A brief review of basic probability

**Uncorrelated:**

- Two random variables $X$ and $Y$ are uncorrelated if
  \[ \text{cov}(X,Y) = 0. \]

- The average value of the product of the variables equals the product of their individual averages:
  \[ \text{E}[XY] = \text{E}[X] \cdot \text{E}[Y]. \]

**Independent:**

- Two random variables $X$ and $Y$ are independent if
  \[ \text{P}(X,Y) = \text{P}(X) \cdot \text{P}(Y). \]

- Their joint probability equals the product of their individual probabilities:
  \[ \text{P}(X,Y) = \text{P}(X) \cdot \text{P}(Y). \]

- The average value of any function $f(X,Y)$ is the same regardless of the value of $Y$:
  \[ \text{E}[f(X,Y)] = \text{E}[f(X)]. \]

- The average value of any function $g(Y)$ is the same regardless of the value of $X$:
  \[ \text{E}[g(Y)] = \text{E}[g(Y)]. \]

- Their joint probability equals the product of their individual probabilities:
  \[ \text{P}(X,Y) = \text{P}(X) \cdot \text{P}(Y). \]

- The average value of their product is the product of their individual averages:
  \[ \text{E}[XY] = \text{E}[X] \cdot \text{E}[Y]. \]

Which of the above represent uncorrelated RVs?

Which of the above represent independent RVs?

Which of the above represent uncorrelated and independent RVs?

Which of the above represent independent RVs?

**Analysis:**

Instructor: Bhiksha Raj

Class 20. 8 Nov 2012
A brief review of basic probability

The expected value of an odd function of an RV is 0 if the PDF is of the RV is symmetric around 0.

A brief review of basic info. theory

Joint entropy: The minimum average number of bits to convey a symbol

Conditional entropy: The minimum average number of bits to convey a symbol, after symbol X has already been conveyed

The entropies of X and Y if they are independent

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

A brief review of basic info. theory

To convey sets (pairs here) of symbols

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

The expected value of an odd function of an RV is 0.
We're actually computing a score

\[ M = H = \frac{11755}{18797} \]

\[ W = \text{pinv}(W)M \]

\[ \text{for} \ W \text{ that minimizes the error} \]

\[ \text{arg min}_{W,H} \| M - WH \| \]

\[ \text{a least squares solution} \]

\[ \text{When both parameters are unknown} \]

\[ H = \text{approx}(H) \]

\[ W = \text{approx}(W) \]

\[ \text{must estimate both} \ H \text{and} \ W \text{to best approximate} \ M \]

\[ \text{ideally, must learn both the notes and their transcription of given notes} \]

\[ \text{for any} \ W \text{ that minimizes the error,} \ W' = W \text{ and} \ H' = A^{-1}H \]

\[ \text{for any invertible} \ A \]

\[ \text{For our problem, let's consider the "truth"..} \]

\[ H \text{ is in an approximation} \]

\[ \text{When one note occurs, the other does not} \]

\[ H_i \cdot H_j = 0 \text{ for all} \ i \neq j \]

\[ \text{the rows of} \ H \text{ are uncorrelated} \]

\[ \text{for all} \ i \neq j \]

\[ \text{Must ideally find the notes corresponding to the}\]

\[ \text{given}\ H, \text{estimate} \ W \text{ to minimize error} \]

\[ \text{an approximation} \]

\[ \text{Going the other way..} \]

\[ W \text{ is in an approximation} \]

\[ H \text{ is in an approximation} \]

\[ \text{So what are we doing here?} \]

\[ \text{We're actually computing a score} \]

\[ \text{How about the other way?} \]

\[ \text{Must estimate both} \ H \text{ and} \ W \text{ to best} \]
A least squares solution

Assume: $HH^T = I$

Normalizing all rows of $H$ to length 1:

Projecting $M$ onto $H$

$W = M \text{pinv}(H) = MH^T$

$$WH = M H^T H$$

$W$ minimizes least squares error with the constraint that the rows of $H$ are length 1 and orthogonal to

Since $HH^T = I$ and orthogonal to

$$W_i = \frac{1}{\sqrt{\text{trace}(HH^T)}} H_i$$

are mutually orthogonal to

Equivalences

$$W_i^T W_j = 0$$

Note: $W_i^T W_i = 1$

Minimize least squares error with the constraint

$$\|H_W W - W\|^2_2 = \text{min}$$

Differentiating and equating to 0:

$$W = H$$

So how does that work?

There are 12 notes in the segment; hence we try

Finding the notes

The first three “notes” and their contributions

The spectrograms of the notes are statistically uncorrelated

The notes in the segment, hence we try

Finding the notes

The first three “notes” and their contributions

The spectrograms of the notes are statistically uncorrelated
Finding the notes

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

\[ \gamma(t) = \sum_{i=1}^{2} \alpha_i(t) \beta_i(t) \]

\( \text{Solving the above with the constraints that the columnsof W are orthonormal gives you the eigen vectorsof the data in M} \)

**Formulating it with Independence**

\[ \text{impose statistical independence constraints on decomposition} \]

\[ W^* = \arg \min \| H - WH \|_F \]

**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, playing independently of one another or in a multi-instrument piece, instruments are dependent on what else is playing
- Our notes are not orthogonal

**Finding the notes**

To estimate 12 notes,

There are 12 notes in the segment, hence we try

\[ \text{Correlation}(W) = \max \text{Correlation}(W) \]

\( W \) is the desired solution

More generally, simple orthogonality will not give

- Harmonic continuance to associate to previous note
- Notes occur concurrently
- Overlapping frequencies

\[ \text{Our notes are not orthogonal} \]

**Changing problems for a bit**

\[ \text{Recorded by two microphones} \]

\[ \text{E} \text{h} \text{d} \text{d} \text{i} \text{l} \text{i} \text{f} \text{b} \text{t} \text{h} \text{i} \text{l} \]

\[ \text{E} \text{h} \text{c} \text{h} \text{r} \text{e} \text{c} \text{d} \text{e} \text{d} \text{s} \]

\[ \text{i} \text{s} \text{g} \text{n} \text{a} \text{l} \text{s} \]

\[ \text{Two people speak simultaneously} \]

So how does this work?
Imposing Statistical Constraints

\( W \hat{H} = \hat{M} \)

First step of ICA: Set the mean of \( M \) to 0

0 = \( \text{mean}(M) \) = \( \text{mean}(W \hat{H}) \)

\( W \) and \( \hat{H} \) are independent

Statistical independence

\( W \) and \( \hat{H} \) are independent

If \( W = H \), \( \text{mean}(M) \) = 0

Multiple approaches...

Eigenvectors of the covariance matrix

Statistical independence

\( W \) and \( H \) are independent

Given any \( W \) estimate \( H \)

Given any \( H \) estimate \( W \)

\( W \) and \( H \) are independent

Signal from speaker 1

\( M = WH \)

Signal from microphone 1

\( W = \text{notes} \)

\( H = \text{transcription} \)

Signal from speaker 2

\( W = \text{mixed} \)

Signal at microphone 1

\( W = \text{notes} \)

Signal at microphone 2
ICA: Freeing Fourth Moments

Let \( X = C \mathbf{M} \) be the whitened version of \( X \).

The process of whitening \( X \) is called whitening, which is the first step in the whitening process.

Let \( C = \mathbf{S} \) be a white noise matrix.

The whitening matrix \( \mathbf{W} \) is defined as 

\[
\mathbf{W} = \mathbf{S}^{-1/2}
\]

and \( \mathbf{X} = \mathbf{W} \mathbf{X} \).

The fourth moments of \( \mathbf{X} \) are decoupled.

The fourth moments of \( \mathbf{X} \) are decoupled.

This is one of the signatures of independent RVs.

This does not ensure higher order moments are also

Uncorrelated. 

Therefore, the process of whitening \( X \) is called whitening.

The process of whitening \( X \) is called whitening.

For the process of whitening \( X \) is called whitening.

The process of whitening \( X \) is called whitening.

The process of whitening \( X \) is called whitening.
Note that the autocorrelation matrix of \( H \) will also be diagonal.

The fourth moment matrix of \( H = \text{vec} \left( X \right) \) is diagonal:

\[
V = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}
\]

\[
\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
\]

The fourth moment matrix of \( \Sigma \) is:

\[
D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
\]

The fourth moment matrix of \( \Sigma \) is diagonal.

Procedure:

1. Compute \( V = X \Sigma X^T \).
2. Compute \( B = U^T V U \).
3. Compute \( C = S \Sigma E \).
4. Obtain \( C \) via Eigendecomposition.
5. Compute the whitening matrix \( D' = U \Sigma U^T \).
6. Obtain independent components via Eigendecomposition.

Goal: to derive a matrix \( A \) such that the rows of \( AM \) are independent components analyses.

Overall Solution

There's III

1. \( B = U^T \).
2. \( D = \Sigma V \).

Diagonalize \( D \) via Eigendecomposition.

Compose \( D' = \Sigma [X X X] \).

Recall: \( X = CM, \quad H = BX = BCM \).

Compose \( D \) a fourth order matrix from \( X \).

Diagonalizing D

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

\[
\frac{1}{m} \sum_{i=1}^{m} \frac{1}{n} \sum_{j=1}^{n} \left[ \frac{1}{p} \sum_{k=1}^{p} \frac{1}{q} \sum_{l=1}^{q} H_{ijkl} \right] = D
\]

IC: The D matrix
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.
- Enforcing independence of the output components can often be achieved in practice.
- Linear functions of the mixing matrix are not considered.
- The procedure is only efficient for a small number of sources.

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

$|\mathbf{W}| \log - (\mathbf{y}^t \mathbf{H}^t \mathbf{y}) = (\mathbf{h})^t$

Jointly diagonalizes several fourth-order moment matrices

$\exp(-\log(b) \mathbf{H} \mathbf{y} = (\mathbf{h})^t$

Individual columns of the $\mathbf{H}$ and $\mathbf{x}$ matrices are independent.

$\mathbf{y} = \mathbf{Bx}$

Linear Functions

$\mathbf{h} = \mathbf{Bx}$

$\text{diag} \left( \text{betw. \text{Lin}} \right) \left( \text{and Lin} \right)$

$\text{diag} \left( \text{Lin} \right)$

$|\mathbf{W}| \log - (\mathbf{y}^t \mathbf{H}^t \mathbf{y}) = (\mathbf{h})^t$

Contrast function: An non-linear function that has

An explicit contrast function

$|\mathbf{B}| \log + (\mathbf{x})^t \mathbf{H} (\mathbf{y})^t

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

$|\mathbf{B}| \log - (\mathbf{x})^t \mathbf{H} (\mathbf{y})^t

(\mathbf{H})^t (\mathbf{y})^t \mathbf{y} = (\mathbf{h})^t

The contrast function

Minimize the above to obtain $\mathbf{H}$.

$\text{diag} \left( \text{Lin} \right)$

$\text{diag} \left( \text{Lin} \right)$

$\text{diag} \left( \text{Betw. Lin} \right)$

$|\mathbf{W}| \log - (\mathbf{y}^t \mathbf{H}^t \mathbf{y}) = (\mathbf{h})^t$

Contrast function: A non-linear function that has

An explicit contrast function

$|\mathbf{B}| \log + (\mathbf{x})^t \mathbf{H} (\mathbf{y})^t

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

$|\mathbf{B}| \log - (\mathbf{x})^t \mathbf{H} (\mathbf{y})^t

(\mathbf{H})^t (\mathbf{y})^t \mathbf{y} = (\mathbf{h})^t

The contrast function

Minimize the above to obtain $\mathbf{H}$.

$\text{diag} \left( \text{Lin} \right)$

$\text{diag} \left( \text{Lin} \right)$

$\text{diag} \left( \text{Betw. Lin} \right)$

$|\mathbf{W}| \log - (\mathbf{y}^t \mathbf{H}^t \mathbf{y}) = (\mathbf{h})^t$

Contrast function: A non-linear function that has

An explicit contrast function

$|\mathbf{B}| \log + (\mathbf{x})^t \mathbf{H} (\mathbf{y})^t

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

$|\mathbf{B}| \log - (\mathbf{x})^t \mathbf{H} (\mathbf{y})^t

(\mathbf{H})^t (\mathbf{y})^t \mathbf{y} = (\mathbf{h})^t

The contrast function

Minimize the above to obtain $\mathbf{H}$.

$\text{diag} \left( \text{Lin} \right)$

$\text{diag} \left( \text{Lin} \right)$

$\text{diag} \left( \text{Betw. Lin} \right)$

$|\mathbf{W}| \log - (\mathbf{y}^t \mathbf{H}^t \mathbf{y}) = (\mathbf{h})^t$

Contrast function: A non-linear function that has

An explicit contrast function

$|\mathbf{B}| \log + (\mathbf{x})^t \mathbf{H} (\mathbf{y})^t

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

$|\mathbf{B}| \log - (\mathbf{x})^t \mathbf{H} (\mathbf{y})^t

(\mathbf{H})^t (\mathbf{y})^t \mathbf{y} = (\mathbf{h})^t

The contrast function

Minimize the above to obtain $\mathbf{H}$.
An alternate approach

\[ A = B \]

\[ E = \text{Diag}(B(Bx)(Bx)^T) \]

leads to trivial Widrow Hopfield iterative rules:

\[ \theta \|O - P\|_F = \theta \]

Minimize error

This is a square matrix

\[ (\gamma y)(\gamma y) \sum y = \gamma O \]

\[ f \not= ! \quad (\gamma y) \int \sum y = \gamma O \]

Must ideally be

Define (H) = (Bx) (component-wise function)

Define (H) = (H) \sum (H) = \sum (H) \sum = \sum \]

An alternate approach

\[ \text{Error} = \| \sum \not= \gamma O \]

An alternate approach

Independence

Define an “error” objective that represents

 framework for ICA

Can we arrive at an error minimization

leads: \[ |E| \sum (B) - W |E| = \sum (B) - W \]

Independent

the columns of \( W \) must be statistically

Recall PCA
Update Rules

- Multiple solutions under different assumptions for \( g() \) and \( f() \)
- \( H = BX \)
- Jutten Herault: Online update
  \[ \Delta B = \eta \nabla_{B} \mathbb{E} \]
  \[ \nabla_{B} \mathbb{E} = (f(B^T h) - g(h^T X)) X \]
- Bell Sejnowski
  \[ \Delta B = (f(B^T h) - g(h^T X)) X \]

What are \( G() \) and \( H() \)?

- Must be odd symmetric functions
- Multiple functions proposed
- \( g(x) = \sqrt{1 + x^2} \)
- Audio signals in general
  \[ \Delta B = (1 - H^T K^{-1} H) W \]
- Of simply
  \[ \Delta B = (1 - K^{-1} H) W \]

Another Example

- Three instruments.

Another Example

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

So how does it work?

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

Aanother example!

- Input
  - Mix
  - Output
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
- XA
- YA
ICA can be used to separate them out.

EEG signals are frequently corrupted by heartbeats and physiological signals.

Very commonly used to enhance EEG signals.

ICA for Signal Enhancement

Finding useful transforms with ICA

Example case: ICA-faces vs. Eigenfaces

ICA vs PCA bases

Three instruments:

so can ICA be used to do

Very successfully used

ideally notes

blocks.

represent the "building blocks" in PCA.

The Notes
So how does that work?

- There are 12 notes in the segment, hence we try to estimate 12 notes. If one note plays, other notes are not playing.
- Notes are not independent.
- Still this didn't affect the three instruments case.
- Not symmetric – negligible values never happen.
- Note energy here around mean.
- Assume distribution of signals is symmetric.

PCA solution

- Better.
- But not much.
- But the issues here?

ICA issues

- No sense of order.
- Unlike PCA scaling the signal does not affect independence.
- Outputs are scaled versions of desired signals.
- Scaling the signal does not affect independence after permutation.
- So the sources can come in any order.
- Get K independent directions, but does not have a notion.
- No sense of order.

Continue in next class.

NMF

Factor analysis.

What else went wrong?

- Assumed distribution of signals is symmetric.
- Note energy here.

So how does this work? ICA solution.

But not much.

Better.

So what does that work?

So how does this work? PCA solution.