonal directions of maximal variance
  - May not align with the data!
- ICA finds directions that are independent
  - More likely to “align” with the data

Non-Gaussian data



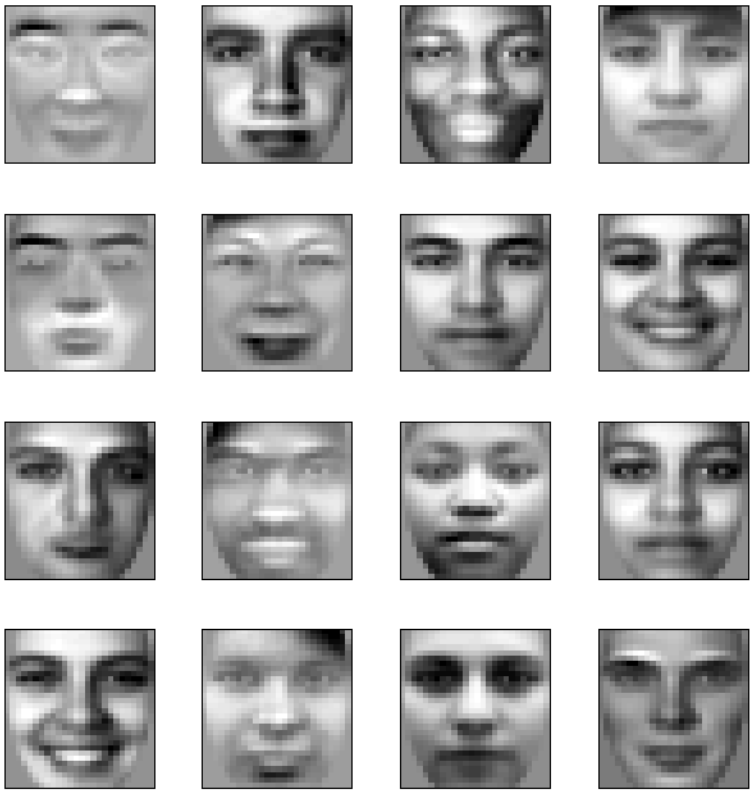
Finding useful transforms with ICA

- 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

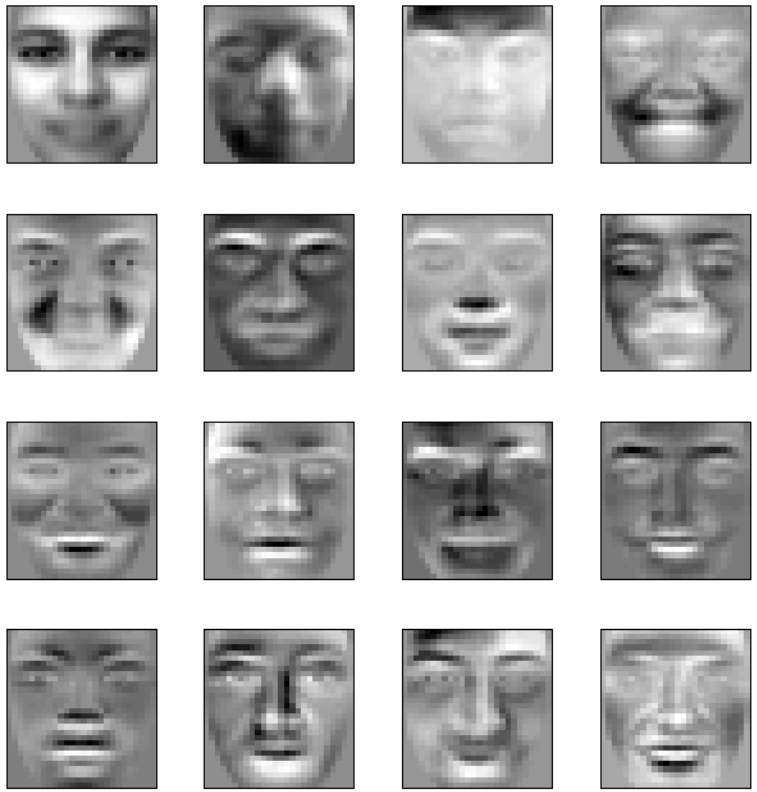


Example case: ICA-faces vs. Eigenfaces

ICA-faces



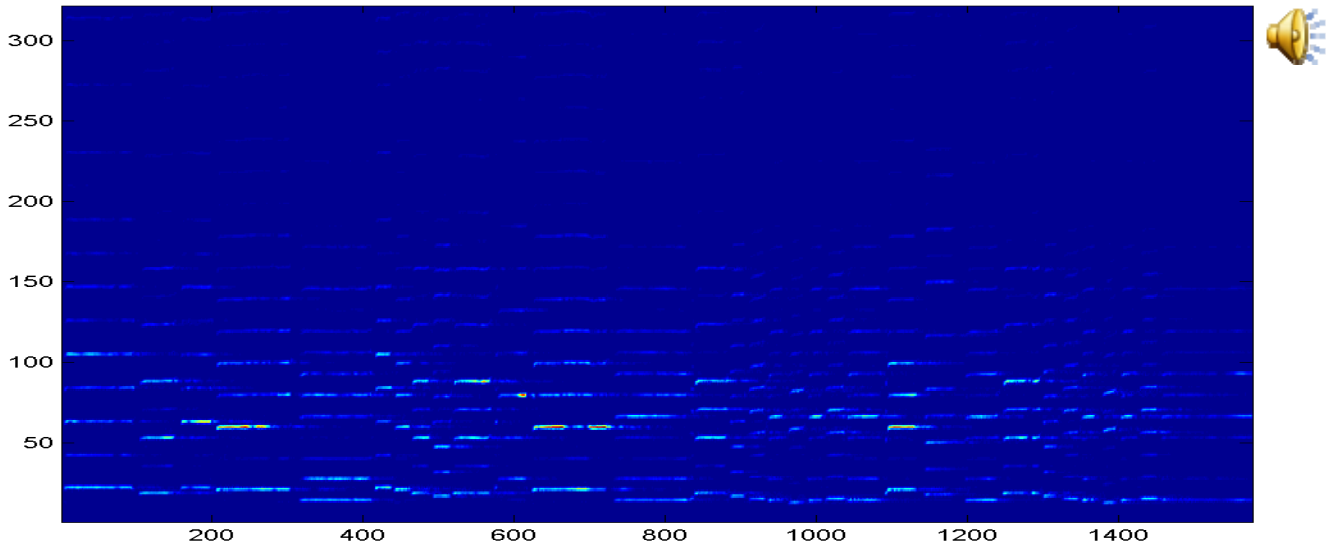
Eigenfaces



ICA for Signal Enhncement

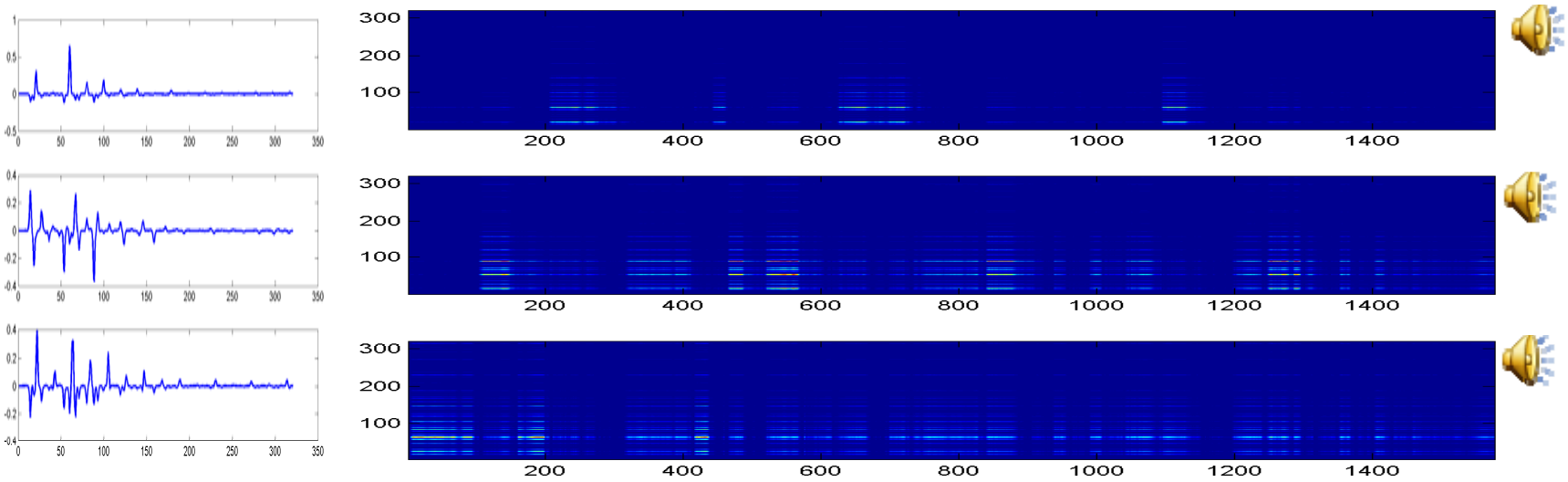
- Very commonly used to enhance EEG signals
- EEG signals are frequently corrupted by heartbeats and biorhythm signals
- ICA can be used to separate them out

So how does that work?



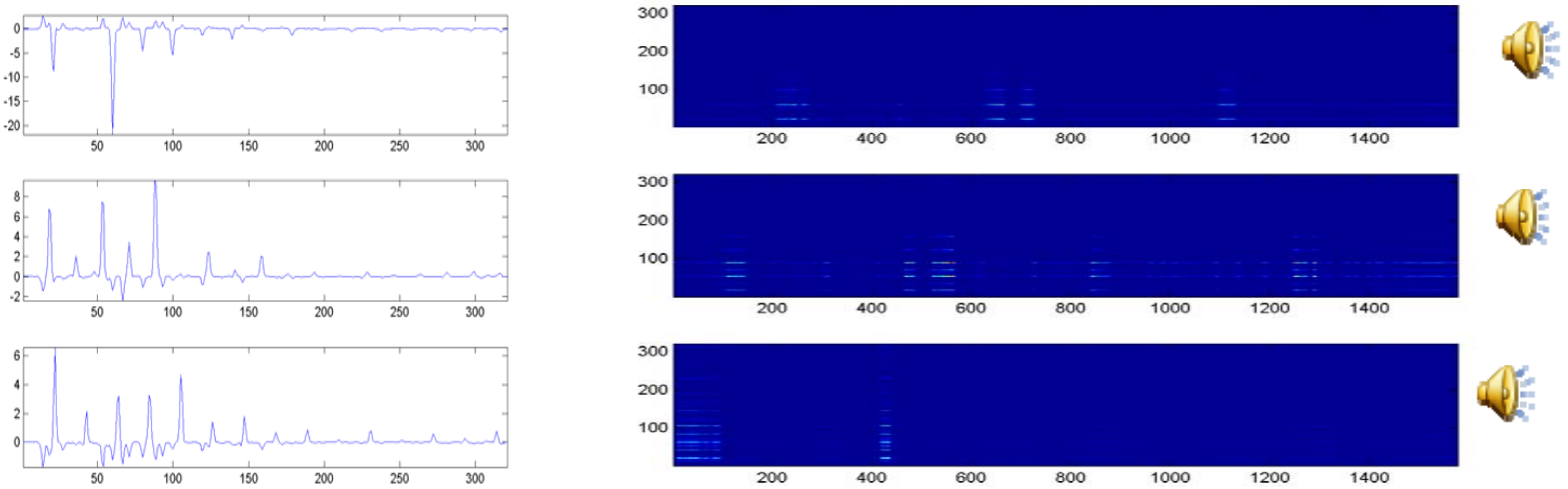
- There are 12 notes in the segment, hence we try to estimate 12 notes..

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?

- Assume distribution of signals is symmetric around mean
  - Note energy here
  - *Not* symmetric – negative values never happen
  - Still this didn't affect the three instruments case..

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

Continue in next class..

- NMF
- Factor analysis..