Machine Learning and NeuroEngineering

机器学习与神经工程

Lecture 6 – EEG signal process & Probability Review

Quanying Liu （刘泉影）

SUSTech, BME department

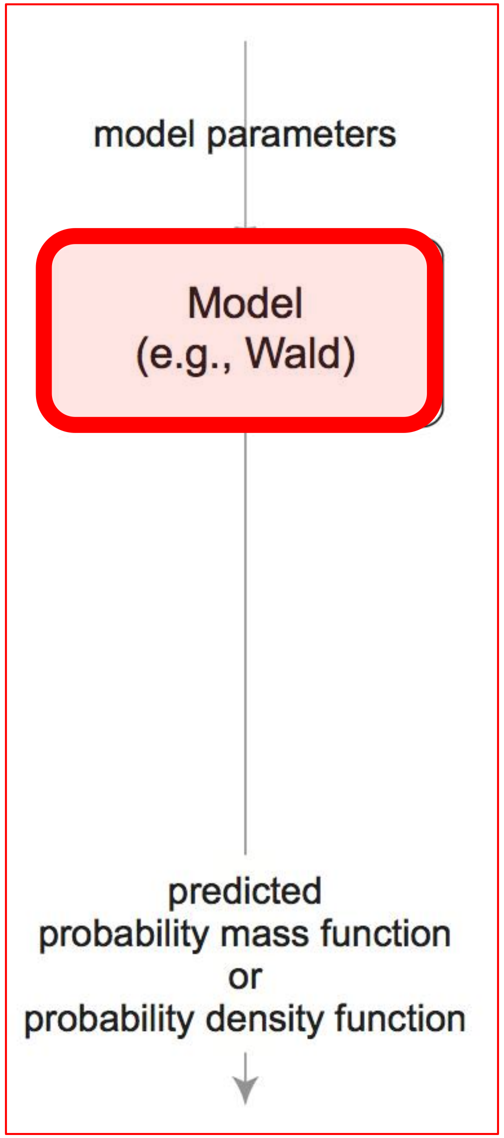
Email: liuqy@sustech.edu.cn

BMEB212, Spring term, 2021

Lecture 5 – Recap

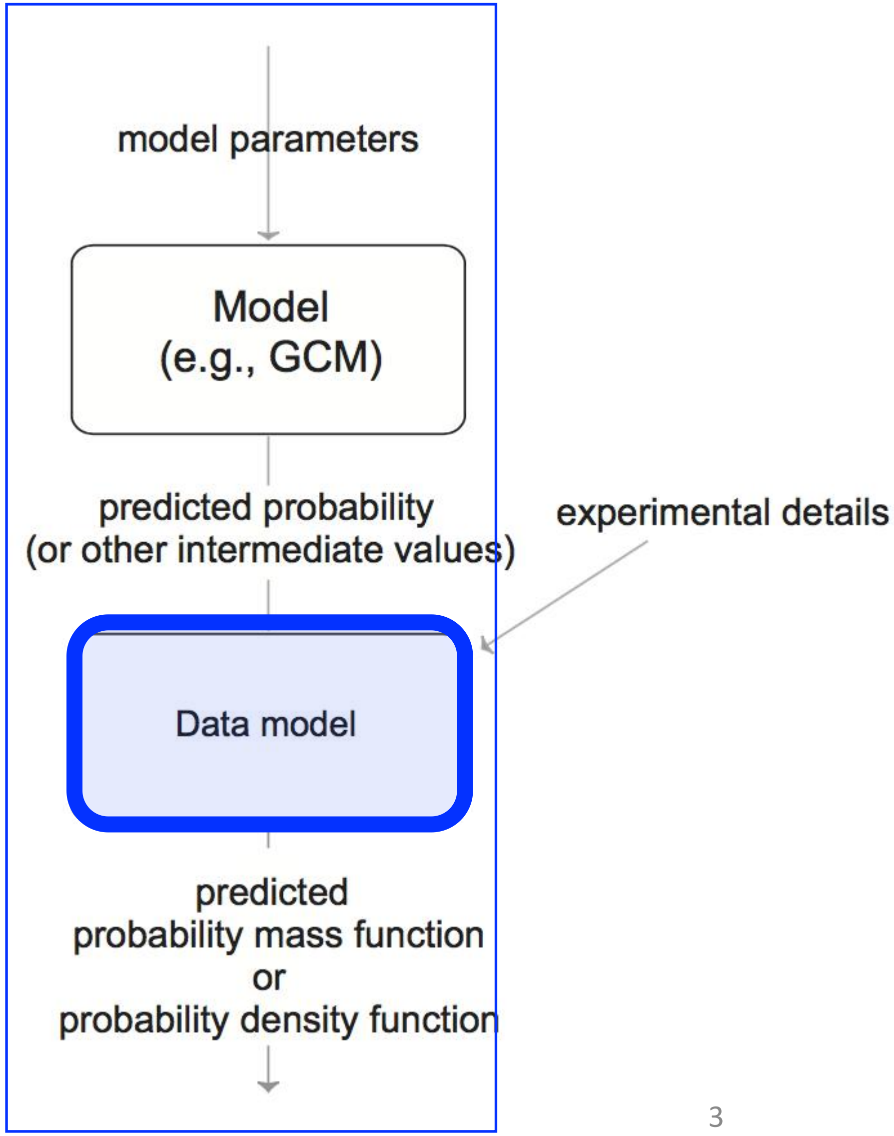
- Motivations for MLE
- Basics of Probabilities: properties of probability; probability functions
- What is a likelihood $P(\boldsymbol{\theta} \mid data)$
- Defining a Probability Distribution
  - Specified by the NeuroPsychological Model
  - Specified by by Data Models (binomial distribution)
- Finding the Maximum Likelihood

Two types of probability functions



Directly relate the model to the data

GCM with a binomial data model



Lecture 6 (part 1) – EEG data analysis

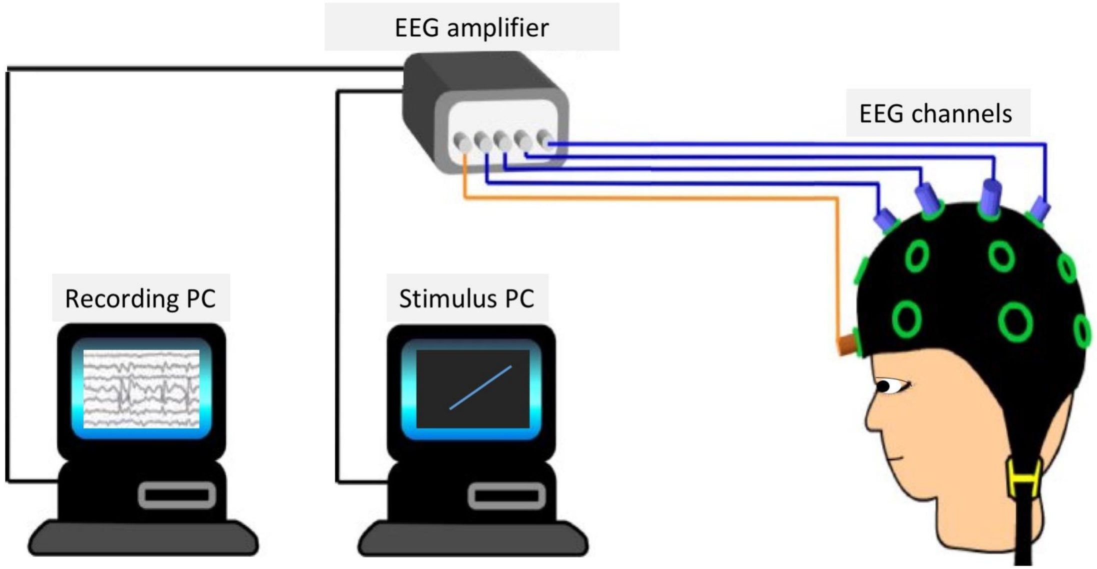
- Electroencephalography (EEG) measurements
- EEG preprocessing
  - Bad channel detection/repairmen
  - Filtering
  - Artifact removal
  - Re-referencing
- EEG sensor-level analysis
  - ERP
  - ERD/ERS

EEG measurements

Electroencephalography (EEG) is an electrophysiological process to record the electrical activity of the brain.

EEG measures changes in the electrical activity of the brain produced. Voltage changes come from **ionic current** within and between some brain cells (neurons).

An illustration of typical EEG equipment



Low-density EEG

Pros:
Cheap
Portable
Easy to setup

Cons:
Lost information

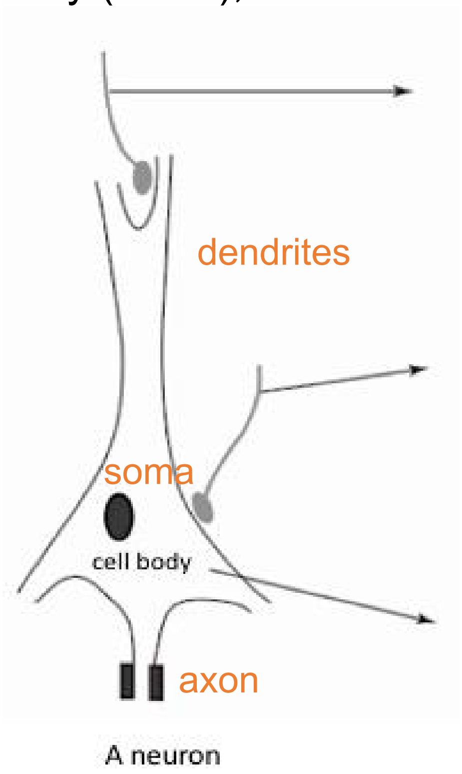
High-density EEG

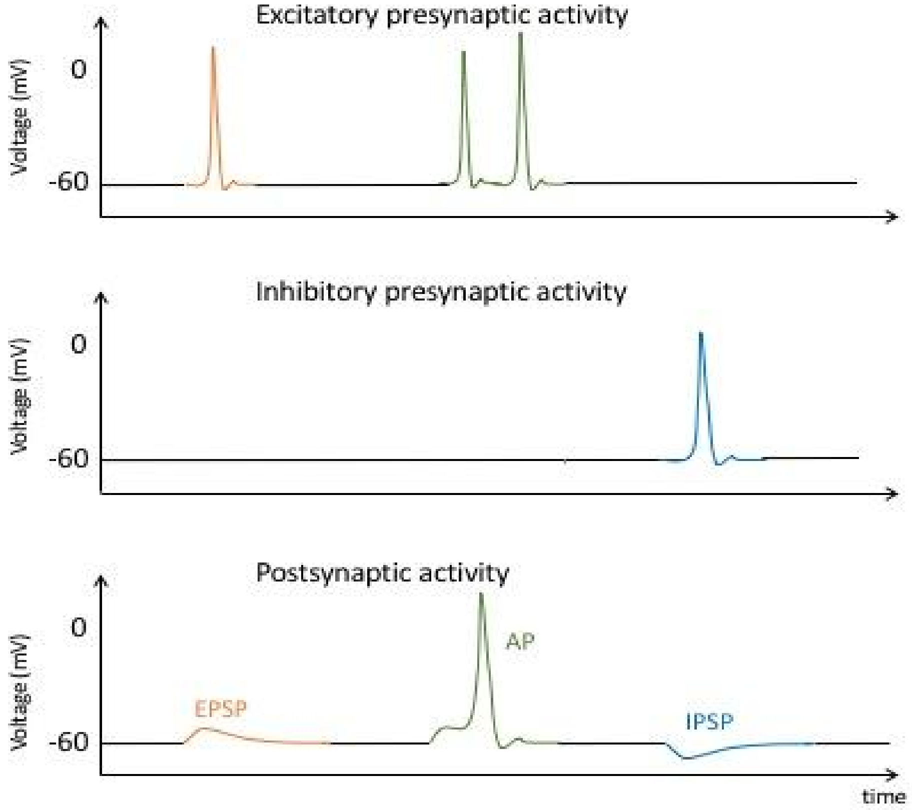
Pros:
Rich information

Cons:
Tedious to setup
Heavy to analyze data

EPSP and IPSP

A neuron consists of dendrites, cell body (soma), and axon.





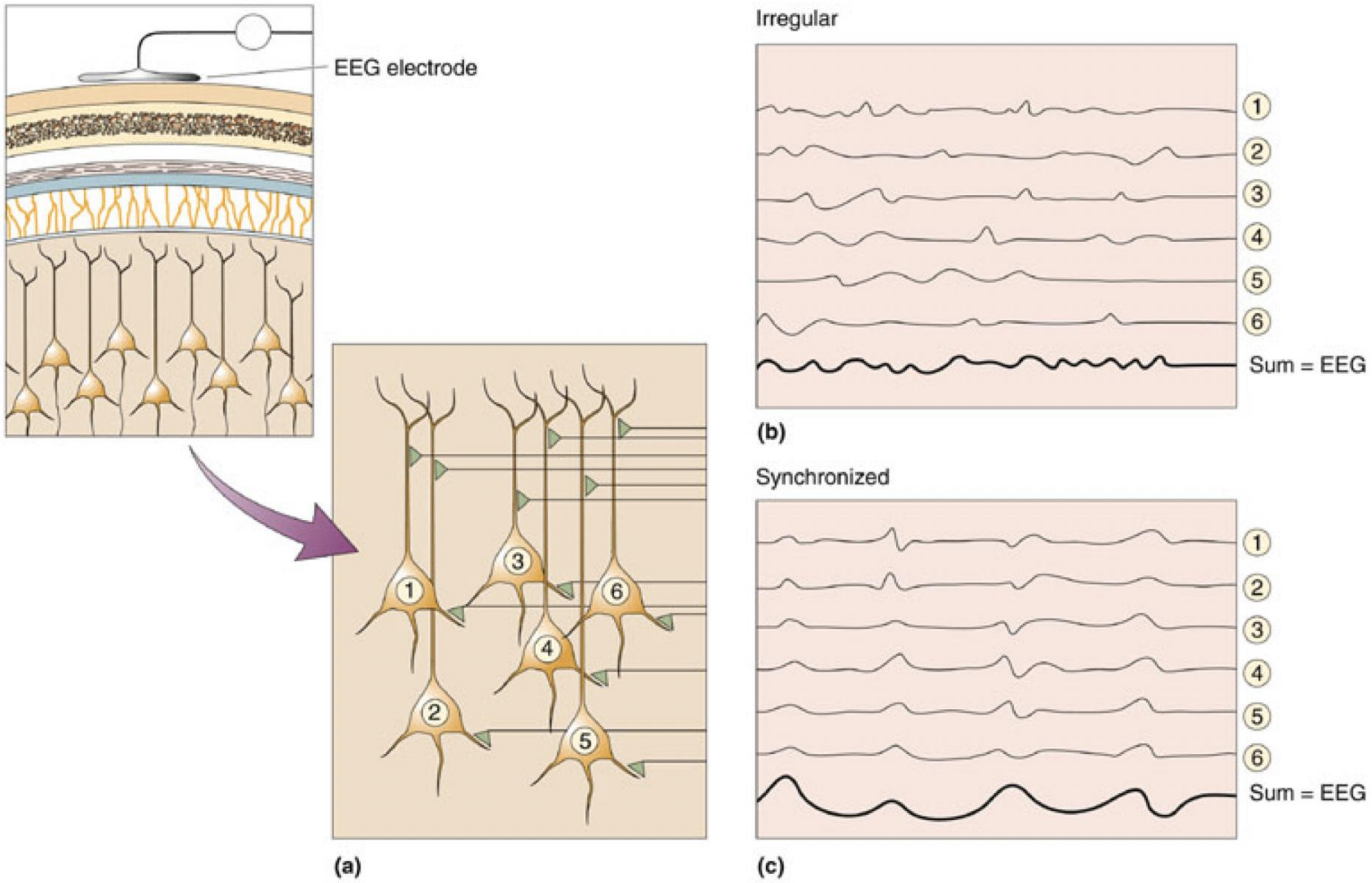
Activity measured at different locations of the neuron

If an action potential travels along the fiber, which ends in an **excitatory** synapse, an excitatory postsynaptic potential (EPSP) occurs in the following neuron.

If two or more action potentials travel along the same fiber over a short distance, there will be a summation of EPSPs producing an action potential (AP) on the postsynaptic neuron which may reach a certain threshold of membrane potential.

If the fiber ends in an **inhibitory** synapse, then hyperpolarization will occur, indicating an inhibitory postsynaptic potential (IPSP).

Generation of EEG signals



The generation of EEG signals by **synchronous/asynchronous activity** in neurons.

(a) pyramidal cells under an EEG electrode, each neuron receives many synaptic inputs.

(b) If the inputs fire at **irregular** intervals, the pyramidal cell responses are not synchronized, and the summed activity detected by the electrode may have relatively small amplitude and high frequency.

(c) If the same number of inputs fire within a narrow time window so that the pyramidal cell responses are **synchronized**, the resulting EEG may have relatively high amplitude and low frequency.

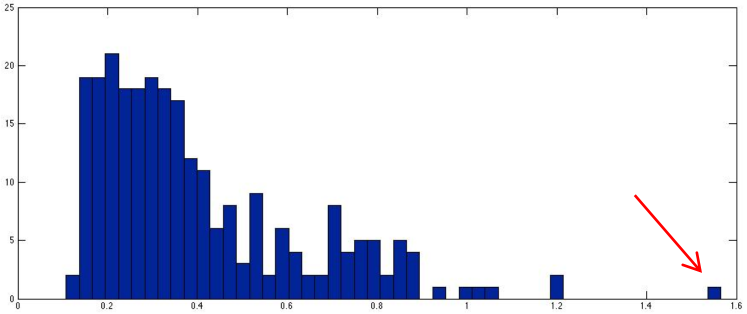
EEG preprocessing

EEG preprocessing – bad channel detection

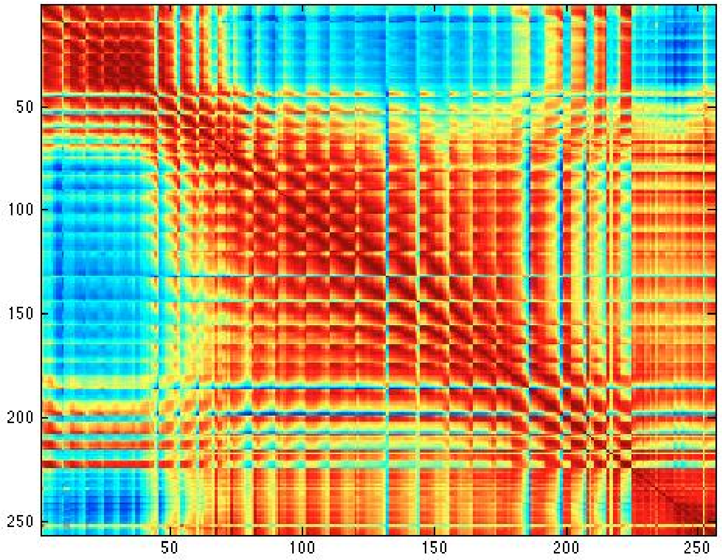
Bad channel repair → Filtering → Artifact removal → Re-referencing

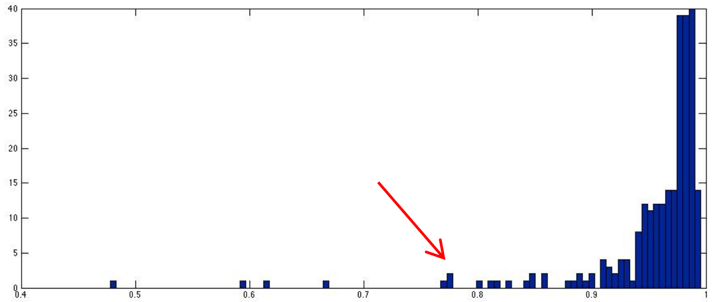
1. Bad channel detection
2. Channel interpolation

Covariance of signals in 200-250Hz



Correlation matrix between signals (0.5hz-99.5Hz)





Ref:

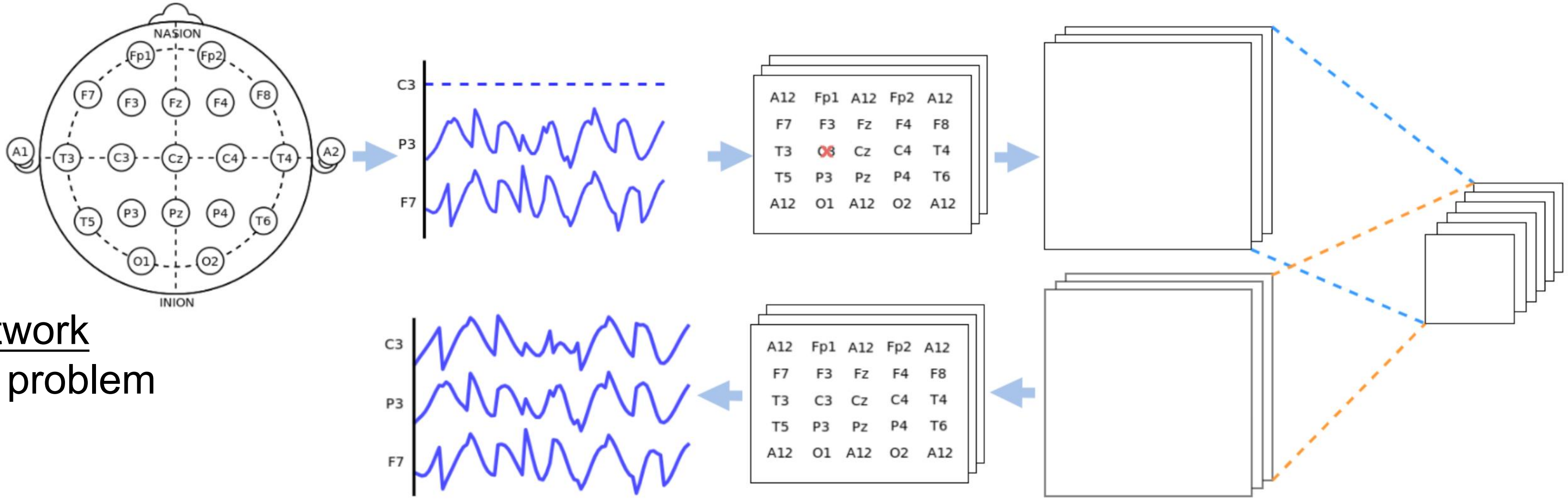
Oliveira et al (2017). A channel rejection method for attenuating motion-related artifacts in EEG recordings during walking
Tuyienge et al (2018). Automatic bad channel detection in intracranial electroencephalographic recordings using ensemble machine learning

EEG preprocessing – bad channel interpolation

Bad channel repair → Filtering → Artifact removal → Re-referencing

1. Bad channel detection
2. **Channel interpolation**

<u>Spherical splines</u>: The interpolant can be projected to the surface of the sphere to form a spherical triangulation of a polygonal domain of the spherical surface.



<u>Deep encoder-decoder network</u>
Similar to image inpainting problem in computer vision

Ref:
Perrin et al (1989). Spherical splines for scalp potential and current density mapping
Saba-Sadiya et al (2020) EEG Channel Interpolation Using Deep Encoder-decoder Networks

EEG preprocessing – filtering

Bad channel repair → **Filtering** → Artifact removal → Re-referencing

EEG preprocessing – filtering

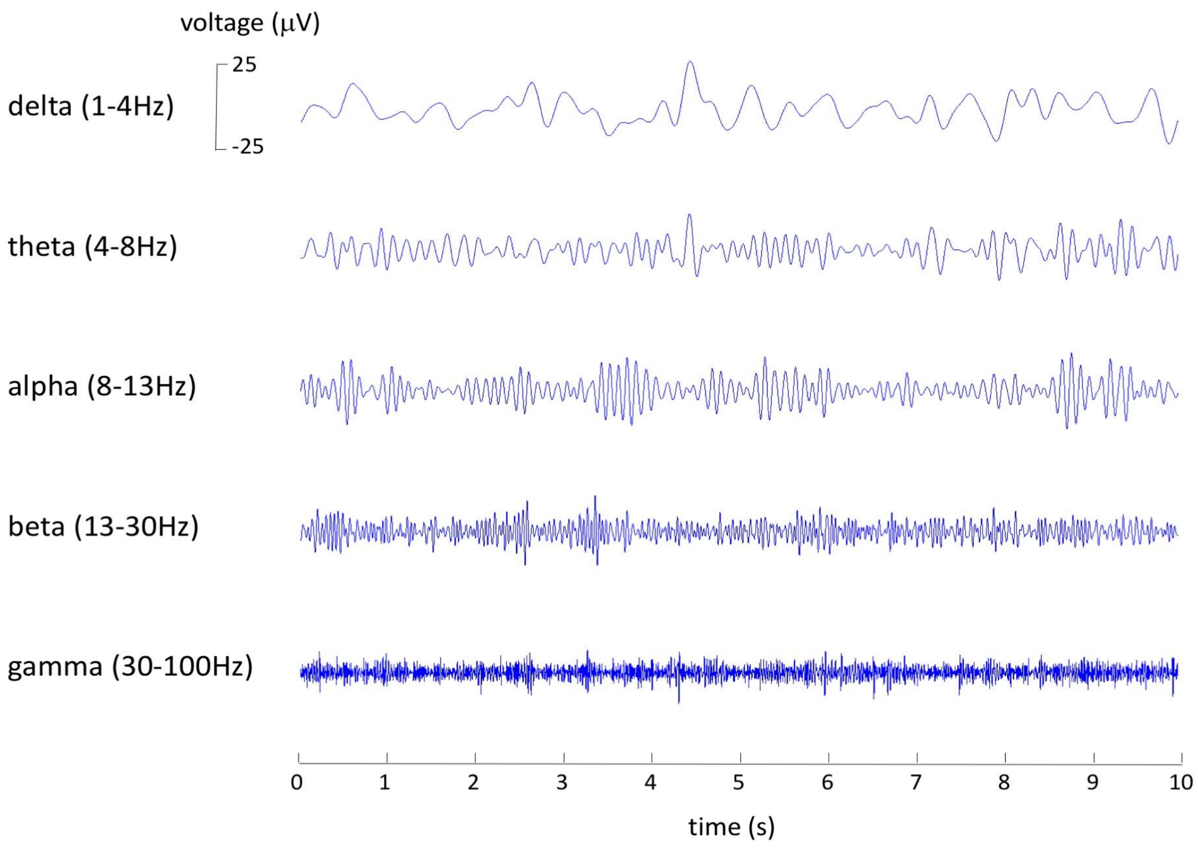
Digital filters are of two kinds:

- Finite impulse response (FIR) filters
- Infinite impulse response (IIR) filters

FIR filters have an impulse response of finite duration.
IIR filters do not become exactly zero after a certain point.
The most common IIR filters are: Butterworth filter, Chebyshev I or II type filter and Elliptic filter (Bhogeshwar et al 2014).

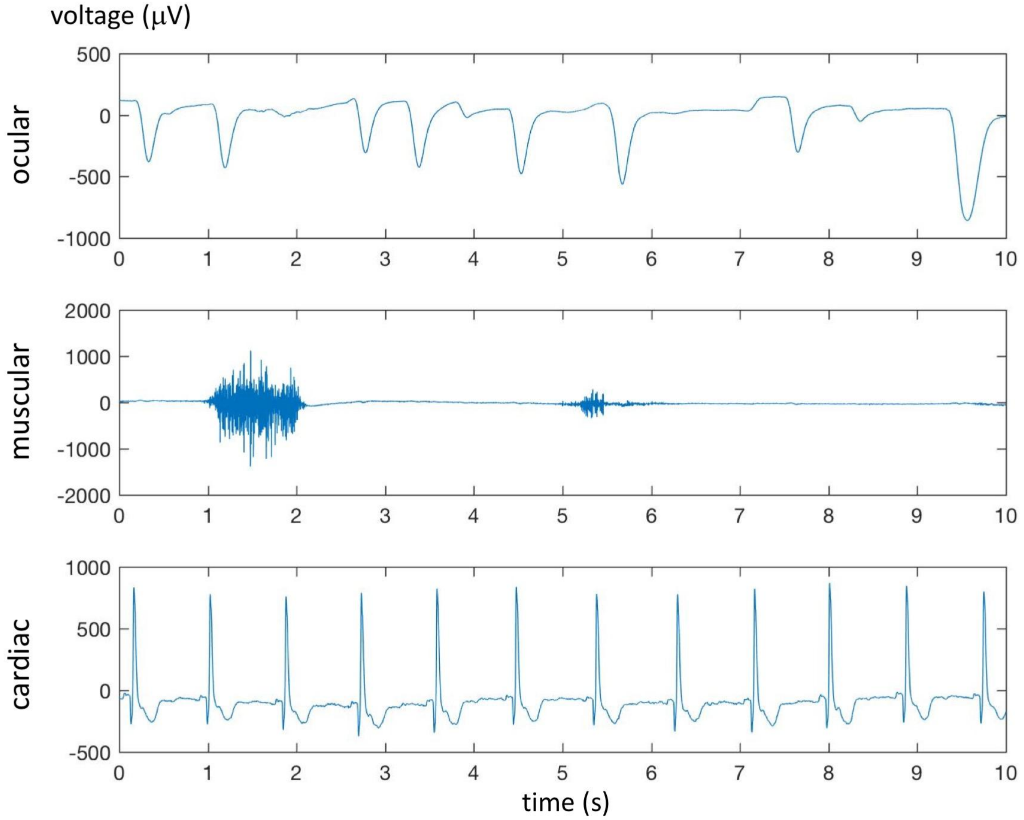
Filter parameters: cutoff frequency, filter order, roll-off, phase delay (zero-phase, linear-phase, non-linear phase).
Roll-off is the rate at which attenuation increases beyond the cut-off frequency.

EEG rhythms



The **phase** distortions and delays might be added to the signals, depending on the frequency response of the filter.

EEG preprocessing – artifact removal

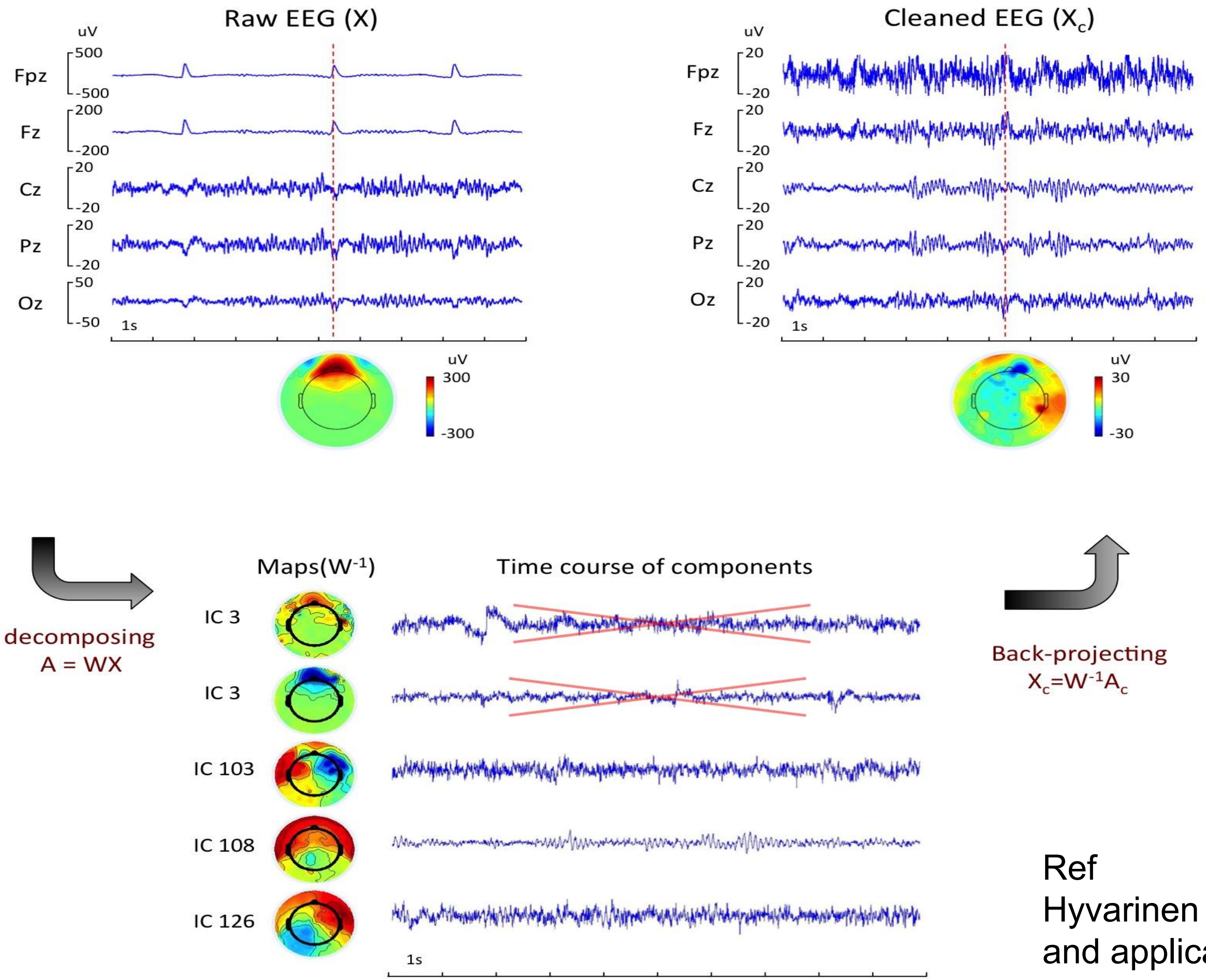


The waveform of three types of non-neuronal signals mixed in raw EEG recordings:

1. Ocular artifacts
2. Muscular artifacts
3. Cardiac artifacts

Their frequencies overlap with the frequency of EEG signals.
Therefore, filtering cannot completely remove these artefacts.

EEG preprocessing – ICA for EEG artifact removal



ICA is a computational method for separating a multivariate signal into *additive* subcomponents.

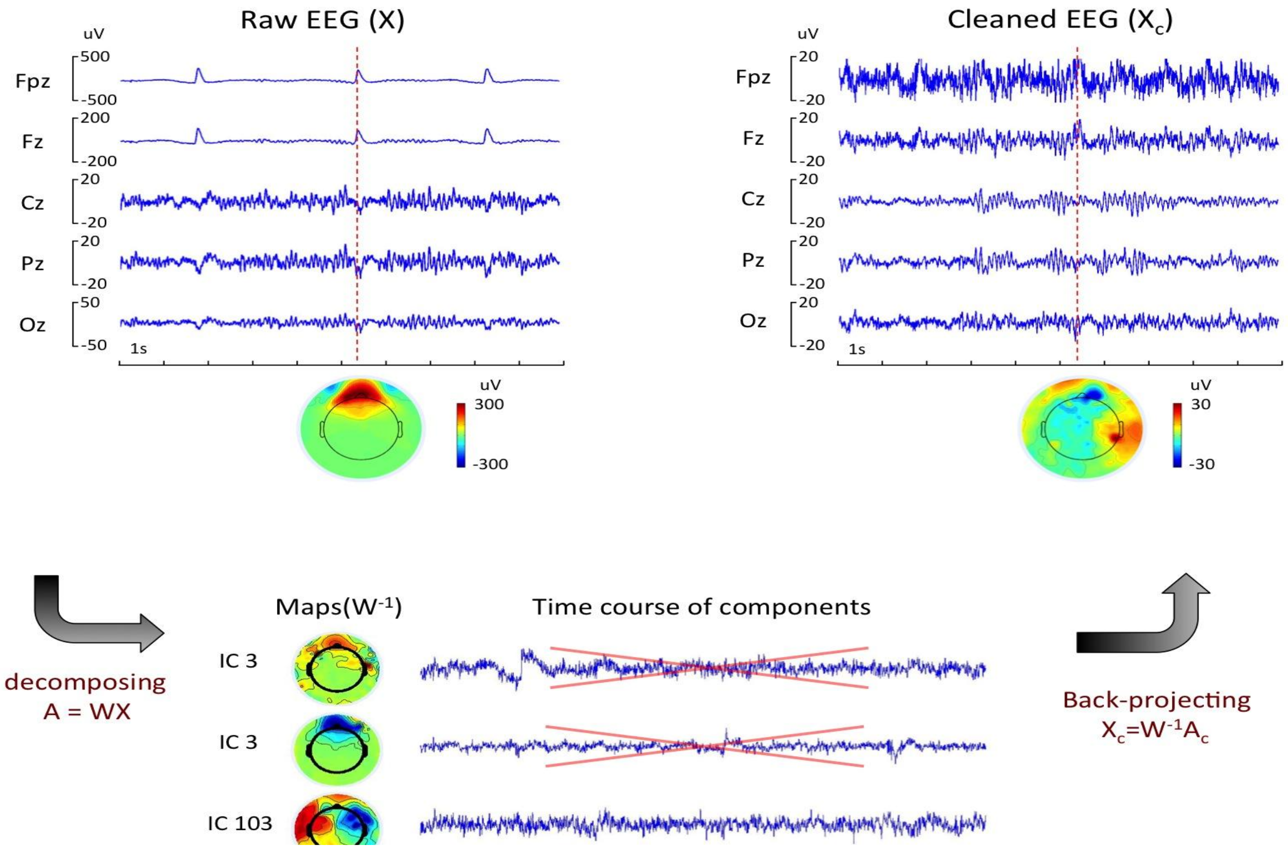
This can be done by **assuming** that the subcomponents are non-Gaussian signals and that they are statistically independent from each other (Hyvarinen & Oja 2000).

The core **mathematical concept** of ICA is to minimize the mutual information among the data projections or maximize their joint entropy.

Ref

Hyvarinen A, Oja E. (2000). Independent component analysis: algorithms and applications. *Neural networks*

EEG preprocessing – ICA for EEG artifact removal



ICA is a computational method for separating a multivariate signal into *additive* subcomponents.

This can be done by **assuming** that the subcomponents are non-Gaussian signals and that they are statistically independent from each other (Hyvarinen & Oja 2000).

The core **mathematical concept** of ICA is to minimize the mutual information among the data projections or maximize their joint entropy.

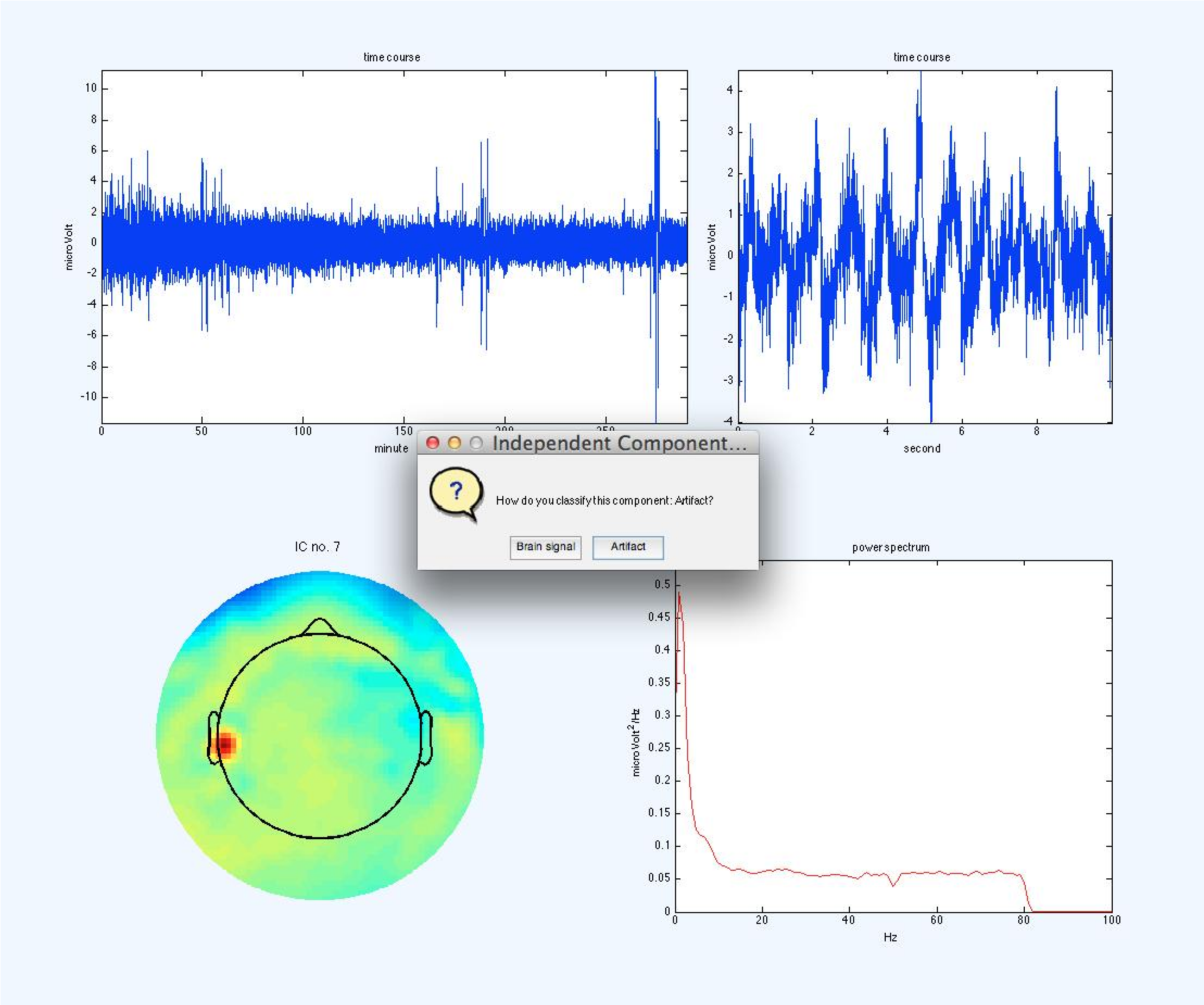
Question:
Which ICs shall be removed?

Question:
Which ICs shall be removed?

- Check IC time course
- Check IC spectrum
- Check IC topography （脑地形图）

- Machine learning to learn the latent features of artefactual ICs

Radüntz et al (2017). Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features

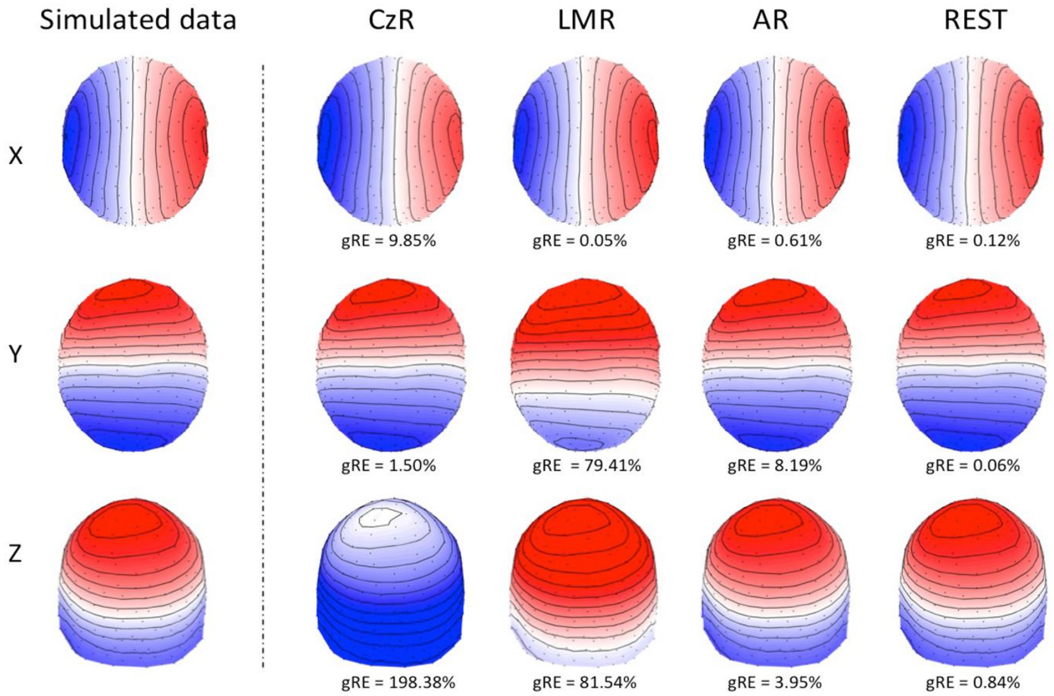


EEG preprocessing – re-referencing

Bad channel repair → Filtering → Artifact removal → **Re-referencing**

EEG is a relative measure that compares the recording site with another (reference) site.

- Cz reference
- Linked-mastoid reference (LMR)
- Average reference (AR)
- Reference standardization technique (REST)



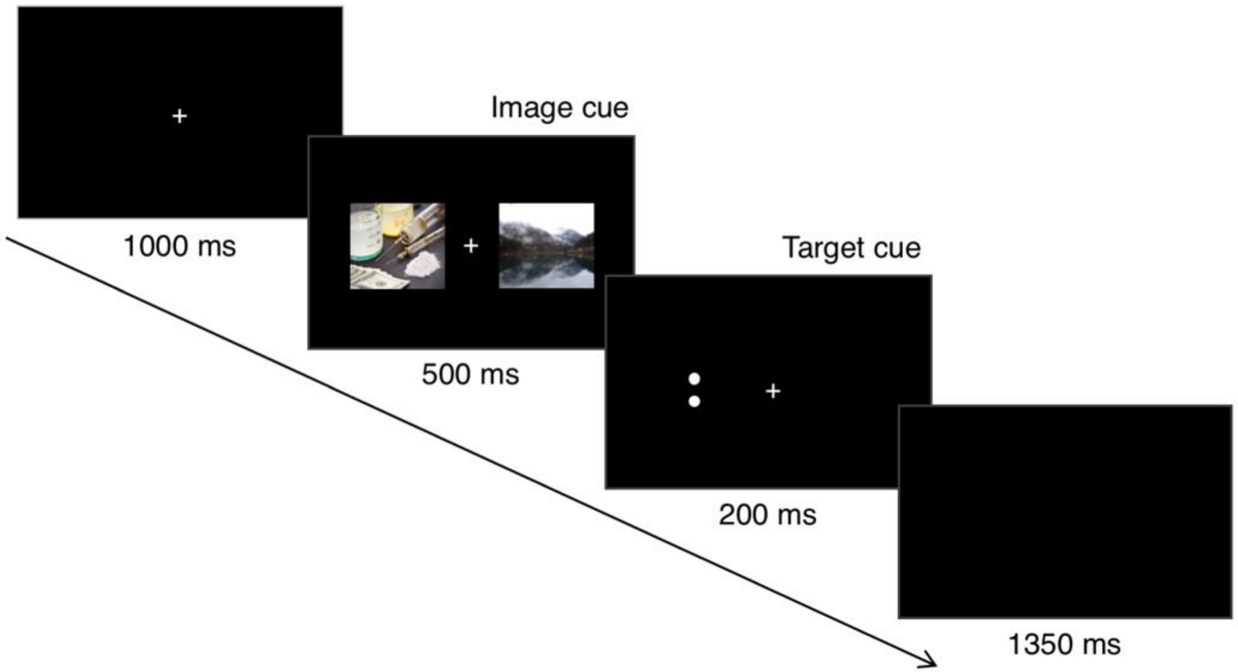
Liu et al (2015). Estimating a neutral reference for electroencephalographic recordings: the importance of using a high-density montage and a realistic head model

EEG sensor-level analysis

Resting EEG: Functional connectivity analysis; Resting-State Networks (RSNs)

Task EEG:

Event-related potential (ERP) analysis;



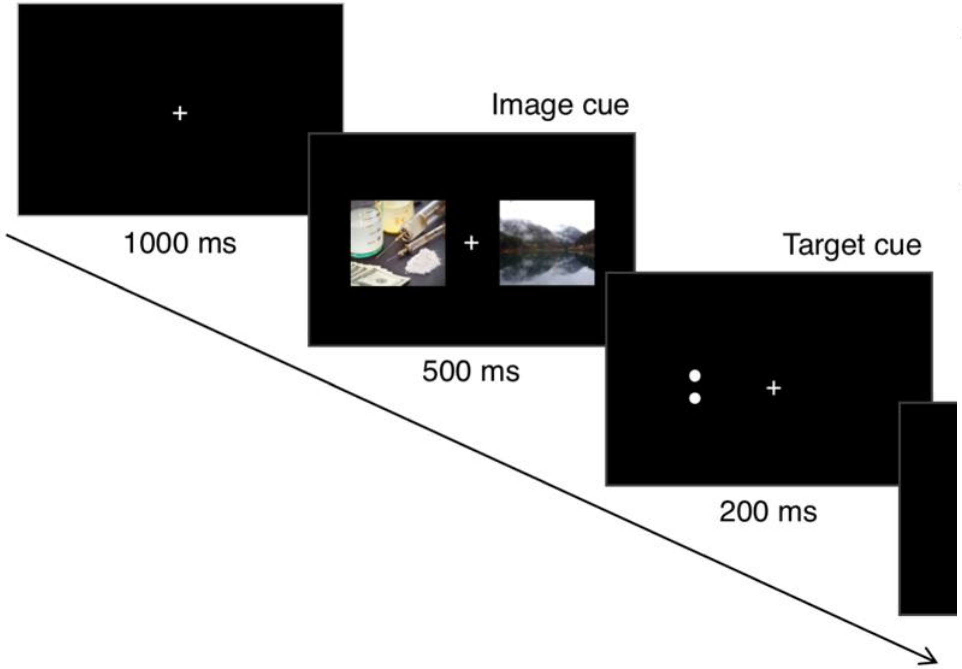
EEG sensor-level analysis

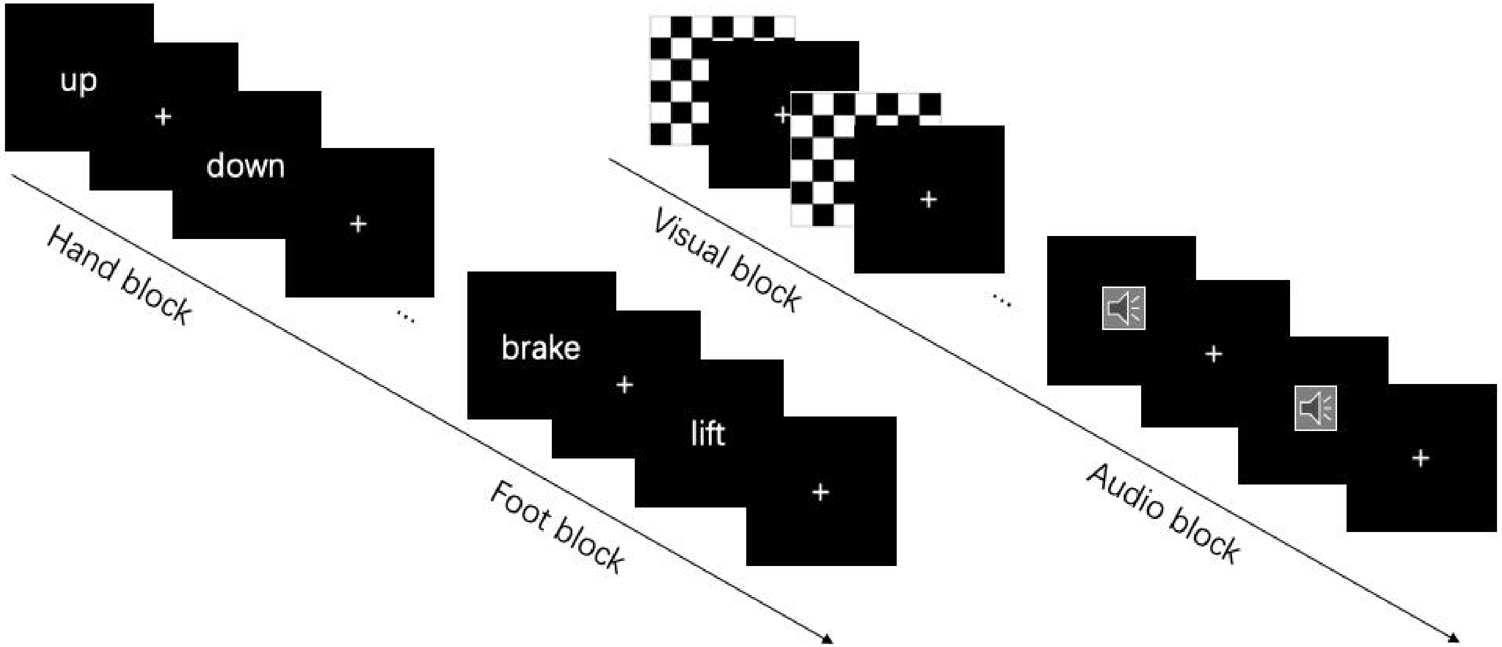
Resting EEG: Functional connectivity analysis; Resting-State Networks (RSNs)

Task EEG:

Event-related potential (ERP) analysis;

Event-Related Desynchronization (ERD) and Event-Related Synchronization (ERS)



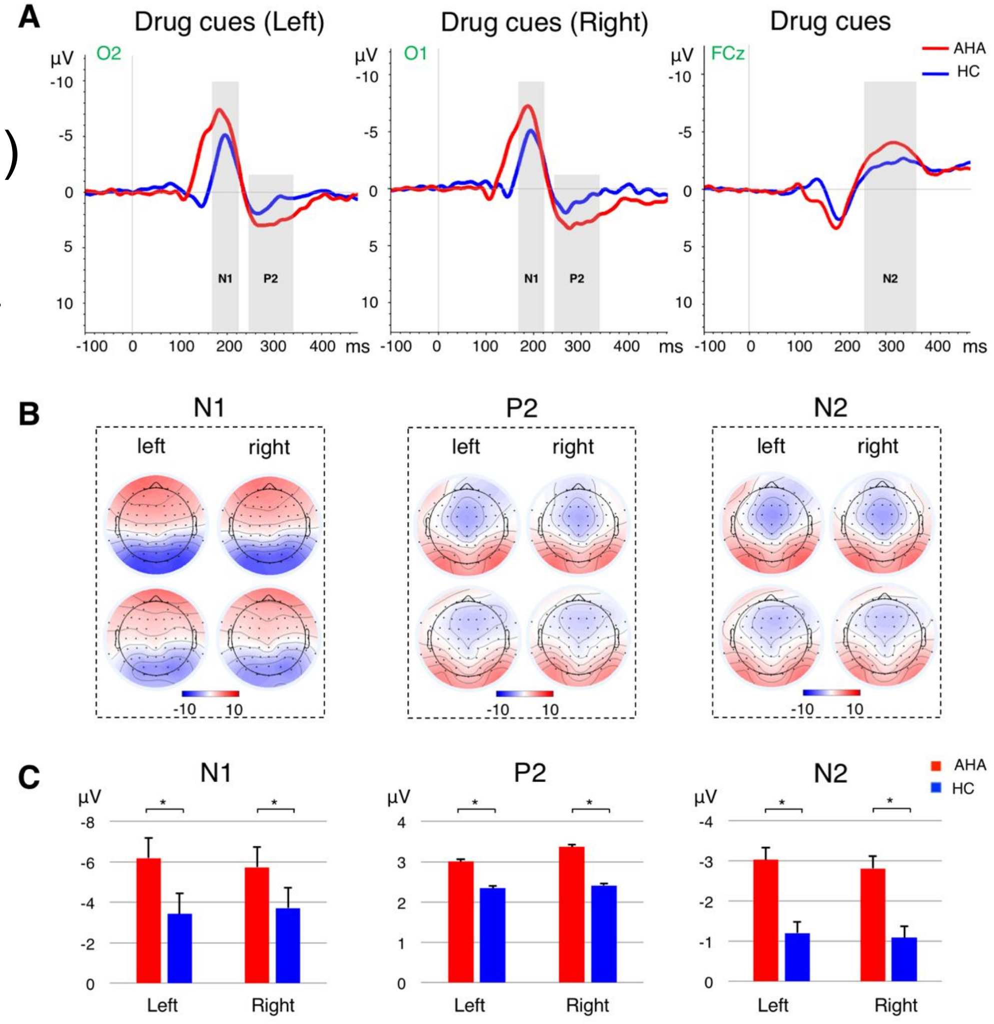


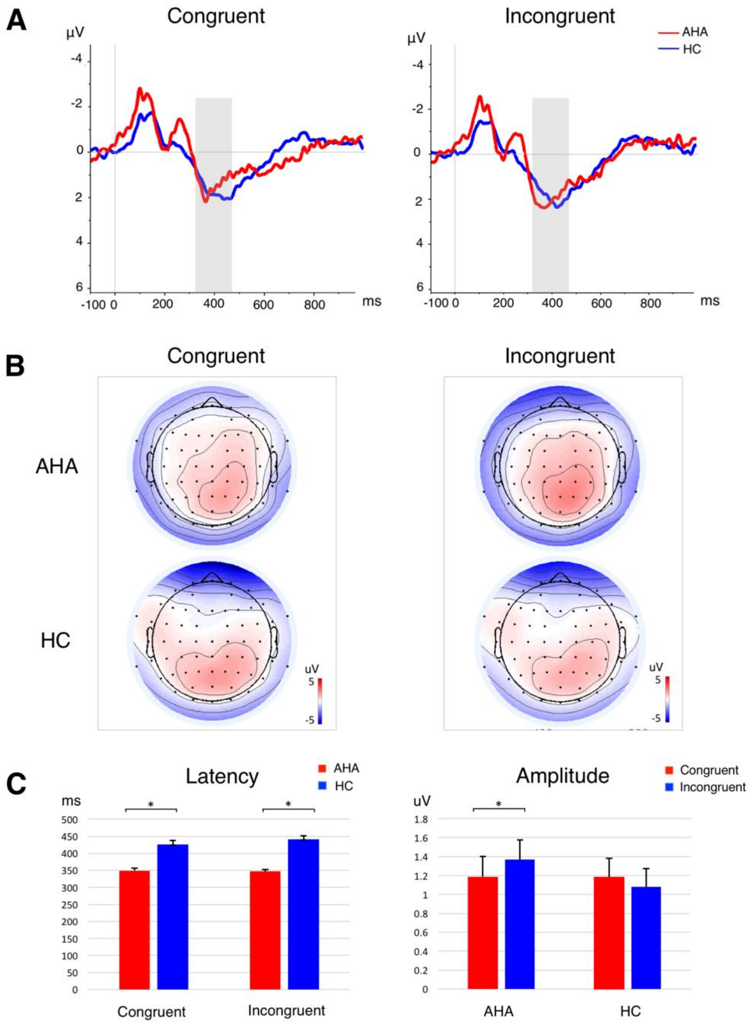
ERP analysis

Procedure to get ERP:
- Filter
- Epoch (-100ms to 1000ms)
- Bad trial removal
- Average across <u>conditions</u>

ERP components
- N1 (negative: ~100ms)
- P2
- N2
- P3 or P300

Different ERP components are associated with different cognitive processes.
- Latency
- Amplitude





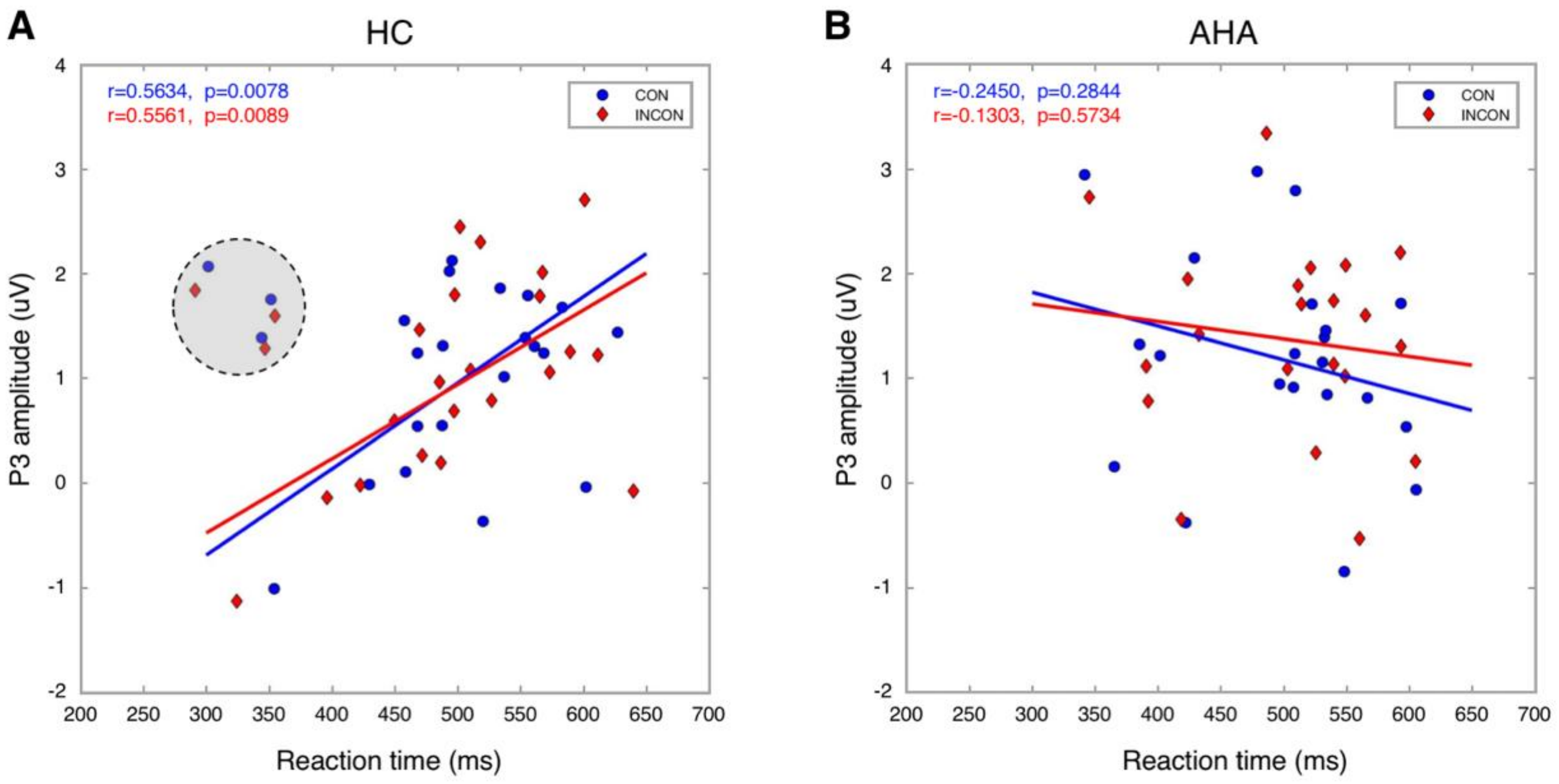
Zhao et al. (2018), Neural correlates of drug-related attentional bias in heroin dependence

ERP analysis

Procedure to get ERP:
- Filter
- Epoch (-100ms to 1000ms)
- Bad trial removal
- Average across conditions

ERP components
- N1 (negative: ~100ms)
- P2
- N2
- P3 or P300

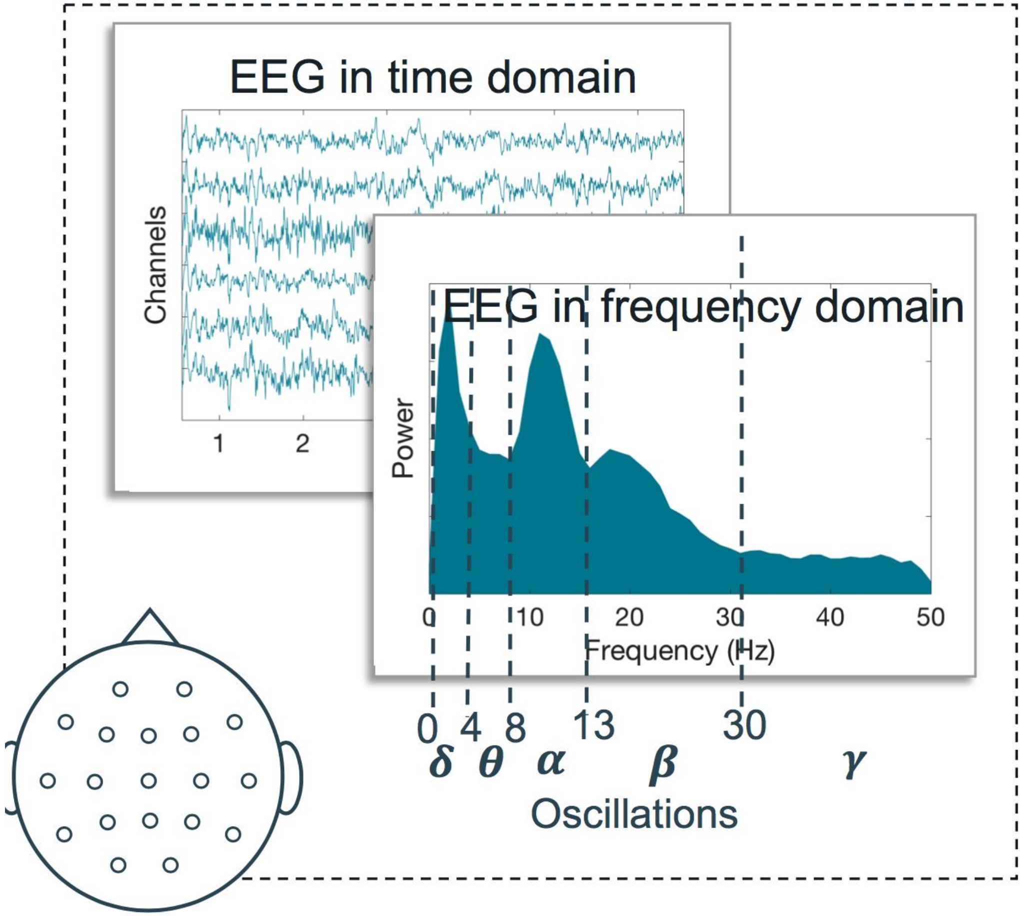
Different ERP components are associated with different cognitive processes.
- Latency
- Amplitude



Zhao et al. (2018), Neural correlates of drug-related attentional bias in heroin dependence

ERD/ERS analysis

Multi-dimensionality: • **space** • **time** • **frequency** • **feature**



Some movement-related features

- Event-related desynchronization/synchronization (ERS/ERD)[3]

$$ERD/ERS(f,t) = \frac{P(f,t) - P_B(f)}{P_B(f)} \times 100\%$$

- Cortico-kinematic coherence (CKC)[4]

$$CKC(f) = \frac{|PEEG_{KIN}(f)|^2}{P_{EEG}(f)P_{KIN}(f)}$$

- Cortico-muscular coherence (CMC)[4]

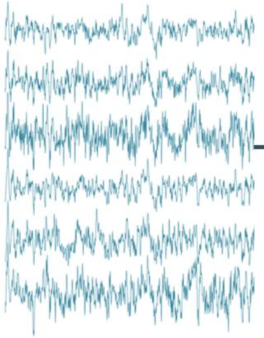
$$CMC(f) = \frac{|PEEG_{EMG}(f)|^2}{P_{EEG}(f)P_{EMG}(f)}$$

Other dimensions? Condition, trial ...

[3] Pfurtscheller, G., et al (1999)
[4] Bourguignon, M., et al (2019)

ERD/ERS analysis

An example of ERD/ERS

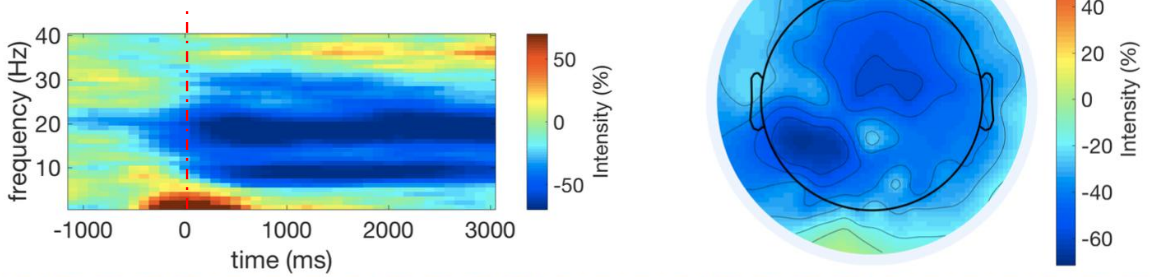


Time-frequency decomposition

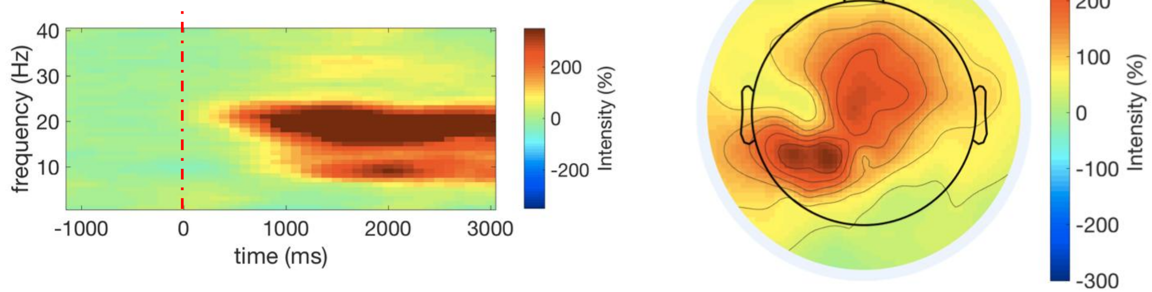
$p_i(f, t)$

$$ERD/ERSi(f,t) = \frac{P_i(f,t) - P_{Bi}(f)}{P_{Bi}(f)} \times 100\%$$

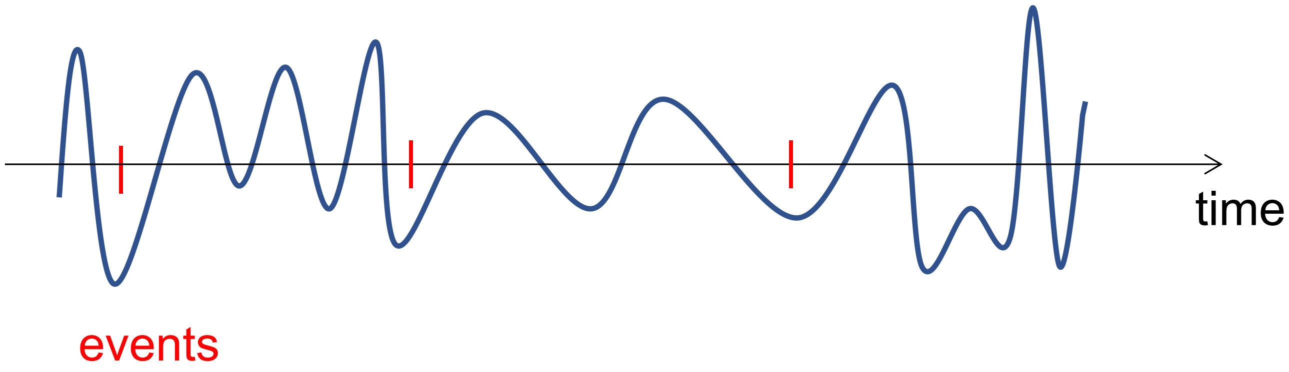
ERD after the start of movement



ERS after the stop of movement

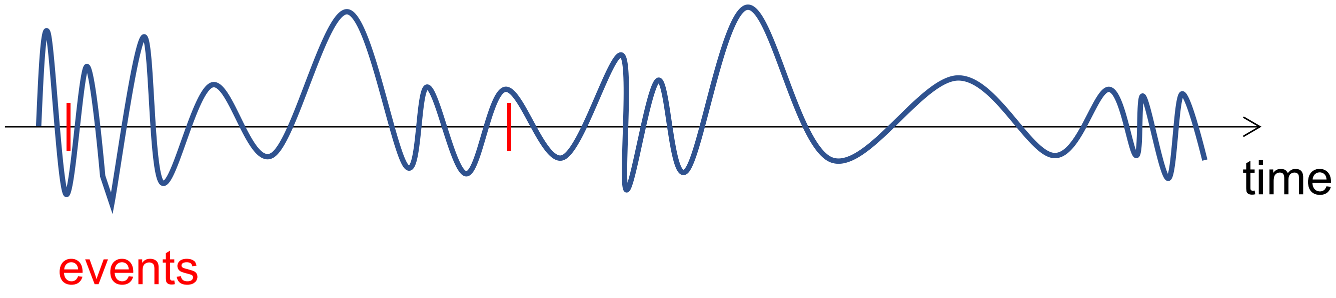


Differences between ERP and ERD/ERS



ERP

- Potential average
- Time-locked and phase-locked
- Evoked response



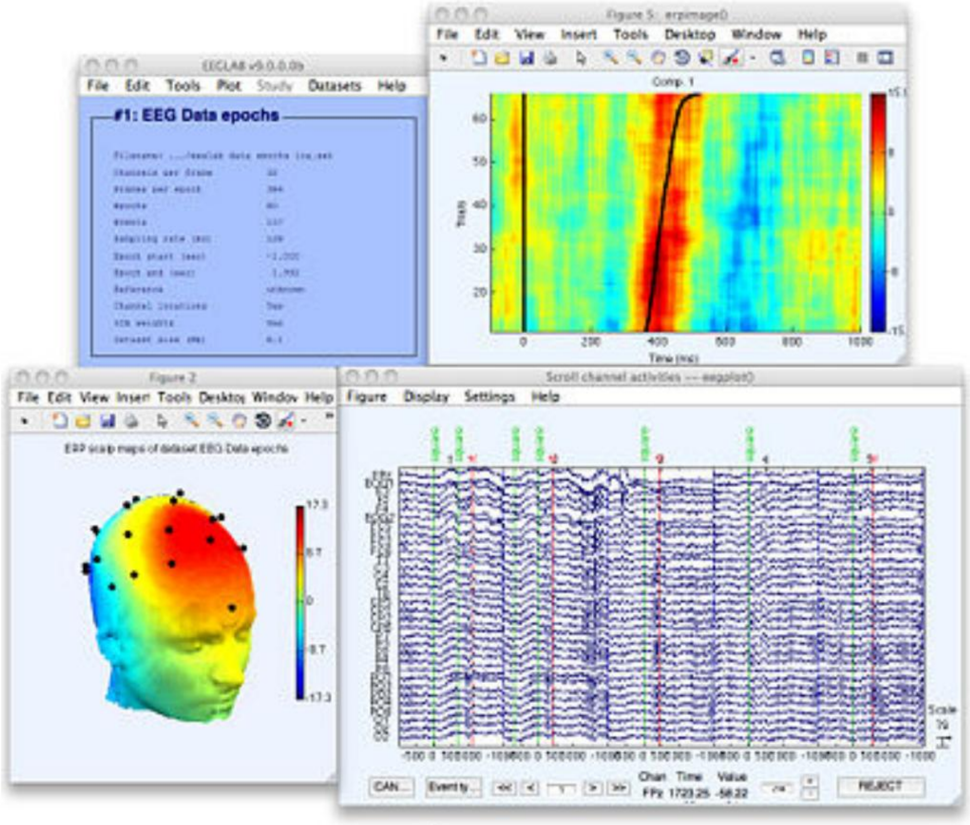
ERD/ERS

- Power average
- Time-locked, **not** phase-locked
- Induced response

Toolbox for EEG analysis

EEGlab (matlab)

Official webpage https://sccn.ucsd.edu/eeglab/index.php
Tutorial https://www.bilibili.com/video/BV1mJ411s7vH



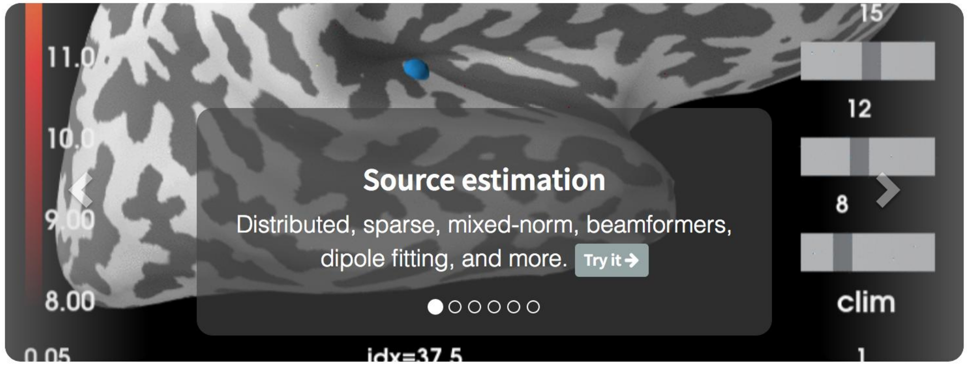
MNE-python

Official webpage https://mne.tools/stable/index.html
Tutorial https://www.bilibili.com/video/BV1YK411T7H8

MNE

MEG + EEG ANALYSIS & VISUALIZATION

Open-source Python package for exploring, visualizing, and analyzing human neurophysiological data: MEG, EEG, sEEG, ECoG, NIRS, and more.



Homework 2

Tips:

- Watch the tutorial video
  https://www.bilibili.com/video/BV1YK411T7H8
- Read the official website of MNE-python
  https://mne.tools/stable/index.html

EEG data processing with MNE-python

DDL: March 29, 2021，上课前

把code和结果图做成pdf发邮件给曲由之 12031145@mail.sustech.edu.cn

Requirements

1. Read the paper https://www.nature.com/articles/s41597-020-0535-2
2. Download the raw EEG data from [26] in the paper (https://doi.org/10.7910/DVN/RBN3XG)
   学生ID为奇数的同学下载sub-001； 学生ID为偶数的同学下载sub-002
3. Plot the **time course** of raw EEG signals with 10-second window (as Figure 4 in the paper).
4. Data preprocessing, artifacts removal
5. Plot the **time course** of preprocessed EEG signals
6. Plot the **time-frequency maps** of the subject (as Figure 6 in the paper)
7. Plot the **topographical distribution** of power of the subject (as Figure 7 in the paper)
8. **Comparison** of power (in dB) changes with time (in s) during hand, elbow motor imagery, and resting state for electrode C3, and electrode C4 (as Figure 8 in the paper)

Lecture 6 (part 2) – Probability Review

- Random variable:
  - Discrete (Bernoulli, Binomial, Categorical, Geometric, Poisson)
  - Continuous (Gaussian, Uniform, Exponential, Beta)
- Joint probability p(x, y)
- Marginalization / Law of Total Probability
- Conditional Probability p(x | y)
- Bayes' Rule
- Expectation & Conditional expectation
- Extension to N random variables

What is a random variable?

- A random variable x is a quantity that is uncertain

- It may be the result of experiment (e.g., flipping a coin) or real world measurement (e.g., measuring temperature)

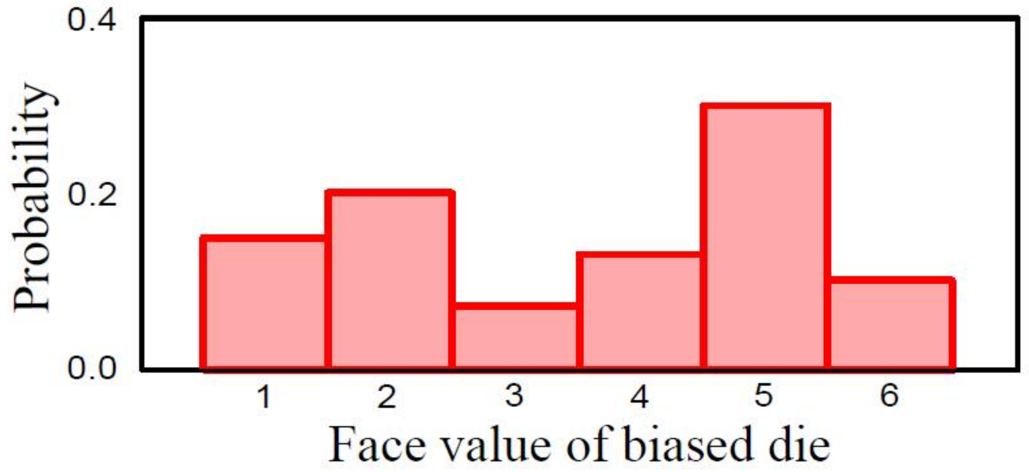
- If observe *x* multiple times, we get different values
- Some values occur more than others; this information captured by *probability distribution* p(x)

- If x is **discrete**, then "p" is "probability mass function" (or pmf).
- If x is **continuous**, then "p" is "probability density function" (or pdf).

Adapted from S. Prince

Discrete Random Variable

- Take on discrete values



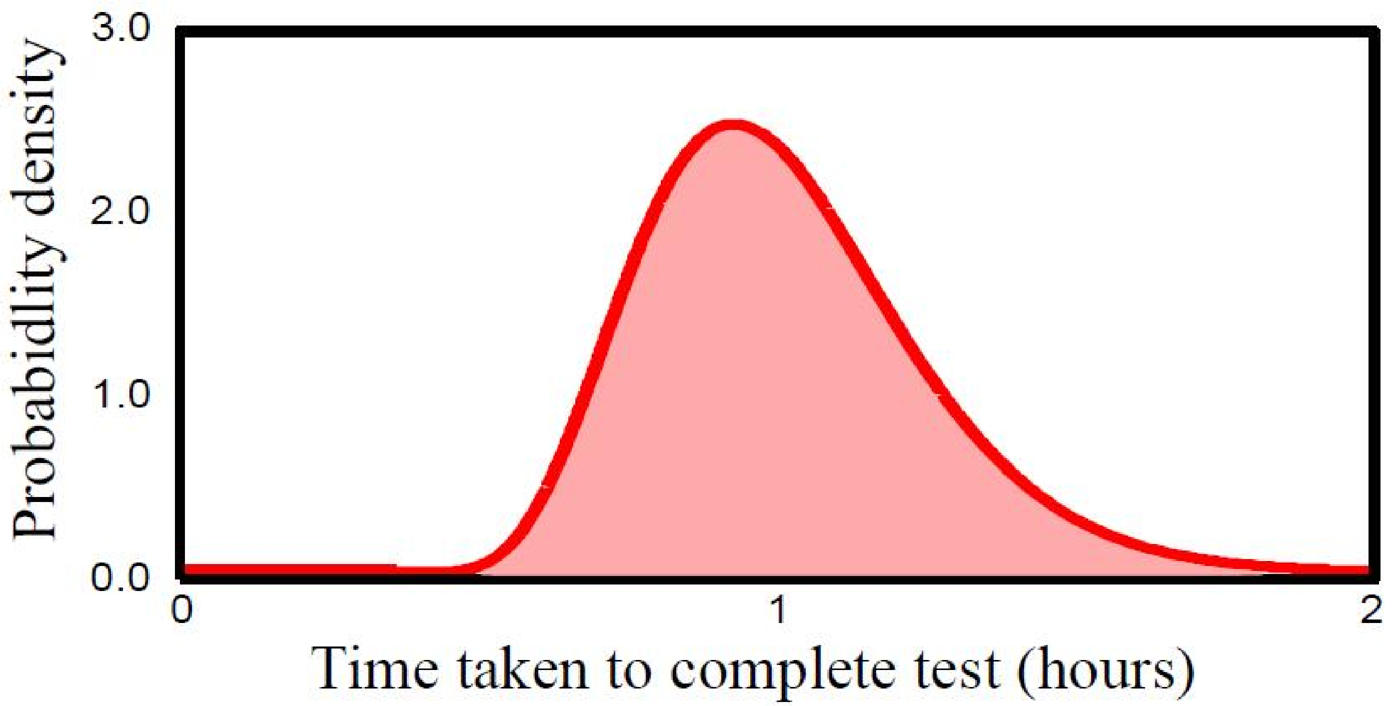
$$\sum_x p(x) = 1$$

1. $0 \leq P(A) \leq 1,$
2. $P(\Omega) = 1,$
3. $P(A \cup B) = P(A) + P(B) \quad \text{if} \quad A \cap B = \emptyset.$

Adapted from S. Prince

Continuous Random Variable

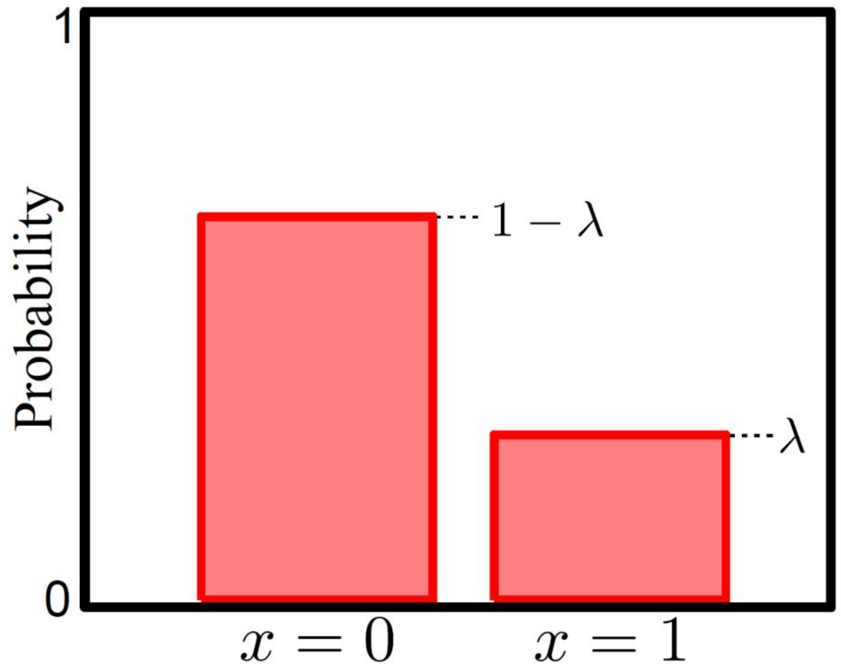
- Take on continuous values or uncountably infinite number of values



Famous Discrete Random Variables

- Bernoulli: http://en.wikipedia.org/wiki/Bernoulli_distribution
- Categorical: http://en.wikipedia.org/wiki/Categorical_distribution
- Binomial: http://en.wikipedia.org/wiki/Binomial_distribution
- Geometric: http://en.wikipedia.org/wiki/Geometric_distribution
- Poisson: http://en.wikipedia.org/wiki/Poisson_distribution
- …

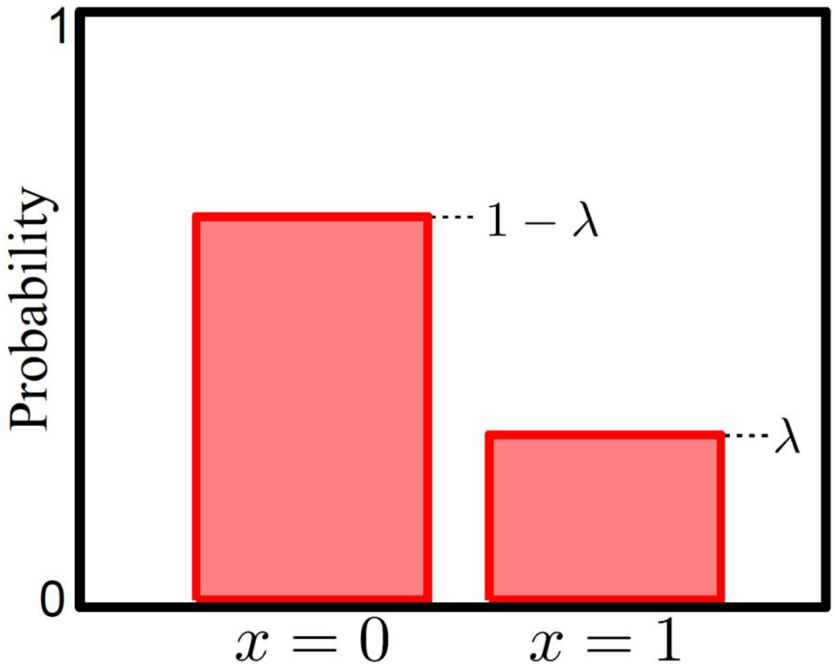
Bernoulli Distribution



Bernoulli distribution describes situation where only two possible outcomes x = 0 or x = 1 (e.g. failure/success)

Takes a single parameter: $\lambda \in [0, 1]$

Bernoulli Distribution

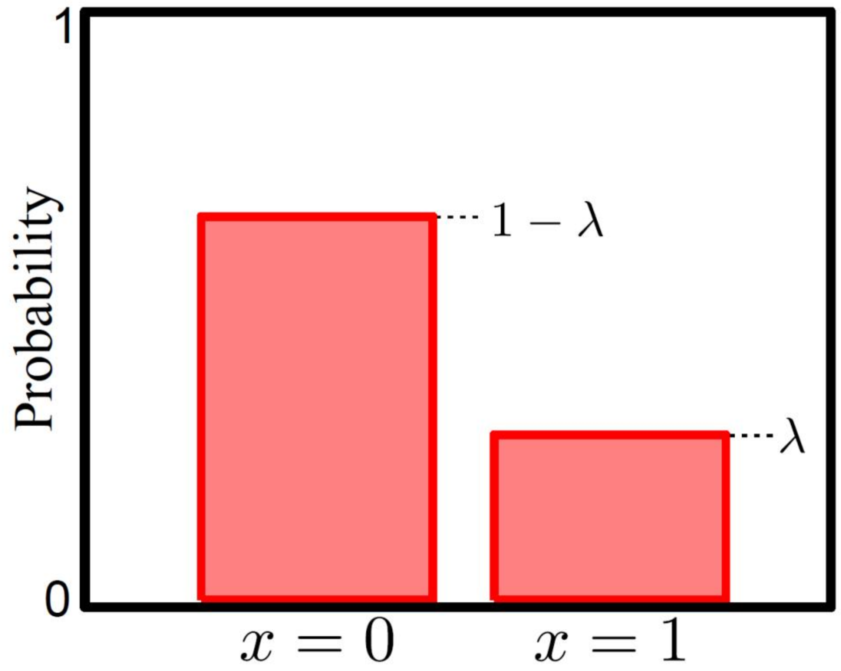


$$
\begin{aligned}
Pr(x=0) &= 1-\lambda \\
Pr(x=1) &= \lambda.
\end{aligned}
$$

Bernoulli distribution describes situation where only two possible outcomes x = 0 or x = 1 (e.g. failure/success)

Takes a single parameter: $\lambda \in [0, 1]$

Bernoulli Distribution



$$
\begin{aligned}
Pr(x=0) &= 1-\lambda \\
Pr(x=1) &= \lambda.
\end{aligned}
$$

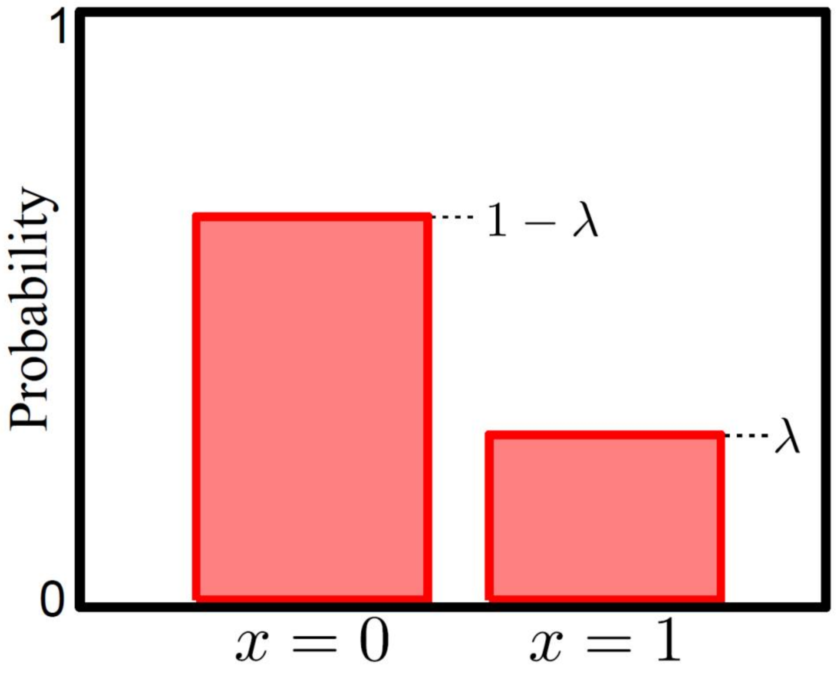
or

$$Pr(x) = \lambda^{x}(1-\lambda)^{1-x}$$

Bernoulli distribution describes situation where only two possible outcomes x = 0 or x = 1 (e.g. failure/success)

Takes a single parameter: $\lambda \in [0,1]$

Bernoulli Distribution



$$
\begin{aligned}
Pr(x=0) &= 1-\lambda \\
Pr(x=1) &= \lambda.
\end{aligned}
$$

or

$$Pr(x) = \lambda^x (1-\lambda)^{1-x}$$

For short we write:

$$\boxed{p(x) = \mathrm{Ber}(x|\lambda)}$$

Bernoulli distribution describes situation where only two possible outcomes x = 0 or x = 1 (e.g. failure/success)

Takes a single parameter: $\lambda \in [0,1]$

Binomial Distribution

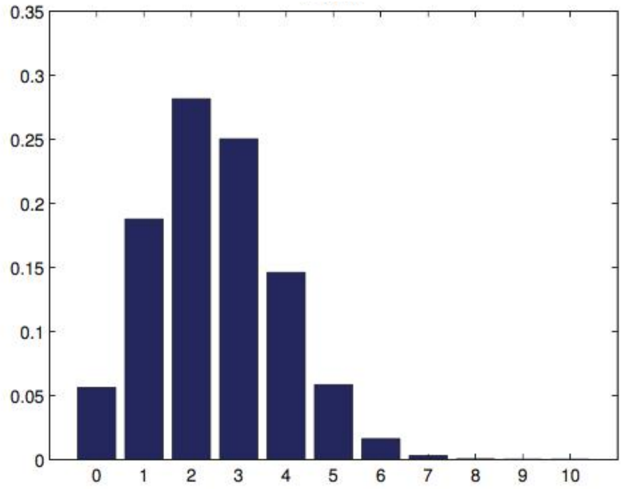
Suppose we toss a coin n times. Let $X \in \{0, \ldots, n\}$ be the number of heads. If the probability of heads is θ, then we say X has a **binomial** distribution, written as $X \sim \text{Bin}(n, \theta)$. The pmf is given by

$$\text{Bin}(k|n, \theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

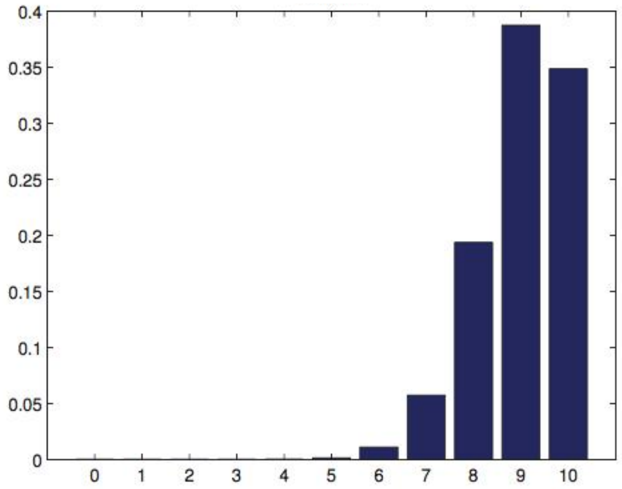
where

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$

n=10, theta = 0.25



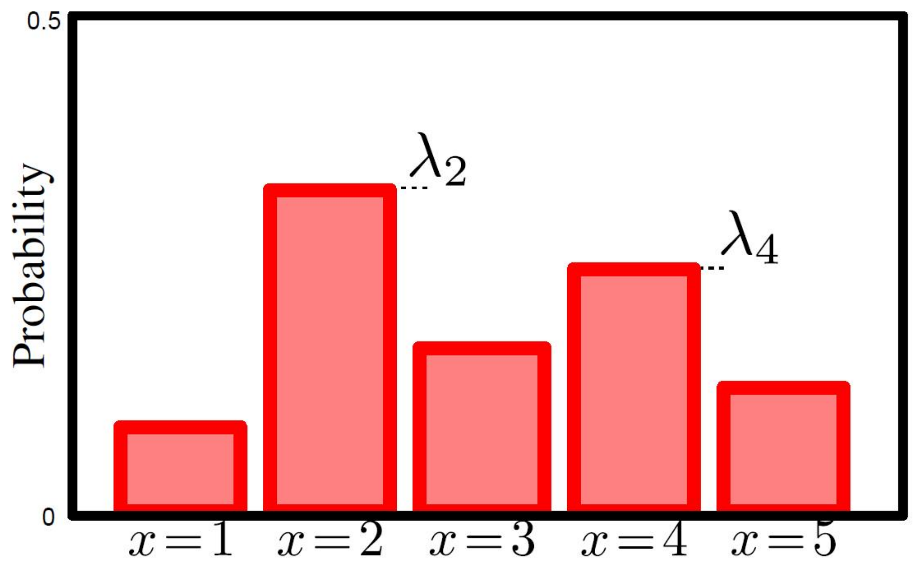
n=10, theta = 0.80



$$\text{mean} = \theta, \quad \text{var} = n\theta(1-\theta)$$

```
# Generate data
data1 <- rbinom(n=1000, size=10, prob=0.25)
```

Categorical Distribution

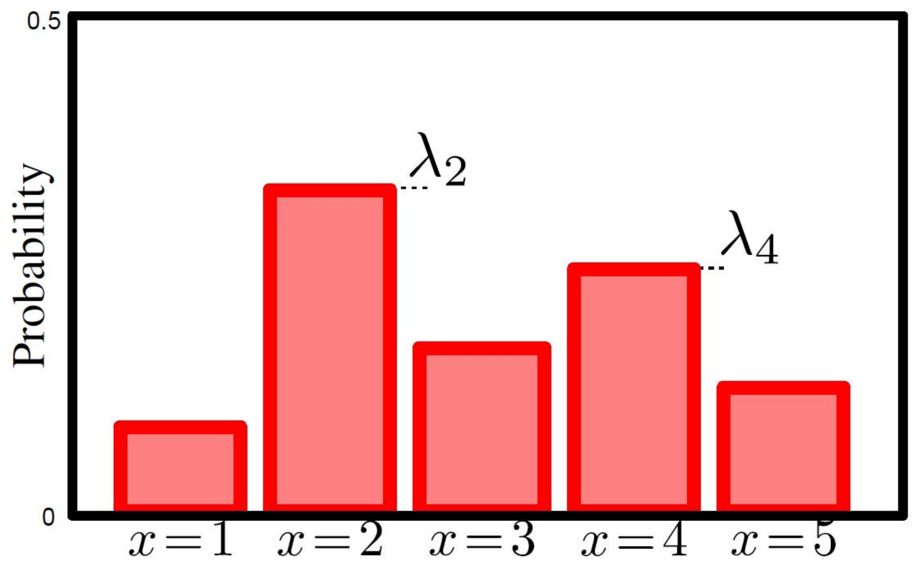


Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0,1]$ where $\sum_{k=1}^{K} p(X=k) = \sum_{k=1}^{K} \lambda_k = 1$

$$\lambda = \{\lambda_1, \cdots, \lambda_K\}$$

Categorical Distribution



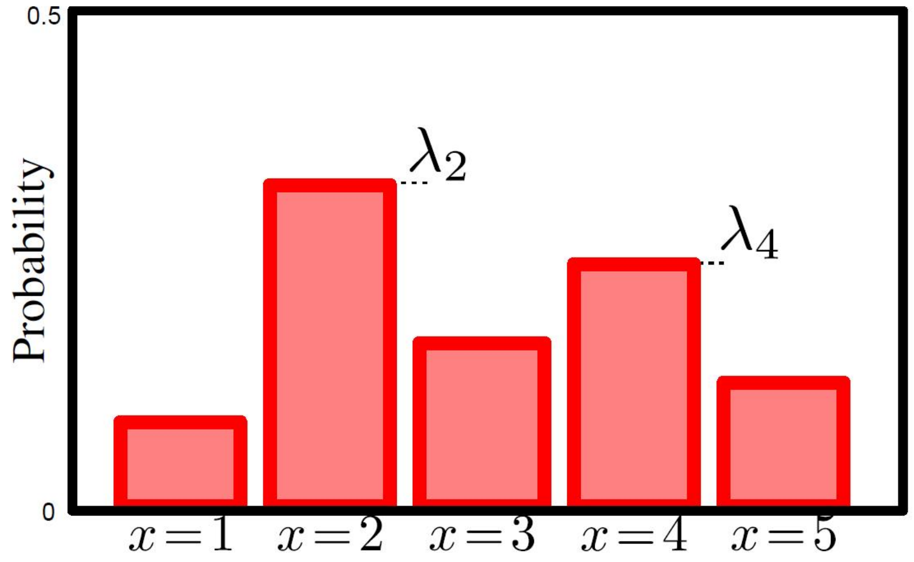
$$Pr\left(x = k\right) = \lambda_k$$

Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0, 1]$ where $\sum_{k=1}^{K} p(X = k) = \sum_{k=1}^{K} \lambda_k = 1$

$$\lambda = \{\lambda_1, \cdots, \lambda_K\}$$

Categorical Distribution



$$Pr(x = k) = \lambda_k$$

or can think of data as vector with all elements zero except kth e.g. $\mathbf{e}_4$ = [0,0,0,1,0]

$$Pr(x = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{\mathbf{e}_{kj}} = \lambda_k$$

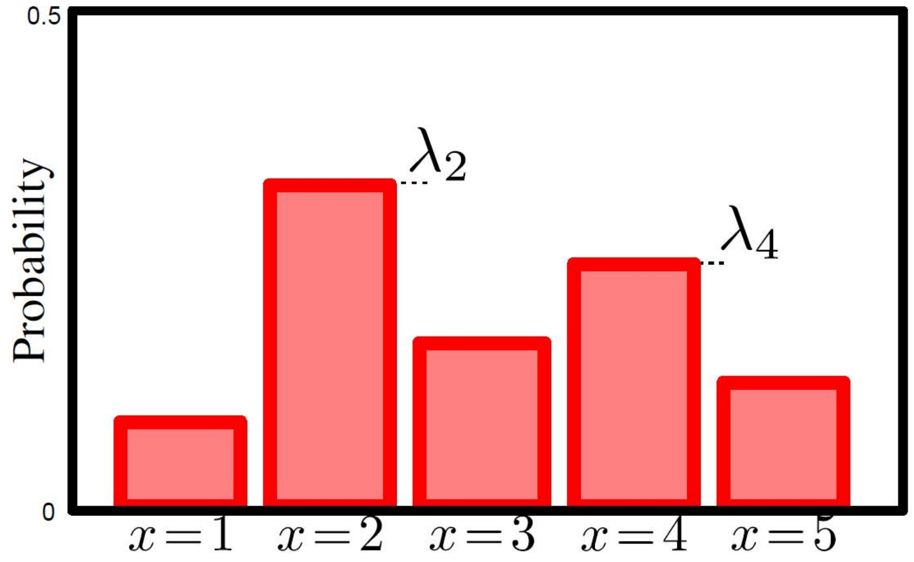
where $\mathbf{e}_{kj}$ is the j-th element of $\mathbf{e}_k$

Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0, 1]$ where $\sum_{k=1}^{K} p(X = k) = \sum_{k=1}^{K} \lambda_k = 1$

$$\lambda = \{\lambda_1, \cdots, \lambda_K\}$$

Categorical Distribution



$$Pr\,(x=k)=\lambda_k$$

or can think of data as vector with all elements zero except k[th] e.g. $\mathbf{e}_4$ = [0,0,0,1,0]

$$Pr\,(x=\mathbf{e}_k)=\prod_{j=1}^{K}\lambda_j^{\mathbf{e}_{kj}}=\lambda_k$$

where $\mathbf{e}_{kj}$ is the j-th element of $\mathbf{e}_k$

For short we write: $\boxed{p(x)=\text{Cat}(x|\lambda)}$

Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k\in[0,1]$ where $\sum_{k=1}^{K}p(X=k)=\sum_{k=1}^{K}\lambda_k=1$

$$\lambda=\{\lambda_1,\cdots,\lambda_K\}$$

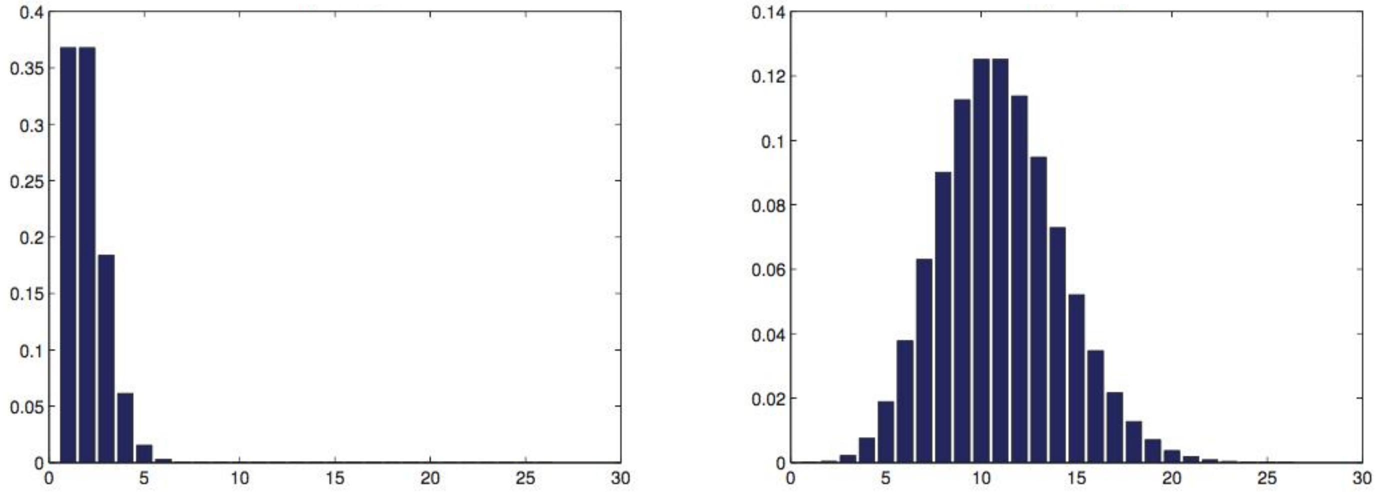
Poisson Distribution

We say that $X \in \{0, 1, 2, \ldots\}$ has a **Poisson** distribution with parameter $\lambda > 0$, written $X \sim \mathrm{Poi}(\lambda)$, if its pmf is

$$\mathrm{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

normalization constant
to ensure the distribution sums to 1

Lambda = 1 Lambda = 10



(a) (b)

The Poisson distribution is often used as a model for **counts of rare events** like radioactive decay and traffic accidents.

In neuroscience, we used Poisson distribution to model **the counts of neuron spikes**.

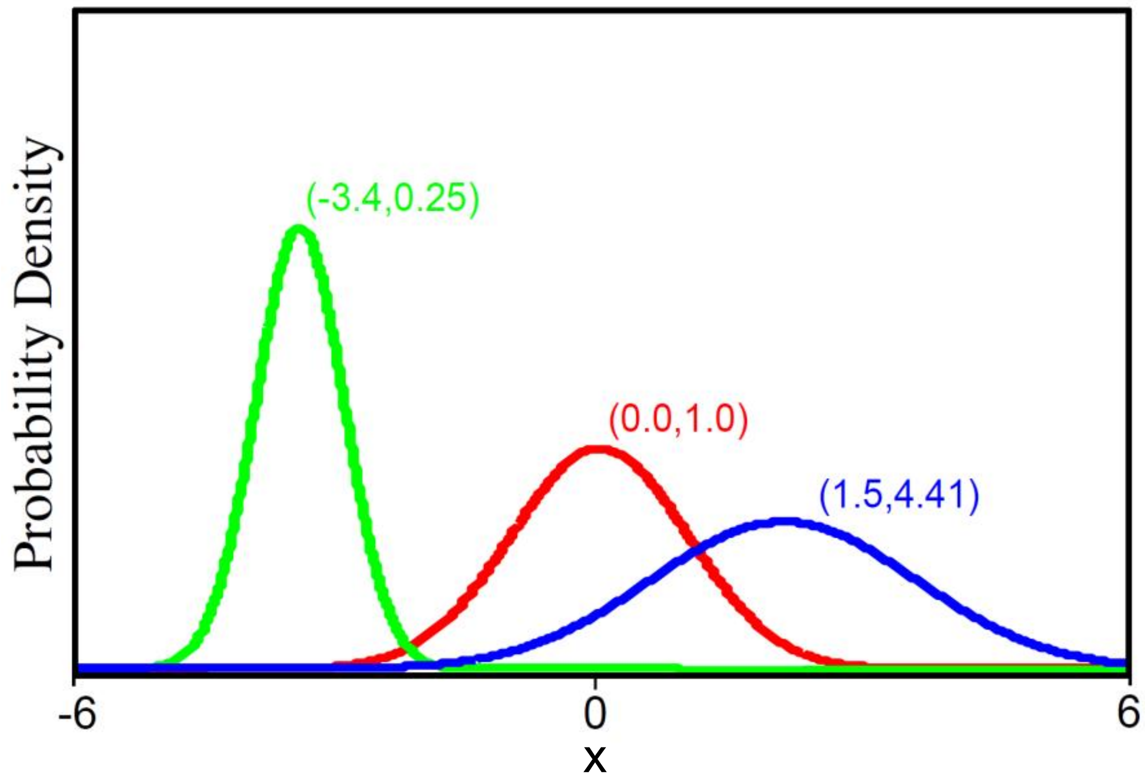
MLaPP, Kevin Murphy

Famous Continuous Random Variables

- Gaussian: http://en.wikipedia.org/wiki/Normal_distribution
- Uniform: http://en.wikipedia.org/wiki/Uniform_distribution_(continuous)
- Exponential: http://en.wikipedia.org/wiki/Exponential_distribution
- Beta: http://en.wikipedia.org/wiki/Beta_distribution
- …

Gaussian / Normal Distribution

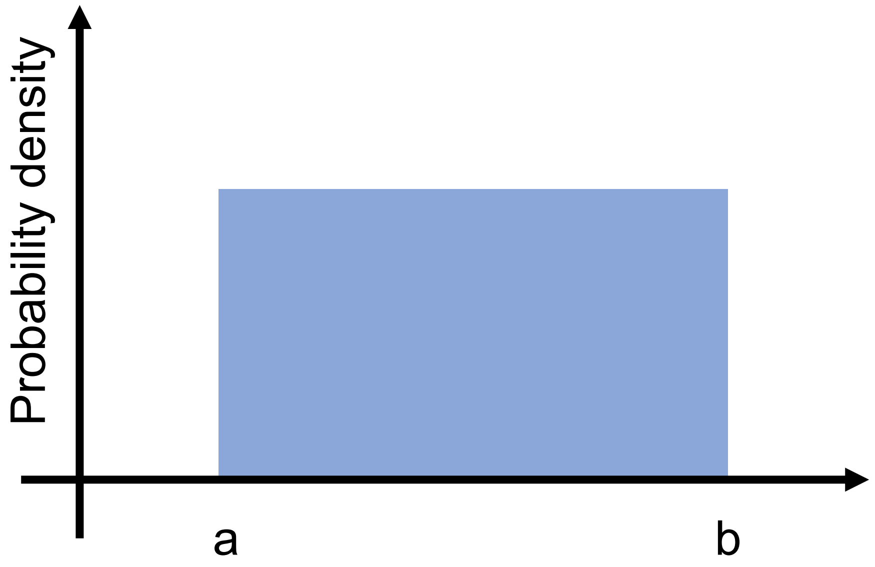
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



2 parameters:
mean μ and variance $\sigma^2 > 0$

Adapted from S. Prince

Uniform Distribution



$$\mathrm{Unif}(x|a,b) = \frac{1}{b-a}\mathbb{I}(a \le x \le b)$$

Mean = (b+a)/2

var = (b-a)^2/12

We usually use it as an uninformative prior.

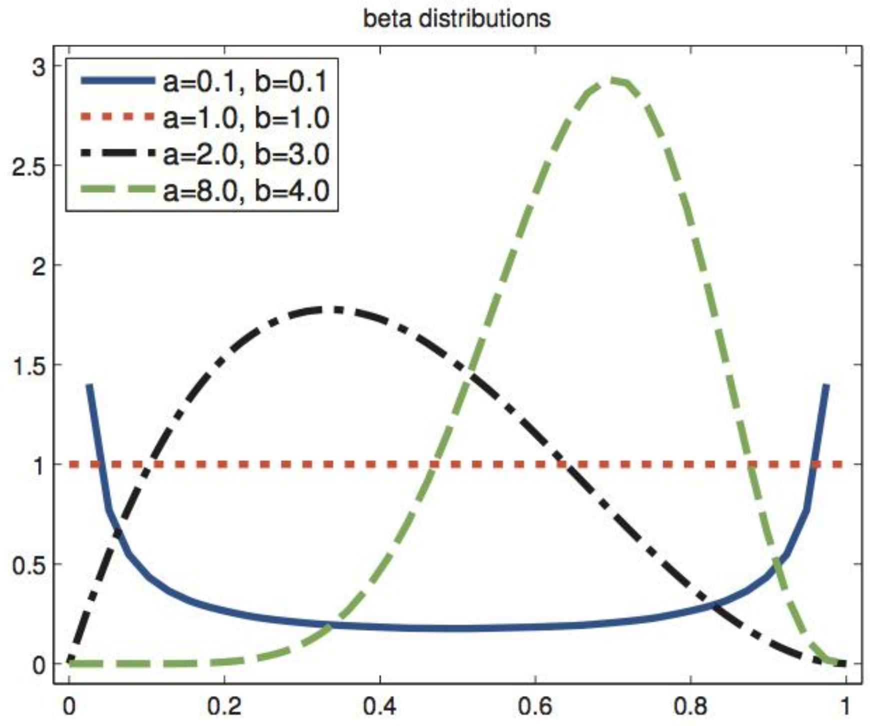
$$\mathrm{var}[X] \triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx$$

MLaPP, Kevin Murphy

Beta Distribution

The **beta distribution** has support over the interval $[0, 1]$ and is defined as follows:

$$\text{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \qquad B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$



$$\boxed{\begin{aligned} \text{mean} &= \frac{a}{a+b} \\ \text{mode} &= \frac{a-1}{a+b-2} \\ \text{var} &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}}$$

Beta distribution is important, for we usually use it as a conjugate prior of Binomial (or Bernoulli) process.

MLaPP, Kevin Murphy

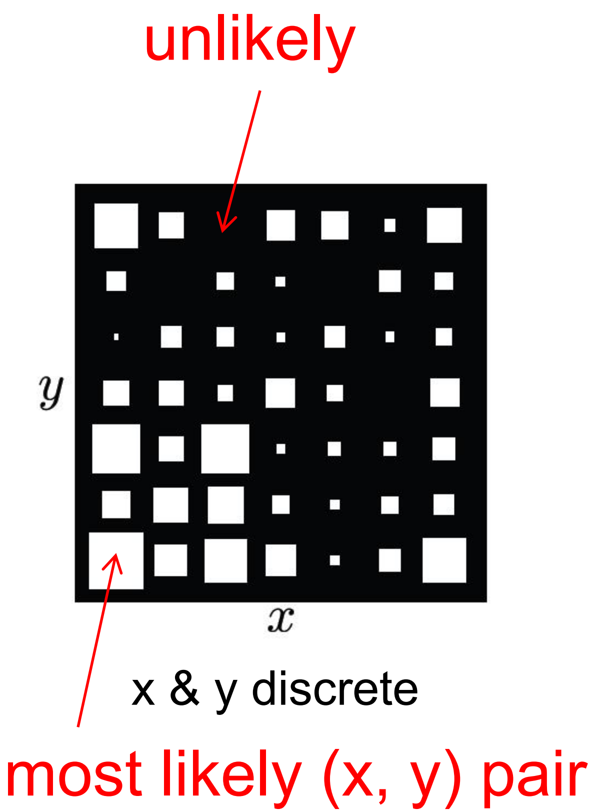
Joint Probability

- If we observe two random variables x & y multiple times, then some combinations of outcomes more likely than others

- This information captured by *joint probability distribution*

- Written as p(x, y), which is read as “joint probability distribution of x and y”

Joint Probability p(x, y)



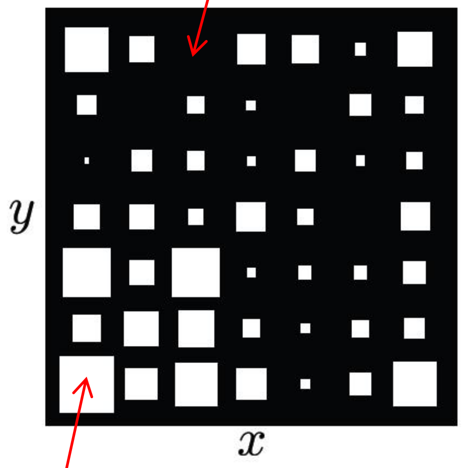
x & y discrete

$$\sum_y \sum_x p(x, y) = 1$$

Adapted from S. Prince

Joint Probability p(x, y)

unlikely

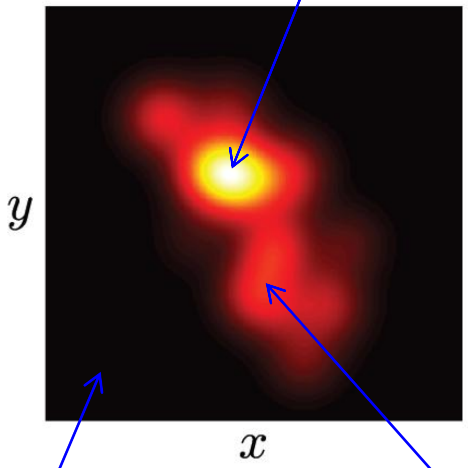


x & y discrete

most likely (x, y) pair

$$\sum_y \sum_x p(x, y) = 1$$

most likely (x, y) pair



x & y continuous

unlikely

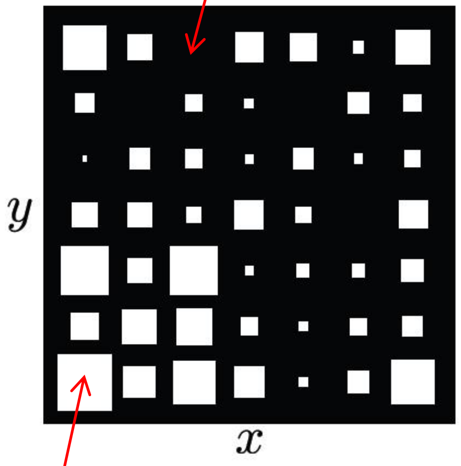
somewhat likely

$$\int_y \int_x p(x, y) dx dy = 1$$

Adapted from S. Prince

Joint Probability p(x, y)

unlikely

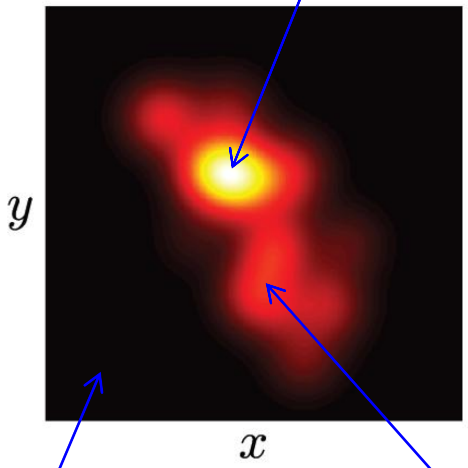


x & y discrete

most likely (x, y) pair

$$\sum_y \sum_x p(x, y) = 1$$

most likely (x, y) pair

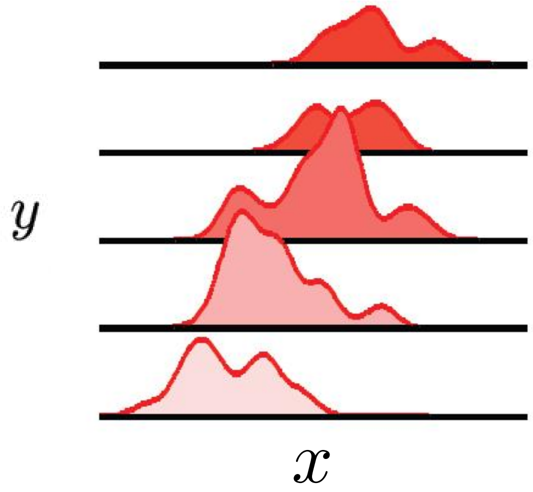


x & y continuous

unlikely

somewhat likely

$$\int_y \int_x p(x, y) dx dy = 1$$



x is continuous,
y is discrete

$$\sum_y \int_x p(x, y) dx = 1$$

Adapted from S. Prince

Marginalization / Law of Total Probability

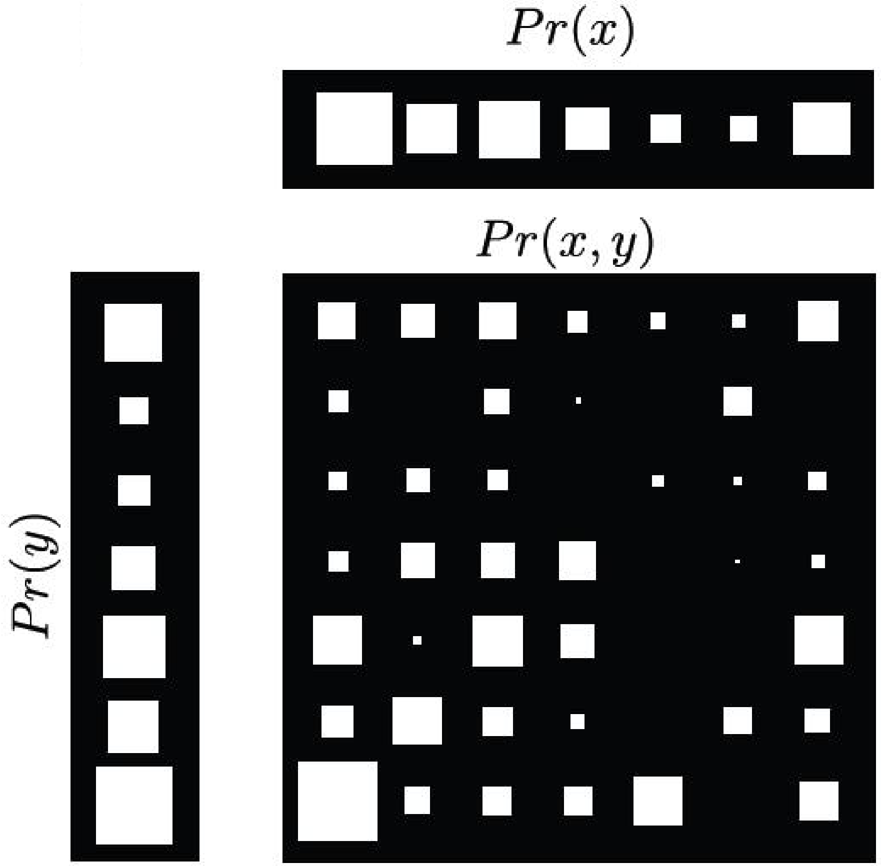
Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variable(s).

This is called marginalization.

$$p(x) = \sum_y p(x, y)$$

$$p(y) = \sum_x p(x, y)$$



Adapted from S. Prince

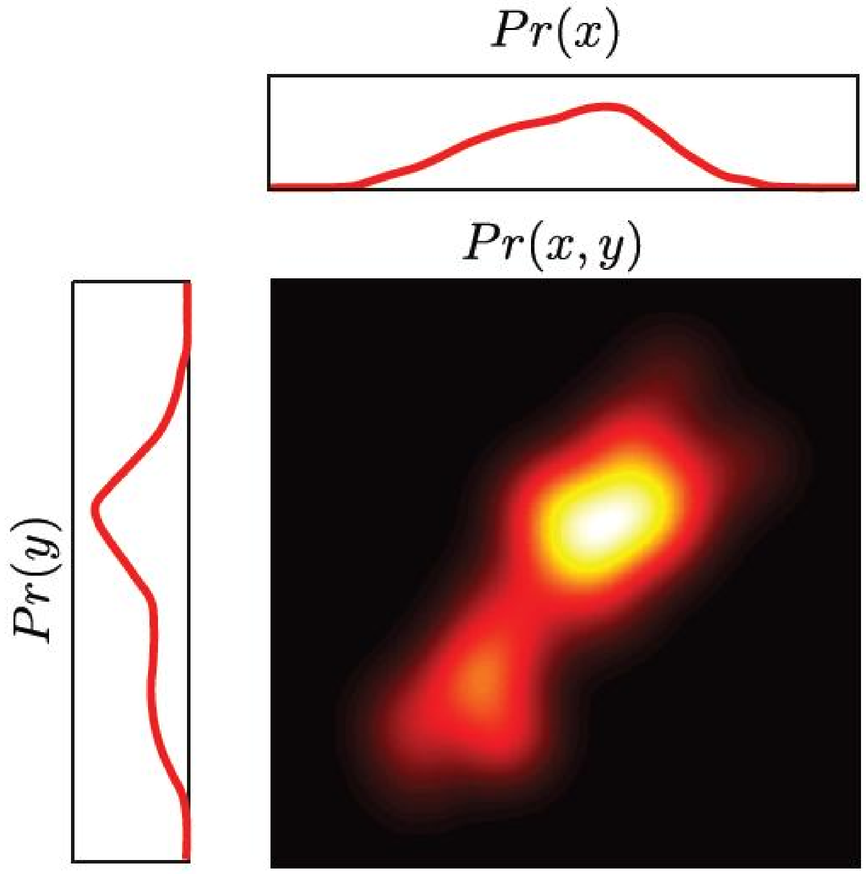
Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variable(s).

This is called marginalization.

$$p(x) = \int_y p(x, y)dy$$

$$p(y) = \int_x p(x, y)dx$$



Adapted from S. Prince

Marginalization Example

p(x, y)

		x	
		0	2.5
	-3	0	1/2
y	-1	1/8	1/4
	2	1/8	0

Marginalization Example

$$p(y) = \sum_x p(x, y)$$

p(x, y)

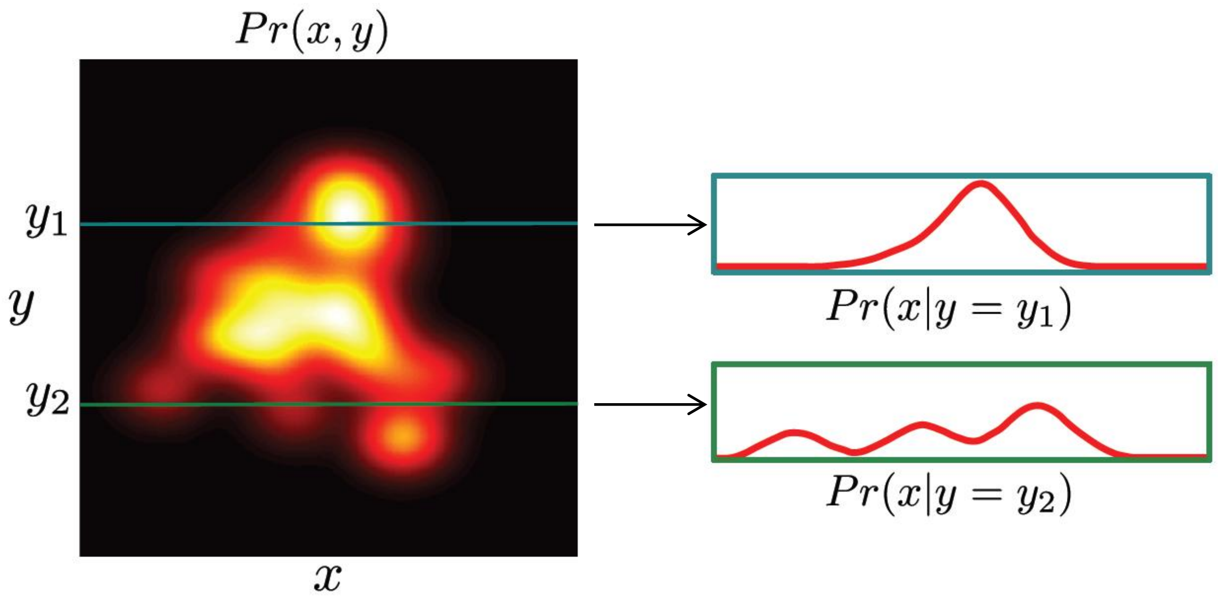
		x		
		0	2.5	p(y)?
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)?	1/4	3/4	

$$p(x) = \sum_y p(x, y)$$

Conditional Probability

Conditional Probability

- Suppose we observe y to be y_1, then $p(x \mid y = y_1)$ is how likely x will take on various values given this observation
- $p(x \mid y = y_1)$ read as "conditional probability of X given Y is equal to y_1"



$$\int_x p(x|y=y_1)dx = 1$$

$$\int_x p(x|y=y_2)dx = 1$$

Adapted from S. Prince

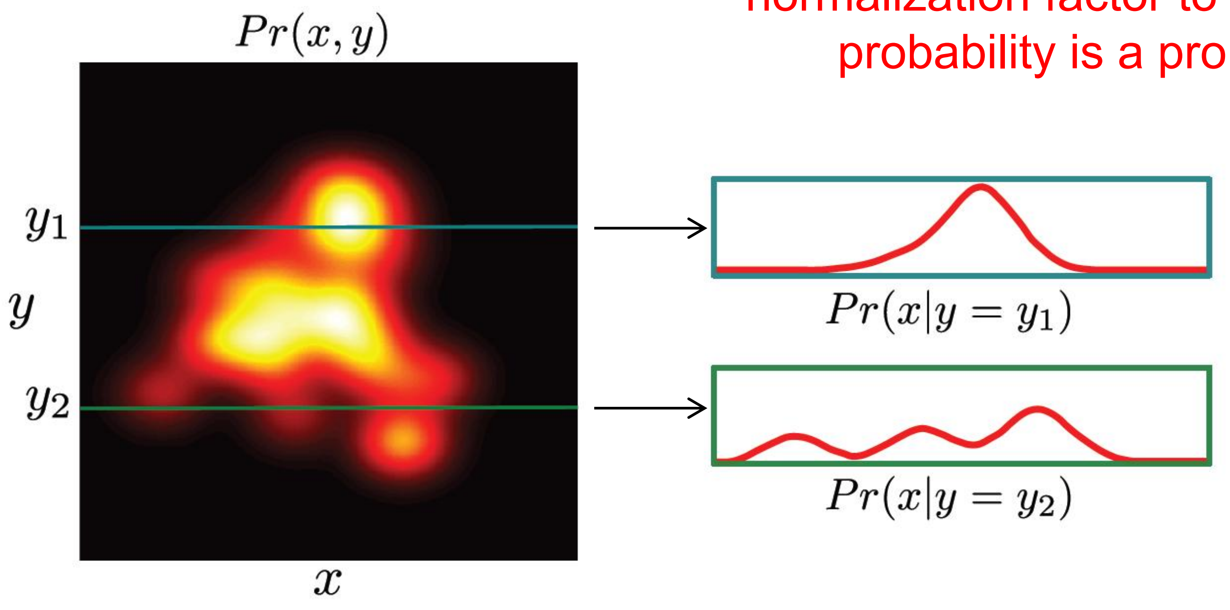
Conditional Probability

- Conditional probability can be computed from joint probability.

slice of joint distribution

$$p(x|y = y^*) = \frac{p(x, y = y^*)}{p(y = y^*)}$$

normalization factor to ensure conditional probability is a proper distribution



$$\int_x p(x|y = y_1)dx = 1$$

$$\int_x p(x|y = y_2)dx = 1$$

Adapted from S. Prince

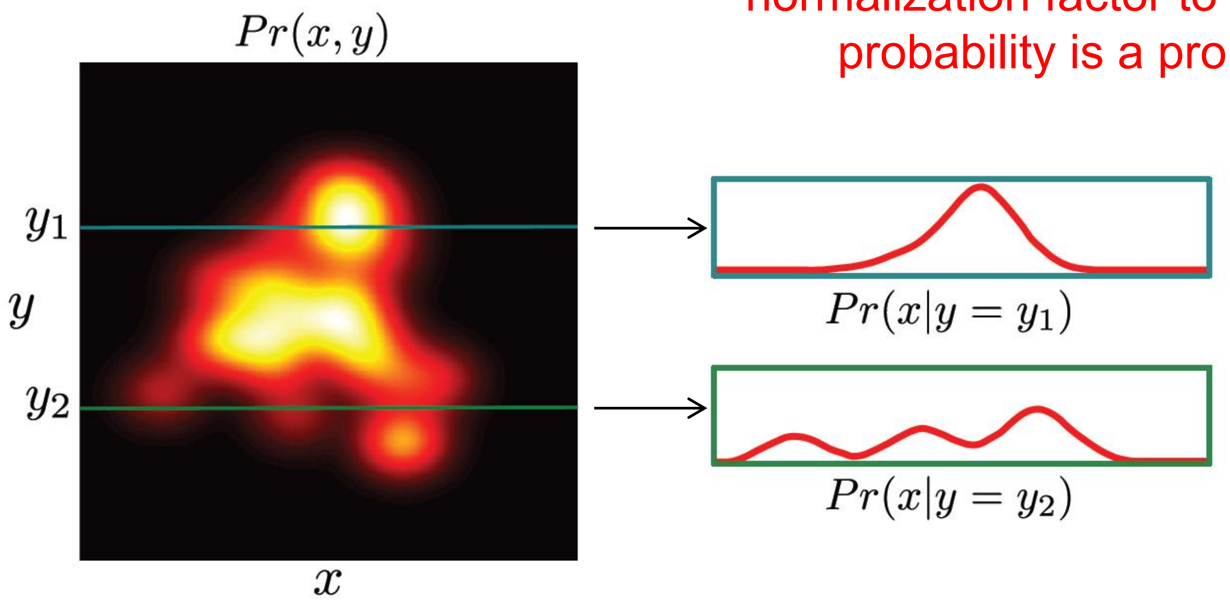
Conditional Probability

- Conditional probability can be computed from joint probability.

slice of joint distribution

$$p(x|y=y^*) = \frac{p(x,y=y^*)}{p(y=y^*)} = \frac{p(x,y=y^*)}{\int p(x,y=y^*)dx}$$

normalization factor to ensure conditional probability is a proper distribution



$$\int_x p(x|y=y_1)dx = 1$$

$$\int_x p(x|y=y_2)dx = 1$$

Adapted from S. Prince

Conditional Probability

$$p(x|y = y^*) = \frac{p(x, y = y^*)}{p(y = y^*)} = \frac{p(x, y = y^*)}{\int p(x, y = y^*)dx}$$

- More usually written in compact form

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

Adapted from S. Prince

Conditional Probability

$$p(x|y=y^*) = \frac{p(x,y=y^*)}{p(y=y^*)} = \frac{p(x,y=y^*)}{\int p(x,y=y^*)dx}$$

- More usually written in compact form

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

- Can be re-arranged to give

$$p(x,y) = p(y)p(x|y)$$
$$p(x,y) = p(x)p(y|x)$$

Adapted from S. Prince

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$p(x|y=-1) =$???

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

Conditional Probability Example

		x = 0	x = 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) =$$

$$p(x=2.5|y=-1) =$$

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) = \frac{p(x=0, y=-1)}{p(y=-1)}$$

$$p(x=2.5|y=-1) =$$

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) = \frac{p(x=0, y=-1)}{p(y=-1)} = \frac{1/8}{3/8} = \frac{1}{3}$$

$$p(x=2.5|y=-1) =$$

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) = \frac{p(x=0, y=-1)}{p(y=-1)} = \frac{1/8}{3/8} = \frac{1}{3}$$

$$p(x=2.5|y=-1) = \frac{p(x=2.5, y=-1)}{p(y=-1)} = \frac{1/4}{3/8} = \frac{2}{3}$$

Bayes' Rule

Deriving Bayes' Rule (y continuous)

From before:

$$p(x,y) = p(y)p(x|y)$$

$$p(x,y) = p(x)p(y|x)$$

Equate RHS

Combining:

$$p(y)p(x|y) = p(x)p(y|x)$$

Re-arranging:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

$$p(x) = \sum_y p(x,y)$$

$$= \frac{p(y)p(x|y)}{\sum_y p(x,y)}$$

$$p(x,y) = p(y)p(x|y)$$

$$= \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

Adapted from S. Prince

Bayes' Rule

Prior – what we know about y BEFORE seeing x

Likelihood – propensity for observing a certain value of x given a certain value of y

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

Posterior – what we know about y AFTER seeing x

Evidence – a constant to ensure that the left hand side is a valid distribution

Adapted from S. Prince

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

$$p(y = 1|x = 1) = \frac{p(y = 1)p(x = 1|y = 1)}{p(y = 1)p(x = 1|y = 1) + p(y = 0)p(x = 1|y = 0)}$$

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

$$
\begin{aligned}
p(y = 1|x = 1) &= \frac{p(y = 1)p(x = 1|y = 1)}{p(y = 1)p(x = 1|y = 1) + p(y = 0)p(x = 1|y = 0)} \\
&= \frac{0.004 \times 0.8}{0.004 \times 0.8 + 0.996 \times 0.1} = 0.031
\end{aligned}
$$

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

$$
\begin{aligned}
p(y = 1|x = 1) &= \frac{p(y = 1)p(x = 1|y = 1)}{p(y = 1)p(x = 1|y = 1) + p(y = 0)p(x = 1|y = 0)} \\
&= \frac{0.004 \times 0.8}{0.004 \times 0.8 + 0.996 \times 0.1} = 0.031
\end{aligned}
$$

- US government no longer recommend mammogram for women in 40s

Independence

Independence

- If x & y are independent, then knowing x tells us nothing about y (and vice versa):

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

Independence

- If x & y are independent, then knowing x tells us nothing about y (and vice versa):

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

- If x & y are independent, then joint distribution factorizes into product of marginal distributions:

$$\begin{aligned} p(x, y) &= p(x)p(y|x) \\ &= p(x)p(y) \end{aligned}$$

Independence

- If x & y are independent, then knowing x tells us nothing about y (and vice versa):

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

- If x & y are independent, then joint distribution factorizes into product of marginal distributions:

$$\begin{aligned} p(x,y) &= p(x)p(y|x) \\ &= p(x)p(y) \end{aligned}$$

- Conversely, if joint distribution can be factorized into product of marginal distributions, then x & y are independent

Expectation of One Random Variable

Expectation

Expectation tells us the average (i.e., expected) value of some function f(x) taking into account the distribution of x

Definition:

$$E[f(x)] = \sum_{x} f(x)p(x)$$

$$E[f(x)] = \int f(x)p(x)dx$$

Adapted from S. Prince

Expectation: Mean and Variance

$$E[f(x)] = \int f(x)p(x)dx$$

- If $f(x) = x$
  - $E[f(x)] = E(x) = \mu_x$, the "mean of x"
  - If we observe x many (infinite) times and average, we get μ_x

Expectation: Mean and Variance

$$E[f(x)] = \int f(x)p(x)dx$$

- If $f(x) = x$
  - $E[f(x)] = E(x) = \mu_x$, the "mean of x"
  - If we observe x many (infinite) times and average, we get μ_x
- If $f(x) = (x - \mu_x)^2$
  - $E[f(x)] = E[(x - \mu_x)^2] = \sigma_x^2$
  - $\sigma_x^2 = Var(x)$ called "variance"; σ_x called "standard deviation"
  - If we observe x many (infinite) times and average square of difference between each observation and μ_x, we get σ_x^2
  - Measure how likely x is going to be far away from mean

Expectation Example: Mean

		x		
		0	2.5	p(y)
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E(x) = \sum_x xp(x)$$

Expectation Example: Mean

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$\begin{aligned} E(x) &= \sum_x xp(x) \\ &= 0 \times 1/4 + 2.5 \times 3/4 \\ &= 0 + 1.875 \\ &= 1.875 \end{aligned}$$

Expectation Example: Mean

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$\begin{aligned} E(x) &= \sum_x xp(x) \\ &= 0 \times 1/4 + 2.5 \times 3/4 \\ &= 0 + 1.875 \\ &= 1.875 \end{aligned}$$

$$\begin{aligned} E(y) &= \sum_y yp(y) \\ &= -3 \times 1/2 + (-1) \times 3/8 + 2 \times 1/8 \\ &= -3/2 - 3/8 + 1/4 \\ &= -1.625 \end{aligned}$$

Expectation Example: Variance

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$\begin{aligned} E(x) &= \sum_x xp(x) \\ &= 0 \times 1/4 + 2.5 \times 3/4 \\ &= 0 + 1.875 \\ &= 1.875 \end{aligned}$$

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E[(x - \mu_x)^2] = \sum_x (x - \mu_x)^2 p(x)$$

Expectation Example: Variance

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$\begin{aligned} E(x) &= \sum_x xp(x) \\ &= 0 \times 1/4 + 2.5 \times 3/4 \\ &= 0 + 1.875 \\ &= 1.875 \end{aligned}$$

$$E[f(x)] = \sum_x f(x)p(x)$$

$$\begin{aligned} E[(x-\mu_x)^2] &= \sum_x (x-\mu_x)^2 p(x) \\ &= (0-1.875)^2 \times 1/4 + (2.5-1.875)^2 \times 3/4 \\ &= 1.171875 \end{aligned}$$

Expectation Example: Variance

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$\begin{aligned} E(x) &= \sum_x xp(x) \\ &= 0 \times 1/4 + 2.5 \times 3/4 \\ &= 0 + 1.875 \\ &= 1.875 \end{aligned}$$

$$E[f(x)] = \sum_x f(x)p(x)$$

$$\begin{aligned} E[(x-\mu_x)^2] &= \sum_x (x-\mu_x)^2 p(x) \\ &= (0-1.875)^2 \times 1/4 + (2.5-1.875)^2 \times 3/4 \\ &= 1.171875 \end{aligned}$$

Expectation of Two Random Variables

Expectation for X, Y

- Expectation tells us the expected or average value of some function $f(x, y)$ taking into account $p(x, y)$

$$E[f(x,y)] = \int \int f(x,y)p(x,y)dxdy$$

Expectation for X, Y

- Expectation tells us the expected or average value of some function $f(x, y)$ taking into account $p(x, y)$

$$E[f(x,y)] = \int\int f(x,y)p(x,y)dxdy$$

- Special case: $f(x, y) = (x–\mu_x)(y–\mu_y)$
  - $E[f(x, y)] = E[(x–\mu_x)(y–\mu_y)] = Cov(x, y)$, the covariance of x and y
  - Measure how much two variables change together
  - $Cov(x, y)$ positive whenever $x > \mu_x$, then $y > \mu_y$ <u>on average</u> (& vice versa)
  - $Cov(x, y)$ negative whenever $x > \mu_x$, then $y < \mu_y$ <u>on average</u> (& vice versa)

Expectation Example: Covariance

		x: 0	x: 2.5	p(y)
y	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
y	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$

$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$E[(x-\mu_x)(y-\mu_y)]$$

$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y)$$

Expectation Example: Covariance

		x: 0	x: 2.5	p(y)
y	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
y	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$\begin{aligned}
&E[(x-\mu_x)(y-\mu_y)] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \\
&= (0-1.875)(-3+1.625)\times 0 + (2.5-1.875)(-3+1.625)\times 1/2 + \\
&\quad (0-1.875)(-1+1.625)\times 1/8 + (2.5-1.875)(-1+1.625)\times 1/4 + \\
&\quad (0-1.875)(2+1.625)\times 1/8 + (2.5-1.875)(2+1.625)\times 0 \\
&= -1.328125
\end{aligned}$$

Expectation Example: Covariance

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$\begin{aligned}
&E[(x-\mu_x)(y-\mu_y)] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \\
&= (0-1.875)(-3+1.625) \times 0 + (2.5-1.875)(-3+1.625) \times 1/2 + \\
&\quad (0-1.875)(-1+1.625) \times 1/8 + (2.5-1.875)(-1+1.625) \times 1/4 + \\
&\quad (0-1.875)(2+1.625) \times 1/8 + (2.5-1.875)(2+1.625) \times 0 \\
&= -1.328125
\end{aligned}$$

Expectation Example: Covariance

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$\begin{aligned}
&E[(x-\mu_x)(y-\mu_y)] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \\
&= (0-1.875)(-3+1.625)\times 0 + (2.5-1.875)(-3+1.625)\times 1/2 + \\
&\quad (0-1.875)(-1+1.625)\times 1/8 + (2.5-1.875)(-1+1.625)\times 1/4 + \\
&\quad (0-1.875)(2+1.625)\times 1/8 + (2.5-1.875)(2+1.625)\times 0 \\
&= -1.328125
\end{aligned}$$

Expectation Example: Covariance

		x: 0	x: 2.5	p(y)
y	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
y	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$\begin{aligned}
&E[(x-\mu_x)(y-\mu_y)] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \\
&= (0-1.875)(-3+1.625)\times 0 + (2.5-1.875)(-3+1.625)\times 1/2 + \\
&\quad (0-1.875)(-1+1.625)\times 1/8 + (2.5-1.875)(-1+1.625)\times 1/4 + \\
&\quad (0-1.875)(2+1.625)\times 1/8 + (2.5-1.875)(2+1.625)\times 0 \\
&= -1.328125
\end{aligned}$$

Expectation Example: Covariance

		x		
		0	2.5	p(y)
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$\begin{aligned}
&E[(x-\mu_x)(y-\mu_y)] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \\
&= (0-1.875)(-3+1.625)\times 0 + (2.5-1.875)(-3+1.625)\times 1/2 + \\
&\quad (0-1.875)(-1+1.625)\times 1/8 + (2.5-1.875)(-1+1.625)\times 1/4 + \\
&\quad (0-1.875)(2+1.625)\times 1/8 + (2.5-1.875)(2+1.625)\times 0 \\
&= -1.328125
\end{aligned}$$

Expectation Example: Covariance

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$

$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$\begin{aligned}
&E[(x-\mu_x)(y-\mu_y)] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \\
&= (0-1.875)(-3+1.625)\times 0 + (2.5-1.875)(-3+1.625)\times 1/2 + \\
&\quad (0-1.875)(-1+1.625)\times 1/8 + (2.5-1.875)(-1+1.625)\times 1/4 + \\
&\quad (0-1.875)(2+1.625)\times 1/8 + (2.5-1.875)(2+1.625)\times 0 \\
&= -1.328125
\end{aligned}$$

Rules of Expectation

1. Expectation of a non-random constant: $E(c) = c$

Rules of Expectation

1. Expectation of a non-random constant: $E(c) = c$
2. Assuming a and b are non-random constants, then
    - $E[af(x) + bg(x)] = aE(f(x)) + bE(g(x))$
    - This is called "linearity of expectation"

Rules of Expectation

1. Expectation of a non-random constant: $E(c) = c$

2. Assuming a and b are non-random constants, then

   - $E[af(x) + bg(x)] = aE(f(x)) + bE(g(x))$
   - This is called "linearity of expectation"

3. $Cov(x, y) = E(xy) - E(x)E(y)$

   - Therefore $Var(x) \equiv Cov(x, x) = E(x^2) - (E(x))^2$

Rules of Expectation

1. Expectation of a non-random constant: $E(c) = c$

2. Assuming a and b are non-random constants, then

   - $E[af(x) + bg(x)] = aE(f(x)) + bE(g(x))$
   - This is called "linearity of expectation"

3. $Cov(x, y) = E(xy) - E(x)E(y)$

   - Therefore $Var(x) \equiv Cov(x, x) = E(x^2) - (E(x))^2$

4. If x and y are independent, then $E[f(x)g(y)] = E[f(x)]E[g(y)]$

   - If x and y are independent, $Cov(x, y) = 0$
   - However $Cov(x, y) = 0 \nRightarrow$ independence

Conditional Expectation

Conditional Expectation

- Remember $p(x|y)$ is conditional probability of x given y?

Conditional Expectation

- Remember $p(x|y)$ is conditional probability of x given y?
- Conditional expectation:

$$\begin{aligned} E[f(x,y)|y] &\triangleq E_{p(x|y)}[f(x,y)] \\ &\triangleq \sum_x f(x,y)p(x|y) \end{aligned}$$

- Read as "expected value of $f(x,y)$ given y"

Conditional Expectation

- Remember $p(x|y)$ is conditional probability of x given y?
- Conditional expectation:

$$\begin{aligned} E[f(x,y)|y] &\triangleq E_{p(x|y)}[f(x,y)] \\ &\triangleq \sum_x f(x,y)p(x|y) \end{aligned}$$

- Read as "expected value of $f(x,y)$ given y"
- Conditional expectation tells us average value of $f(x,y)$ taking into account $p(x|y)$

Conditional Expectation

- Remember $p(x|y)$ is conditional probability of x given y?
- Conditional expectation:

$$
\begin{aligned}
E[f(x,y)|y] &\triangleq E_{p(x|y)}[f(x,y)] \\
&\triangleq \int_x f(x,y)p(x|y)dx
\end{aligned}
$$

- Read as "expected value of $f(x,y)$ given y"
- Conditional expectation tells us average value of $f(x,y)$ taking into account $p(x|y)$

Conditional Expectation Example

		x		
		0	2.5	p(y)
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x = 0|y = -1) = \frac{1}{3}$$

$$p(x = 2.5|y = -1) = \frac{2}{3}$$

$$E(x) = 1.875$$

$$E_{p(x|y=-1)}[x] =$$

Conditional Expectation Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x=0|y=-1)=\frac{1}{3}$$

$$p(x=2.5|y=-1)=\frac{2}{3}$$

$$E(x)=1.875$$

$$E_{p(x|y=-1)}[x]=\sum_x xp(x|y=-1) \leftarrow \sum_x f(x,y)p(x|y)$$

Conditional Expectation Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x = 0 | y = -1) = \frac{1}{3}$$

$$p(x = 2.5 | y = -1) = \frac{2}{3}$$

$$E(x) = 1.875$$

$$\sum_x f(x, y) p(x|y)$$

$$E_{p(x|y=-1)}[x] = \sum_x x p(x|y = -1)$$

$$= 0 \times p(x = 0|y = -1) + 2.5 \times p(x = 2.5|y = -1)$$

Conditional Expectation Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x=0|y=-1)=\frac{1}{3}$$

$$p(x=2.5|y=-1)=\frac{2}{3}$$

$$E(x)=1.875$$

$$E_{p(x|y=-1)}[x]=\sum_x xp(x|y=-1)$$

$$\sum_x f(x,y)p(x|y)$$

$$\begin{aligned}&=0\times p(x=0|y=-1)+2.5\times p(x=2.5|y=-1)\\&=0\times 1/3+2.5\times 2/3\\&=1.667\end{aligned}$$

N Random Variables

N random variables (aka random vector)

- We have focused on 2 random variables x and y
- In real applications, usually more than 2 variables (e.g., photo has > 1M pixels)
- If we observe $x_1, x_2, \ldots, x_N$ multiple times, some combinations of outcomes more likely than others
- This information captured by joint probability distribution function
- Written as $p(x_1, x_2, \ldots, x_N)$, read as probability distribution of x_1 to x_N
- If $x_1, x_2, \ldots, x_N$ are continuous, then p refers to joint probability distribution function (pdf). If discrete, then refers to joint probability mass function (pmf)
- Many properties for two random variables **generalize** naturally to more variables

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables.

$$
\begin{aligned}
Pr(x) &= \int Pr(x,y)\, dy \\
Pr(y) &= \int Pr(x,y)\, dx
\end{aligned}
$$

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables.

$$
\begin{aligned}
Pr(x) &= \int Pr(x,y)\, dy \\
Pr(y) &= \int Pr(x,y)\, dx
\end{aligned}
$$

It works in higher dimensions as well – leaves joint distribution between whatever variables are left.

$$Pr(x,y) = \sum_{w} \int Pr(w,x,y,z)\, dz$$

Conditional Probability

- Two variables

$$p(x, y) = p(x)p(y|x)$$

Conditional Probability

- Two variables

$$p(x, y) = p(x)p(y|x)$$

- Three variables

$$p(a, b, c) = p(a)p(b, c|a) = p(a)p(b|a)p(c|a, b)$$

Conditional Probability

- Two variables

$$p(x, y) = p(x)p(y|x)$$

- Three variables

$$p(a, b, c) = p(a)p(b, c|a) = p(a)p(b|a)p(c|a, b)$$

- N variables

$$\begin{aligned} p(x_1, \cdots, x_N) &= p(x_1)p(x_2, \cdots, x_N|x_1) \\ &= p(x_1)p(x_2|x_1)p(x_3, \cdots, x_N|x_1, x_2) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_N|x_1, \cdots, x_{N-1}) \end{aligned}$$

Independence

- If $x_1, \cdots, x_N$ are independent, then knowing any subset of x's tells us nothing about the remaining x's

Independence

- If $x_1, \cdots, x_N$ are independent, then knowing any subset of x's tells us nothing about the remaining x's
- If $x_1, \cdots, x_N$ are independent if and only if the joint distribution factorizes into product of marginal distributions:

$$p(x_1, \cdots, x_N) = p(x_1)p(x_2)\cdots p(x_N) \triangleq \prod_{n=1}^{N} p(x_n)$$

Independence

- If $x_1, \cdots, x_N$ are independent, then knowing any subset of x's tells us nothing about the remaining x's

- If $x_1, \cdots, x_N$ are independent <u>if and only if</u> the joint distribution factorizes into product of marginal distributions:

$$p(x_1, \cdots, x_N) = p(x_1)p(x_2)\cdots p(x_N) \triangleq \prod_{n=1}^{N} p(x_n)$$

- $x_1, \cdots, x_N$ are <u>independently and identically distributed</u> (i.i.d.) if they are independent and $p(x_1) = p(x_2) = \cdots = p(x_N)$

Conditional Independence

- x_1 and x_2 are conditionally independent given x_3 if and only if

$$p(x_1, x_2|x_3) = p(x_1|x_3)p(x_2|x_3)$$

Knowing x_2 tells us nothing about x_1 (and vice versa) if we already know x_3

Lecture 6 – Probability Summary

- Random variable:
  - Discrete (Bernoulli, Categorical, Binomial, Geometric, Poisson)
  - Continuous (Gaussian, Uniform, Exponential, Beta)
- Joint probability p(x, y)
- Marginalization / Law of Total Probability
- Conditional Probability p(x | y)
- Bayes' Rule
- Expectation & Conditional expectation
- Extension to N random variables

Homework 2

Tips:

- Watch the tutorial video
  https://www.bilibili.com/video/BV1YK411T7H8
- Read the official website of MNE-python
  https://mne.tools/stable/index.html

EEG data processing with MNE-python

DDL: March 29, 2021，上课前

把code和结果图做成pdf发邮件给曲由之 12031145@mail.sustech.edu.cn

Requirements

1. Read the paper https://www.nature.com/articles/s41597-020-0535-2
2. Download the raw EEG data from [26] in the paper (https://doi.org/10.7910/DVN/RBN3XG)
   学生ID为奇数的同学下载sub-001；学生ID为偶数的同学下载sub-002
3. Plot the **time course** of raw EEG signals with 10-second window (as Figure 4 in the paper).
4. Data preprocessing, artifacts removal
5. Plot the **time course** of preprocessed EEG signals
6. Plot the **time-frequency maps** of the subject (as Figure 6 in the paper)
7. Plot the **topographical distribution** of power of the subject (as Figure 7 in the paper)
8. **Comparison** of power (in dB) changes with time (in s) during hand, elbow motor imagery, and resting state for electrode C3, and electrode C4 (as Figure 8 in the paper)

Reading materials

Textbooks

- Chapter 2 (Probability) of MLaPP (beware of typos)

These notes are from Chapters 2 & 3 of Computer Vision: Models, Learning and Inference, by Simmon Prince

free download: www.computervisionmodels.com

- Please review of probability (by John Tsitsiklis) & linear algebra (by Gilbert Strang).