Machine Learning for Signal Processing
Independent Component Analysis

Instructor: Bhiksha Raj
Revisiting the Covariance Matrix

- Assuming centered data

\[ C = \sum_X XX^T \]

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

- 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 the squared length of the corresponding vector**
Covariance Matrix as a transform

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

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

• More vectors...
• Major axis of component ellipses proportional to the squared length of the corresponding vector
Covariance Matrix as a transform

- And still more vectors..
- **Major axis of component ellipses proportional to the squared length of the corresponding vector**
Covariance Matrix as a transform

• The covariance matrix captures the directions of maximum variance
• What does it tell us about trends?
Data Trends: Axis aligned covariance

• Axis aligned covariance
• At any X value, the average Y value of vectors is 0
  – X cannot predict Y
• At any Y, the average X of vectors is 0
  – Y cannot predict X
• The X and Y components are uncorrelated
Data Trends: Tilted covariance

- Tilted covariance
- The average Y value of vectors at any X varies with X
  - X predicts Y
- Average X varies with Y
- The X and Y components are *correlated*
• 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 $X$
• Expected value of $Y$ given $X$:
  – Find average of $Y$ values of all samples at (or close) to the given $X$
  – If this is a function of $X$, $X$ and $Y$ are correlated
Uncorrelatedness

• Knowing X does not tell you what the average value of Y is
  – And vice versa
Uncorrelated Variables

- The average value of Y is the same regardless of the value of X and vice versa
Uncorrelatedness in Random Variables

- Which of the above represent uncorrelated RVs?
Benefits of uncorrelatedness..

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

• We will frequently transform correlated variables to make them uncorrelated
  – “Decorrelating” variables
The notion of *decorrelation*

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

- So how does one transform the correlated variables \((X,Y)\) to the uncorrelated \((X', Y')\)
What does “uncorrelated” mean

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

\[
E \begin{bmatrix} X' \\ Y' \end{bmatrix}E \begin{bmatrix} X' & Y' \end{bmatrix} = E \begin{bmatrix} X'^2 & X'Y' \\ X'Y' & Y'^2 \end{bmatrix} = E \begin{bmatrix} X'^2 \\ Y'^2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \text{diagonal matrix}
\]

- 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
  – For uncorrelated components, $YY^T = \text{Diagonal}$
    $\Rightarrow MXX^TM^T = \text{Diagonal}$
    $\Rightarrow M.\text{Cov}(X).M^T = \text{Diagonal}$
Decorrelation

• Easy solution:
  – Eigen decomposition of Cov(X):
    \[ \text{Cov}(X) = E \Lambda E^T \]
  – \( EE^T = I \)
• Let \( M = E^T \)

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

• PCA: \( Y = E^T X \)
  – Projects the data onto the Eigen vectors of the covariance matrix
  – *Diagonalizes* the covariance matrix
  – “Decorrelates” the data
PCA

\[ X = w_1 E_1 + w_2 E_2 \]

- Projects the data onto the Eigen vectors of the covariance matrix
  - Changes the coordinate system to the Eigen vectors of the covariance matrix
  - Diagonalizes the covariance matrix
  - "Decorrelates" the data
Decorrelating the data

• Are there other decorrelating axes?
Decorrelating the data

- Are there other decorrelating axes?
Decorrelating the data

• Are there other decorrelating axes?
Decorrelating the data

- Are there other decorrelating axes?
- What about if we don’t require them to be orthogonal?
Decorrelating the data

- Are there other decorrelating axes?
- What about if we don’t require them to be orthogonal?
- What is special about these axes?
The statistical concept of *Independence*

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

- X and Y are *independent* if knowing X tells you nothing at all of Y.
A brief review of basic probability

• **Independence**: Two random variables $X$ and $Y$ are independent iff:
  – Their joint probability equals the product of their individual probabilities
  
  $P(X,Y) = P(X)P(Y)$

• Independence implies uncorrelatedness
  – The average value of $X$ is the same regardless of the value of $Y$
    • $E[X|Y] = E[X]$
  – But uncorrelatedness does not imply independence
A brief review of basic probability

• *Independence:* Two random variables $X$ and $Y$ are independent iff:

• The average value of *any function of* $X$ *is the same regardless of the value of* $Y$
  – Or any function of $Y$

• $E[f(X)g(Y)] = E[f(X)] \cdot E[g(Y)]$ for all $f(), g()$
Independence

• Which of the above represent independent RVs?
• Which represent uncorrelated RVs?
A brief review of basic probability

• The expected value of an odd function of an RV is 0 if
  – The RV is 0 mean
  – The PDF is of the RV is symmetric around 0

• $E[f(X)] = 0$ if $f(X)$ is odd symmetric
A note on bits..

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

• How many bits will you have to send?
A note on bits..

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

• How many bits will you have to send?
A note on bits..

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

• How many bits will you have to send?
A note on bits..

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

• How many bits will you have to send?
Batching up 6-sided dice rolls

• Instead of sending individual rolls, you roll the dice *twice*
  – And send the *pair* to your friend

• How many bits do you send *per roll*?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
Batching up 6-sided dice rolls

• Instead of sending individual rolls, you roll the dice *twice*
  – And send the *pair* to your friend

• How many bits do you send *per roll*?

• 36 combinations: 6 bits per pair of numbers
  – Still 3 bits per roll

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
Batching up 6-sided dice rolls

• Instead of sending individual rolls, you roll the dice **three times**
  – And send the *triple* to your friend
• How many bits do you send *per roll*?
• 216 combinations: 8 bits per triple
  – Still 2.666 bits per roll
  – *Now we’re talking!*
Batching up 6-sided dice rolls

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

• Batching *five rolls*
  – 7776 combinations
  – 13 bits per outcome (5 rolls)
  – 2.6 bits per roll
Batching up 6-sided dice rolls

- Where will it end?
Batching up 6-sided dice rolls

- Where will it end?

- \( \lim_{k \to \infty} \frac{[k \log_2(6)]}{k} = \log_2(6) \) bits per roll in the limit
  - This is the absolute minimum – no batching will give you less than these many bits per outcome
Can we do better?

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

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

• You have

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

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• You use:

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

• What is the average number of bits per outcome
Can we do better?

• You have
  \[ P(1) = 0.5, \quad P(2) = 0.25, \quad P(3) = 0.125, \quad P(4) = 0.125 \]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>110</td>
</tr>
<tr>
<td>4</td>
<td>111</td>
</tr>
</tbody>
</table>

• You use:
  – Note receiver is *never in any doubt as to what they received*

• An outcome with probability \( p \) is equivalent to obtaining one of \( 1/p \) equally likely choices
  – Requires \( \log_2(\frac{1}{p}) \) bits on average
Entropy

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

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

You can’t do better
   - Any other code will require more bits

This is the entropy of the random variable
A brief review of basic info. theory

*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

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

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

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

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

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

$$= -\sum_{X,Y} P(X, Y) \log P(X) - \sum_{X,Y} P(X, Y) \log P(Y) = H(X) + H(Y)$$
Onward..
Projection: multiple notes

\[ M = \]

\[ W = \]

- \[ P = W (W^T W)^{-1} W^T \]
- Projected Spectrogram = PM
We’re actually computing a score

\[ M = \]

\[ W = \]

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

\[ M = \]

\[ H = \]

\[ W = ? \quad U = ? \]

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

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

- **Constraint:** \( \mathbf{W} \) is orthogonal
  - \( \mathbf{W}^T \mathbf{W} = \mathbf{I} \)

- **The solution:** \( \mathbf{W} \) are the Eigen vectors of \( \mathbf{MM}^T \)
  - PCA!!

- \( \mathbf{M} \sim \mathbf{WH} \) is an approximation

- Also, the rows of \( \mathbf{H} \) are *decorrelated*
  - Trivial to prove that \( \mathbf{HH}^T \) is diagonal
PCA

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

\[ M \approx WH \]

- 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

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

- 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..
What else can we look for?

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

• Attempting to find statistically independent components of the mixed signal
  – Independent Component Analysis
Formulating it with Independence

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

- 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

\[ M = WH \quad \ldots \quad H = AM \]

• 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?
An ugly algebraic solution

• We \textit{decorrelated} signals by diagonalizing the covariance matrix

• \textit{Is there a simple matrix we could just similarly diagonalize to make them independent?}
An ugly algebraic solution

• We *decorrelated* signals by diagonalizing the covariance matrix

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

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
- $E[h_i h_j h_k h_l] = E[h_i]E[h_j]E[h_k]E[h_l]$
- $E[h_i^2 h_j h_k] = E[h_i^2]E[h_j]E[h_k]$
- $E[h_i h_j h_l^2] = E[h_i]E[h_j^2]$
- Etc.
Zero Mean

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

• \( \mathbf{M} = \mathbf{W}\mathbf{H} \quad \mathbf{H} = \mathbf{A}\mathbf{M} \)

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

\[
\mu_m = \frac{1}{\text{cols} (\mathbf{M})} \sum_i \mathbf{m}_i
\]

\[
\mathbf{m}_i = \mathbf{m}_i - \mu_m \quad \forall i
\]

– \( \mathbf{m}_i \) are the columns of \( \mathbf{M} \)
Emulating Independence..

• Independence $\rightarrow$ Uncorrelatedness
• Find $C$ such that $CM$ is decorrelated
  – PCA
• Find $B$ such that $B(CM)$ is independent
• A little more than PCA
Decorrelating and Whitening

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

- Not merely decorrelated but \textit{whitened}
  - $\mathbf{X X}^T = \mathbf{C M M}^T \mathbf{C}^T = \mathbf{S}^{-1/2} \mathbf{E}^T \mathbf{E S E}^T \mathbf{E S}^{-1/2} = \mathbf{I}$

- $\mathbf{C}$ 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

$H = AM$
$A = BC$
$H = BCM$
$H = BX$
FOBI: Freeing Fourth Moments

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

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

- If the rows of $\mathbf{H}$ 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 $\mathbf{B}$ such that the fourth moments of $\mathbf{H} = \mathbf{B} \mathbf{X}$ are decoupled
  - While ensuring that $\mathbf{B}$ is Unitary

- **FOBI: Fourth Order Blind Identification**
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: the weighted correlation matrix of $H$

$$D = E[\|h\|^2 hh^T] = \sum_k \|h_k\|^2 h_k h_k^T$$

- $h$ are the columns of $H$
- Assuming $h$ is real, else replace transposition with Hermitian
ICA: The D matrix

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

\[ D = E[\|h\|^2 hh^T] \]

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

On the actual matrix

\[ D = \sum_k \|h_k\|^2 h_k h_k^T \]

\[ d_{ij} = \frac{1}{cols(H)} \sum_k \left( \sum_l h_{kl}^2 \right) h_{ki} h_{kj} \]
ICA: The D matrix

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

\[
D = E[\|h\|^2 hh^T]
\]

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

- If the \(h_i\) terms were independent and zero mean
- For \(i \neq j\)

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

- For \(i = j\)

\[
- E[h_i h_j \sum_l h_l^2] = E[h_i^4] + E[h_i^2] \sum_{l \neq i} E[h_i^2] \neq 0
\]

- i.e., if \(h_i\) were independent, \(D\) would be a diagonal matrix
  - Let us diagonalize \(D\)
Diagonalizing D

• Recall: \( \mathbf{H} = \mathbf{B}\mathbf{X} \)
  
  – \( \mathbf{B} \) is what we’re trying to learn to make \( \mathbf{H} \) independent
  
  – Assumption: \( \mathbf{B} \) is unitary, i.e. \( \mathbf{BB}^\top = \mathbf{I} \)

• Note: if \( \mathbf{H} = \mathbf{B}\mathbf{X} \), then each vector \( \mathbf{h} = \mathbf{B}\mathbf{x} \)

• The fourth moment matrix of \( \mathbf{H} \) is

\[
\mathbf{D} = E[\mathbf{h}^\top \mathbf{h} \mathbf{h}^\top] = E[\mathbf{x}^\top \mathbf{BB}^\top \mathbf{x} \mathbf{B}^\top \mathbf{x} \mathbf{x}^\top \mathbf{B}]
= E[\mathbf{x}^\top \mathbf{x} \mathbf{B}^\top \mathbf{x} \mathbf{x}^\top \mathbf{B}]
= \mathbf{B}^\top E[\mathbf{x}^\top \mathbf{x} \mathbf{x}^\top] \mathbf{B}
= \mathbf{B}^\top E[||\mathbf{x}||^2 \mathbf{x}^\top] \mathbf{B}
\]

• Objective: Find a matrix \( \mathbf{B} \) such that the rows of \( \mathbf{H} = \mathbf{B}\mathbf{X} \) are statistically independent

Define a matrix \( \mathbf{D} \) that would be diagonal if the rows of \( \mathbf{B}\mathbf{X} \) are independent

Compute \( \mathbf{B} \) such that this matrix becomes diagonal
**Diagonalizing D**

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

- **Compose** $\mathbf{D}_x = \sum_k \|\mathbf{x}_k\|^2 \mathbf{x}_k \mathbf{x}_k^T$

- **Diagonalize** $\mathbf{D}_x$ via Eigen decomposition
  
  $\mathbf{D}_x = \mathbf{U}\Lambda\mathbf{U}^T$

- **$\mathbf{B} = \mathbf{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)
- \( \mathbf{H} = \mathbf{B}\mathbf{X} = \mathbf{U}^T\mathbf{X} \)
- \( \mathbf{h} = \mathbf{U}^T\mathbf{x} \)

- The fourth moment matrix of \( \mathbf{H} \) is
  \[ \mathbb{E}[||\mathbf{h}||^2 \mathbf{h} \mathbf{h}^T] = \mathbf{U}^T \mathbb{E}[||\mathbf{x}||^2 \mathbf{x}\mathbf{x}^T] \mathbf{U} \]
  \[ = \mathbf{U}^T \mathbf{D}_x \mathbf{U} \]
  \[ = \mathbf{U}^T \mathbf{U} \Lambda \mathbf{U}^T \mathbf{U} = \Lambda \]
- The fourth moment matrix of \( \mathbf{H} = \mathbf{U}^T\mathbf{X} \) is Diagonal!!
Overall Solution

• Objective: Estimate $A$ such that the rows of $H = AM$ are independent

• Step 1: *Whiten $M$*
  – $C$ is the (transpose of the) matrix of Eigen vectors of $MM^T$
  – $X = CM$

• Step 2: Free up fourth moments on $X$
  – $B$ is the (transpose of the) matrix of Eigenvectors of $X.diag(X^TX).X^T$
  – $A = BC$
FOBI for ICA

• 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
     \[
     R_{MM} = E S E^T, \quad C = S^{-1/2} E^T
     \]
  4. Compute \( X = CM \)
  5. Compute the fourth moment matrix \( D' = E[|x|^2 xx^T] \)
  6. Diagonalize \( D' \) via Eigen decomposition
     \[
     D' = U \Lambda U^T
     \]
  7. \( D' = U \Lambda U^T \)
  8. Compute \( A = U^T C \)

• 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

• FOBI is not perfect
  – Only a subset of fourth order moments are considered
  • Diagonalizing the particular fourth-order moment matrix we have chosen is not guaranteed to diagonalize every other fourth-order moment matrix

• JADE: (Joint Approximate Diagonalization of Eigenmatrices), J.F. Cardoso
  – Jointly diagonalizes multiple fourth-order cumulant matrices
Enforcing Independence

• Specifically ensure that the components of $H$ 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 $\mathbf{MM}^T = \mathbf{ESE}^T$

• $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{E}^T$

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

• $\mathbf{X} = \mathbf{CM}$
  – $\mathbf{E}[\mathbf{x}_i\mathbf{x}_j] = \delta_{ij}$ for centered signal
The contrast function

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

• An explicit contrast function

\[ I(H) = \sum_i H(\bar{h}_i) - H(\bar{h}) \]

• 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 *unmixing* matrix

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

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

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

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

\[ I(H) = \sum_i H(\overline{h}_i) - H(\overline{h}) \]

\[ I(H) = \sum_i H(\overline{h}_i) - H(x) - \log |B| \]

- Ignoring \( H(x) \) (Const)

\[ J(H) = \sum_i H(\overline{h}_i) - \log |B| \]

- Minimize the above to obtain \( B \)
An alternate approach

• Recall PCA

• $\mathbf{M} = \mathbf{WH}$, the columns of $\mathbf{W}$ must be orthogonal

• Leads to: $\min_{\mathbf{W}} \| \mathbf{M} - \mathbf{WW}^\top \mathbf{M} \|^2 + \Lambda \text{trace}(\mathbf{W}^\top \mathbf{W})$
  
  – Error minimization framework to estimate $\mathbf{W}$

• Can we arrive at an error minimization framework for ICA

• Define an “Error” objective that represents independence
An alternate approach

- Definition of Independence – if $x$ and $y$ are independent:
  - $E[f(x)g(y)] = E[f(x)]E[g(y)]$
  - Must hold for every $f()$ and $g()$!!
An alternate approach

• Define $g(H) = g(BX)$ (component-wise function)

<table>
<thead>
<tr>
<th>$g(h_{11})$</th>
<th>$g(h_{21})$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g(h_{12})$</td>
<td>$g(h_{22})$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

• Define $f(H) = f(BX)$

<table>
<thead>
<tr>
<th>$f(h_{11})$</th>
<th>$f(h_{21})$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(h_{12})$</td>
<td>$f(h_{22})$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
An alternate approach

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

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

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

This is a square matrix

- \( \mathbf{Q} = \begin{bmatrix}
Q_{11} & Q_{21} & \cdots \\
Q_{12} & Q_{22} & \\
\cdots & \cdots & \\
\cdots & \cdots & \\
\cdots & \cdots & \\
\end{bmatrix} \)

- \( \mathbf{Q}_{ij} = \mathbf{E}[g(h_i)]\mathbf{E}[f(h_j)] \quad i \neq j \)

- \( \mathbf{Q}_{ii} = \mathbf{E}[g(h_i)f(h_i)] \)

- Error = \( \|\mathbf{P}-\mathbf{Q}\|_F^2 \)
An alternate approach

- Ideal value for $\mathbf{Q}$

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

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

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

- If $g()$ and $f()$ are odd symmetric functions

\[
E[g(h_i)] = 0 \text{ for all } i
\]

- Since $= E[h_i] = 0$ (H is centered)

- $\mathbf{Q}$ 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:

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

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

• Multiple solutions under different assumptions for g() and f()
• \( H = BX \)
• \( B = B + \eta \Delta B \)
• Jutten Herraut : Online update
  \(- \Delta B_{ij} = f(h_i)g(h_j); \quad \text{-- actually assumed a recursive neural network} \)
• 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
  - \( \Delta B = (I - g(H)H^T)X^T \)
- Cichoki-Unbehaeven
  - \( \Delta B = (I - g(H)f(H)^T)X^T \)
What are $G()$ and $F()$

- Must be odd symmetric functions
- Multiple functions proposed

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

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

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

• Three instruments..
• Three instruments..
ICA for data exploration

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

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

Non-Gaussian data
Finding useful transforms with ICA

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