Machine Learning for Signal Processing

Independent Component Analysis

Class 8. 24 Sep 2015

Instructor: Bhiksha Raj
Revisiting the Covariance Matrix

- Assuming centered data

\[ C = \sum_X XX^\top \]

\[ = X_1X_1^\top + X_2X_2^\top + \ldots \]

- 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 twice the length of the corresponding vector.
Covariance Matrix as a transform

- Data comprises only 2 vectors.
- *Major axis of component ellipses proportional to twice the length of the corresponding vector*
Covariance Matrix as a transform

- More vectors..
- *Major axis of component ellipses proportional to twice the length of the corresponding vector*
Covariance Matrix as a transform

- More vectors..
- **Major axis of component ellipses proportional to twice the length of the corresponding vector**
Covariance Matrix as a transform

- And still more vectors..
- *Major axis of component ellipses proportional to twice the length of the corresponding vector*
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 \textit{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 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

\[ \text{Burger consumption} \]

\[ \text{Penguin population} \]

\[ P_1 \]

\[ P_2 \]

\[ b_1 \]

\[ b_2 \]
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?
The notion of **decorrelation**

\[
\begin{bmatrix}
X' \\
Y'
\end{bmatrix} = M \begin{bmatrix}
X \\
Y
\end{bmatrix}
\]

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

---

11755/18797
What does “uncorrelated” mean

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

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

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

• Let $X$ be the matrix of correlated data vectors
  – Each component of $X$ informs us of the mean trend of other components

• Need a transform $M$ such that if $Y = MX$ such that the covariance of $Y$ is diagonal
  – $YY^T$ is the covariance if $Y$ is zero mean
  – $YY^T = \text{Diagonal}$
  $\Rightarrow MXX^TM^T = \text{Diagonal}$
  $\Rightarrow M.\text{Cov}(X).M^T = \text{Diagonal}$
Decorrelation

• Easy solution:
  – Eigen decomposition of $\text{Cov}(X)$:
    
    $$\text{Cov}(X) = \mathbf{E} \Lambda \mathbf{E}^T$$
  – $\mathbf{E} \mathbf{E}^T = \mathbf{I}$

• Let $\mathbf{M} = \mathbf{E}^T$

• $\mathbf{M}\text{Cov}(X)\mathbf{M}^T = \mathbf{E}^T \mathbf{E} \Lambda \mathbf{E}^T \mathbf{E} = \Lambda = \text{diagonal}$

• PCA: $\mathbf{Y} = \mathbf{M}^T \mathbf{X}$

• Diagonalizes the covariance matrix
  – “Decorrelates” the data
PCA

\[ X = w_1 E_1 + w_2 E_2 \]

- PCA: \( Y = M^T X \)
- *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?
• 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 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 $Y$
    - $E[X|Y] = E[X]$
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 brief review of basic info. theory

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

\[ H(X) = \sum_X P(X)[-\log P(X)] \]

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

\[ H(X,Y) = \sum_{X,Y} P(X,Y)[-\log P(X,Y)] \]
A brief review of basic info. theory

• 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

\[
H(X \mid Y) = \sum_Y P(Y) \sum_X P(X \mid Y) \left( -\log P(X \mid Y) \right) = \sum_{X,Y} P(X,Y) \left( -\log P(X \mid Y) \right)
\]
A brief review of basic info. theory

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

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

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

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

- 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 = \( PM \)
We’re actually computing a score

\[ M = \]

\[ W = \]

- \[ M \sim WH \]
- \[ H = \text{pinv}(W)M \]
How about the other way?

\[ M = \]

\[ H = \]

\[ W = ? \]

\[ U = ? \]

- \[ M \sim WH \]
- \[ W = M \text{pinv}(H) \]
- \[ U = WH \]
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

\[ W, H = \arg \min_{W,H} \| M - WH \|_F^2 + \lambda (W^T W - I) \]

- Constraint: \( W \) is orthogonal
  - \( W^T W = I \)
- The solution: \( W \) are the Eigen vectors of \( M M^T \)
  - PCA!!

- \( M \sim WH \) is an approximation
- Also, the rows of \( H \) are \emph{decorrelated}
  - Trivial to prove that \( HH^T \) is diagonal
PCA

\[ W, H = \arg \min_{W, H} \| M - WH \|_F^2 \]

\[ M \approx WH \]

• 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

\[ W, H = \arg \min_{W, H} \| M - \overline{H} \|_F^2 + \lambda (\overline{HH}^T - D) \]

- Different constraint: Constraint \( H \) to be decorrelated
  \[ HH^T = D \]
- This will result exactly in PCA too
- Decorrelation of \( H \) 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..
Formulating it with Independence

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

• 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

\[ m_1(t) = w_{11}h_1(t) + w_{12}h_2(t) \]

\[ m_2(t) = w_{21}h_1(t) + w_{22}h_2(t) \]
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 $M$ estimate $H$
- Identical to the problem of “finding notes”
Imposing Statistical Constraints

- \( M = WH \)
- Given only \( M \) estimate \( H \)
- \( H = W^{-1}M = AM \)
- 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

- $M = WH$
- $H = AM$

Remember this form
An ugly algebraic solution

\[ \mathbf{M} = \mathbf{WH} \quad \ldots \ldots \quad \mathbf{H} = \mathbf{AM} \]

- 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 (\( \mathbf{H}\mathbf{H}^T \))

- Can we do the same for independence
  - Is there a linear transform that will enforce independence?
Emulating Independence

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

• The fourth order moments are independent
  - $E[h_i h_j h_k h_l] = E[h_i]E[h_j]E[h_k]E[h_l]$
  - $E[h_i^2 h_j h_k] = E[h_i^2]E[h_j]E[h_k]$
  - $E[h_i^2 h_j^2] = E[h_i^2]E[h_j^2]$
  - Etc.
Zero Mean

• Usual to assume zero mean processes
  – Otherwise, some of the math doesn’t work well

• \( M = WH \quad H = AM \)

• If \( \text{mean}(M) = 0 \implies \text{mean}(H) = 0 \)
  – \( \mathbb{E}[H] = A \cdot \mathbb{E}[M] = A0 = 0 \)
  – First step of ICA: Set the mean of \( M \) to 0

\[
\mu_m = \frac{1}{\text{cols}(M)} \sum_i m_i
\]

\[
m_i = m_i - \mu_m \quad \forall i
\]

– \( m_i \) are the columns of \( M \)
Emulating Independence..

- Independence $\rightarrow$ Uncorrelatedness
- Estimate a $C$ such that $CM$ is decorrelated
- A little more than PCA

H \begin{bmatrix} \text{Diagonal} \\ + \text{rank1 matrix} \end{bmatrix} = H'
Decorrelating

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

- Not merely decorrelated but \textit{whitened}
  - $XX^T = CMM^TC^T = S^{-1/2}E^T ESE^T ES^{-1/2} = I$

- $C$ is the \textit{whitening matrix}
Uncorrelated != Independent

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

\[ E[x_i x_j] = 0 \text{ if } i \neq j \]

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

\[ E[x_i^2 x_j^2] = E[x_i^2]E[x_j^2] \]

• This is one of the signatures of independent RVs
• Lets explicitly decouple the fourth order moments
Decorrelating

- \( X = CM \)
- \( XX^T = I \)

- Will multiplying \( X \) by \( B \) *re-correlate* the components?
- Not if \( B \) is *unitary*
  - \( BB^T = B^TB = I \)
- \( HH^T = BXX^TB^T = BB^T = I \)
- So we want to find a *unitary* matrix
  - Since the rows of \( H \) are uncorrelated
    - Because they are independent
ICA: Freeing Fourth Moments

- $H = AM, \ A = BC, \ X = CM, \ \Rightarrow \ H = BX$

- The fourth moments of $H$ have the form:
  $E[h_i h_j h_k h_l]$

- If the rows of $H$ were independent
  $E[h_i h_j h_k h_l] = E[h_i] E[h_j] E[h_k] E[h_l]$

- 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 $H$ independent and diagonalize it

• A good candidate
  – Good because it incorporates the energy in all rows of $H$

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

– Where
\[d_{ij} = E[\sum_k h_k h_i h_j]\]

– i.e.
\[D = E[h^T h h h^T]\]
  • $h$ are the columns of $H$
  • Assuming $h$ is real, else replace transposition with Hermitian
ICA: The D matrix

\[ D = \begin{bmatrix}
  d_{11} & d_{12} & d_{13} & \ldots \\
  d_{21} & d_{22} & d_{23} & \ldots \\
   & \ldots & \ldots & \ldots \\
\end{bmatrix} \]

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

- Average above term across all columns of \( H \)

Energy-weighted correlation!!
ICA: The D matrix

\[ D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \ldots \\ d_{21} & d_{22} & d_{23} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \]

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

- If the \( h_i \) terms were independent
  - For \( i \neq j \)
    \[ E\left[ \sum_k h_k^2 h_i h_j \right] = E[h_i^3]E[h_j] + E[h_j^3]E[h_i] + \sum_{k \neq i, k \neq j} E[h_k^2]E[h_i]E[h_j] \]
  - Centered: \( E[h_j] = 0 \) \( \Rightarrow \) \( E[ \sum_k h_k^2 h_i h_j] = 0 \) for \( i \neq j \)
  - For \( i = j \)
    \[ E\left[ \sum_k h_k^2 h_i h_j \right] = E[h_i^4] + E[h_i^2] \sum_{k \neq i} E[h_k^2] \neq 0 \]

- Thus, if the \( h_i \) terms were independent, \( d_{ij} = 0 \) if \( i \neq j \)
- i.e., if \( h_i \) were independent, \( D \) would be a diagonal matrix
  - Let us diagonalize \( D \)
Diagonalizing D

• Compose a fourth order matrix from $X$
  
  – Recall: $X = CM$, $H = BX = BCM$
  
  • $B$ is what we’re trying to learn to make $H$ independent

• Note: if $H = BX$, then each $h = Bx$

• The fourth moment matrix of $H$ is

$$D = E[h^T h h h^T] = E[x^T BB^T x B^T x x^T B]$$

$$= E[x^T x B^T x x^T B]$$

$$= B^T E[x^T x xx^T] B$$
Diagonalizing D

• Objective: Estimate $B$ such that the fourth moment of $H = BX$ is diagonal

• Compose $D_x = E[x^T x x x^T]$

• Diagonalize $D_x$ via Eigen decomposition
  $D_x = U\Lambda U^T$

• $B = U^T$
  – That’s it!!!!
B frees the fourth moment

\[ \mathbf{D}_x = \mathbf{U} \Lambda \mathbf{U}^T ; \quad \mathbf{B} = \mathbf{U}^T \]

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

- The fourth moment matrix of \( \mathbf{H} \) is
  \[
  \mathbb{E}[h^T h h h^T] = \mathbf{U}^T \mathbb{E}[x^T x x x^T] \mathbf{U} \\
  = \mathbf{U}^T \mathbf{D}_x \mathbf{U} \\
  = \mathbf{U}^T \mathbf{U} \Lambda \mathbf{U}^T \mathbf{U} = \Lambda
  \]
- The fourth moment matrix of \( \mathbf{H} = \mathbf{U}^T \mathbf{X} \) is Diagonal!!
Overall Solution

• $H = AM = BCM$
  – $C$ is the (transpose of the) matrix of Eigenvectors of $MM^T$

• $X = CM$

• $A = BC = U^TC$
  – $B$ is the (transpose of the) matrix of Eigenvectors of $Xdiag(X^TX)X^T$
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 $H$ are independent
  – $H = AM$

• 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 $M M^T = E S E^T$

• $C = S^{-1/2} E^T$

• Can use first $K$ columns of $E$ only if only $K$ independent sources are expected
  – In microphone array setup – only $K < M$ sources

• $X = C M$
  – $E[x_i x_j] = \delta_{ij}$ 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

\[ I(H) = \sum_{i} H(h_i) - H(h) \]

- With constraint: \( H = BX \)
  - \( X \) is “whitened” \( M \)
Linear Functions

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

\[
P_h(h) = P_x(B^{-1}h) |B|^{-1}
\]

\[
H(x) = -\int P(x) \log P(x) \, dx
\]

\[
\log P(x) = \log P_x(B^{-1}h) - \log(|B|)
\]

\[
H(h) = H(x) + \log |B|
\]
The contrast function

\[
I(H) = \sum_i H(h_i) - H(H)
\]

\[
I(H) = \sum_i H(h_i) - H(x) - \log |B|
\]

• Ignoring \(H(x)\) (Const)

\[
J(H) = \sum_i H(h_i) - \log |B|
\]

• Minimize the above to obtain \(B\)
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 $g(H) = g(BX)$ (component-wise function)

$$
\begin{array}{ccc}
g(h_{11}) & g(h_{21}) & \ldots \\
g(h_{12}) & g(h_{22}) & \\
\vdots & \vdots & \\
\vdots & \vdots & \\
\vdots & \vdots & \\
\end{array}
$$

• Define $f(H) = f(BX)$

$$
\begin{array}{ccc}
f(h_{11}) & f(h_{21}) & \ldots \\
f(h_{12}) & f(h_{22}) & \\
\vdots & \vdots & \\
\vdots & \vdots & \\
\vdots & \vdots & \\
\end{array}
$$
An alternate approach

- \( P = g(H) f(H)^T = g(BX) f(BX)^T \)

\[
\begin{bmatrix}
P_{11} & P_{21} & \cdots \\
P_{12} & P_{22} & \\
\vdots & \vdots & \\
\end{bmatrix}
\]

\( P = \)

This is a square matrix

- Must ideally be

\[
\begin{bmatrix}
Q_{11} & Q_{21} & \cdots \\
Q_{12} & Q_{22} & \\
\vdots & \vdots & \\
\end{bmatrix}
\]

\( Q = \)

- Error = \( \| P - Q \|_F^2 \)

\( P_{ij} = E[g(h_i)f(h_j)] \)

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

\( Q_{ii} = E[g(h_i)f(h_i)] \)
An alternate approach

- Ideal value for $Q$

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

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

- If $g()$ and $h()$ are odd symmetric functions
  - $E[g(h_i)] = 0$ for all $i$
    - Since $= E[h_i] = 0$ (H is centered)
    - $Q$ is a Diagonal Matrix!!!
An alternate approach

• Minimize Error

\[ P = g(BX)f(BX)^T \]

\[ Q = \text{Diagonal} \]

\[ error = \| P - Q \|_F^2 \]

• Leads to trivial Widrow Hopf type iterative rule:

\[ E = \text{Diag} - g(BX)f(BX)^T \]

\[ B = B + \eta EB^T \]
Update Rules

• Multiple solutions under different assumptions for \( g() \) and \( f() \)

• \( H = BX \)

• \( B = B + \eta \Delta B \)

• Jutten Herraut: Online update
  \[- \Delta B_{ij} = f(h_i)g(h_j); \text{ -- actually assumed a recursive neural network} \]

• Bell Sejnowski
  \[- \Delta B = ([B^T]^{-1} - g(H)X^T) \]
Update Rules

• Multiple solutions under different assumptions for \( g() \) and \( f() \)

• \( H = BX \)

• \( B = B + \eta \Delta B \)

• Natural gradient -- \( f() = \) identity function
  \[ \Delta B = (I - g(H)H^T)W \]

• Cichoki-Unbehaeiven
  \[ \Delta B = (I - g(H)f(H)^T)W \]
What are $G()$ and $H()$

- Must be odd symmetric functions
- Multiple functions proposed

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

- Audio signals in general
  - $\Delta B = (I - HH^\text{T} - K\tanh(H)H^\text{T})W$
- Or simply
  - $\Delta B = (I - K\tanh(H)H^\text{T})W$
So how does it work?

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

Input

Mix

Output

11755/18797
Another Example

• Three instruments..
• 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
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 Enhancement

- 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.
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..