Mathematical Basics

Yu Yao

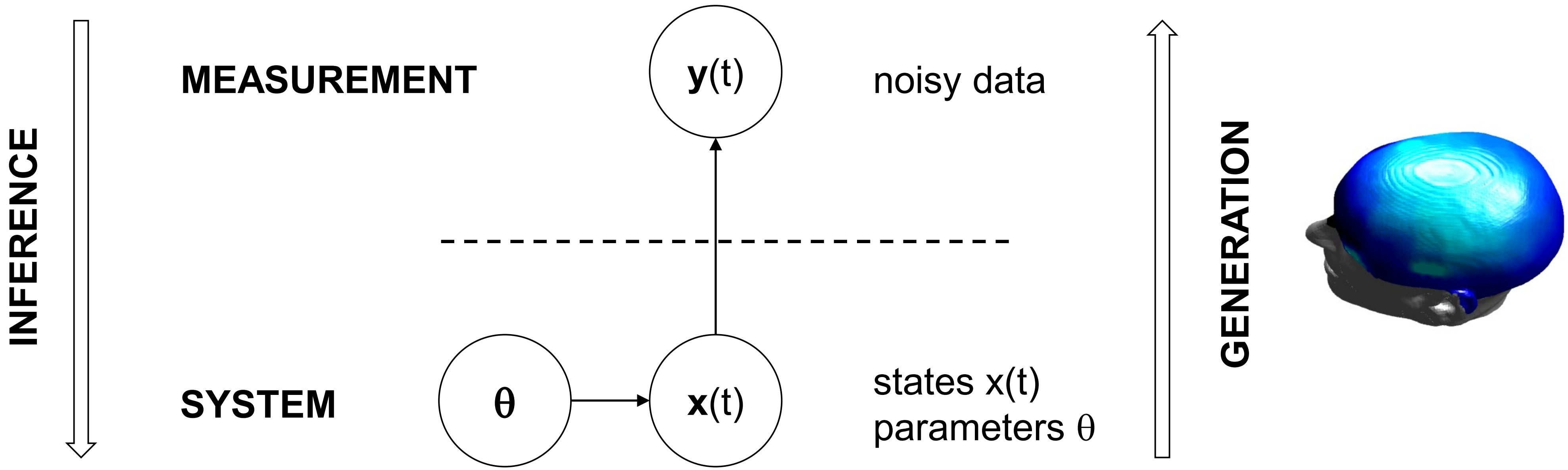
Translational Neuromodeling Unit

Universität Zürich[UZH]

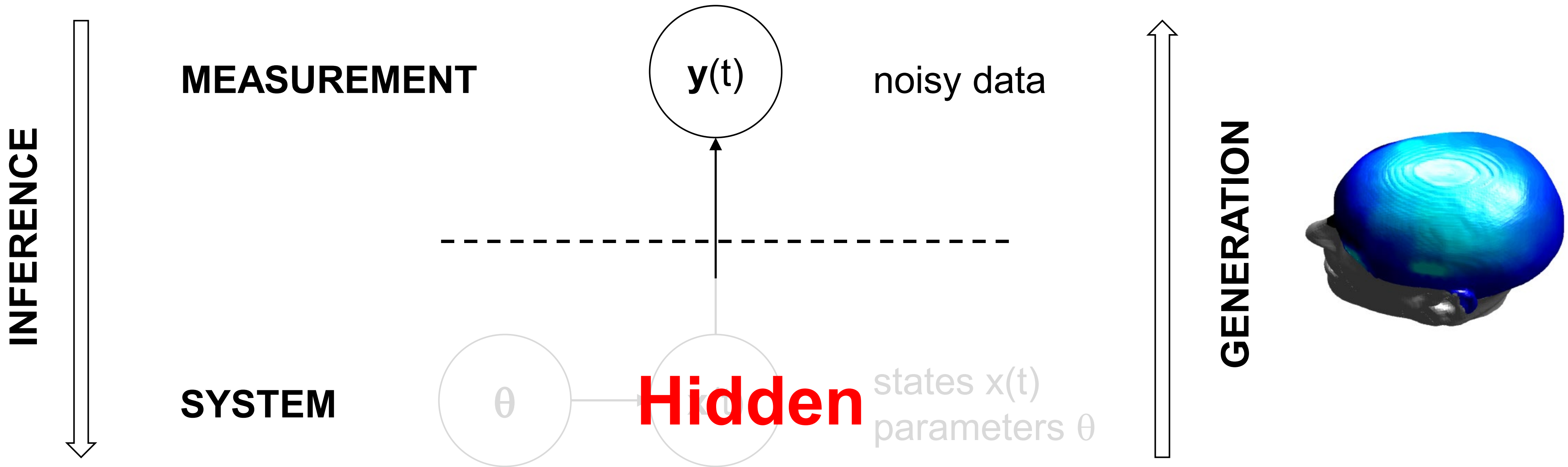
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computational Psychiatry Course 2019
Zurich | 3th September 2019

Introduction



Introduction



Contents

Random Variables and Probability

- **Random variable:** a variable whose possible values are outcomes (events) of a random experiment, e.g.:
  - rolling a dice: eye count (1 … #faces)
  - tossing a coin: side of coin (head or tail)
  - taking the temperature: temperature value (real number)

Random Variables and Probability

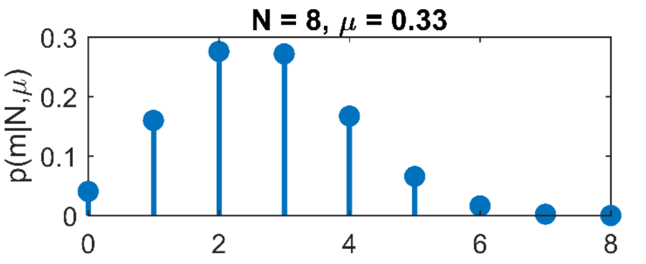
- **Random variable:** a variable whose possible values are outcomes (events) of a random experiment, e.g.:
  - rolling a dice: eye count (1 … #faces)
  - tossing a coin: side of coin (head or tail)
  - taking the temperature: temperature value (real number)
- **Random vector:** a vector of random variables

Random Variables and Probability

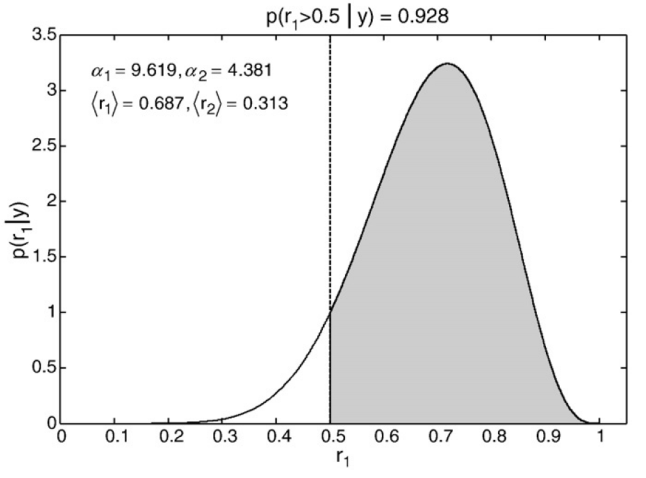
- **Random variable:** a variable whose possible values are outcomes (events) of a random experiment, e.g.:
  - rolling a dice: eye count (1 … #faces)
  - tossing a coin: side of coin (head or tail)
  - taking the temperature: temperature value (real number)
- **Random vector:** a vector of random variables
- **Probability of an event:** number of occurrence of particular outcome (event) / number of times experiment was repeated
  - Map: outcome (event) → number

Random Variables and Probability

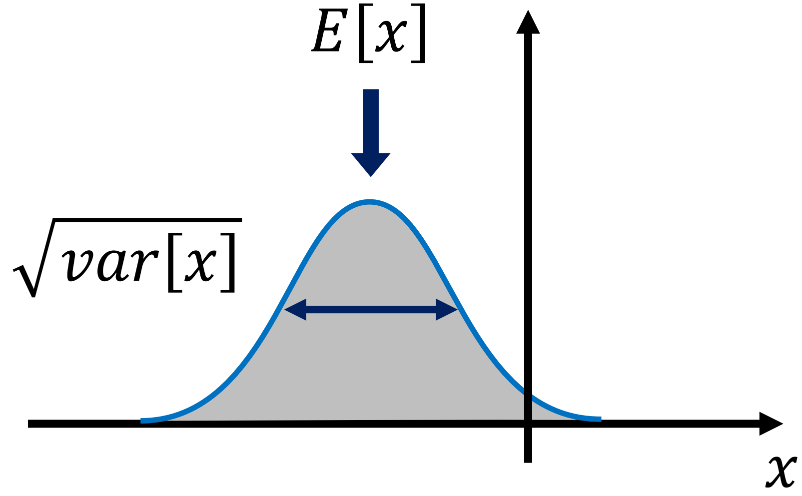
- **Probability distribution:**
  - describes the probability that a **discrete** random variable takes on a particular value



- **Probability density:**
  - describes the probability of a **continuous** random variable falling within a particular range of values



Probability Distributions and Densities



Properties:

$$p(x) \geq 0 \; \forall x$$

$$\sum_x p(x) = 1 \text{ or } \int p(x)dx = 1$$

mean/expectation: variance:

$$E[x] = \sum_x xp(x) \qquad var[x] = \sum_x (x - E[x])^2 p(x)$$

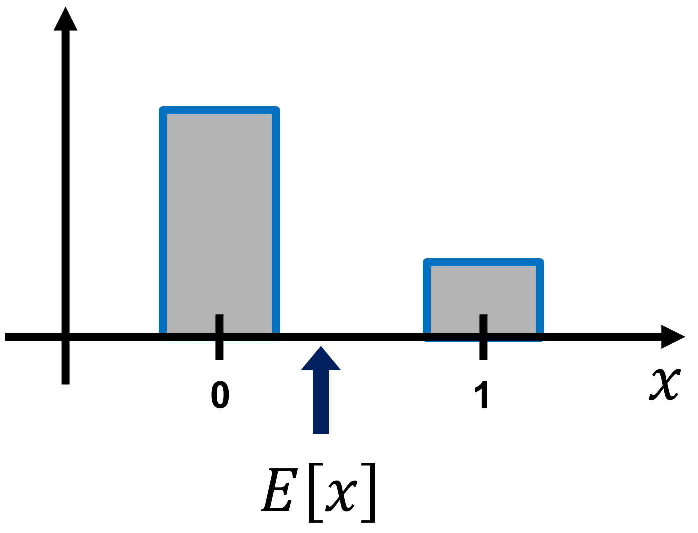
$$E[x] = \int xp(x)dx \qquad var[x] = \int (x - E[x])^2 p(x)dx$$

Probability Distributions and Densities

Properties:

$$p(x) \geq 0 \;\forall x$$

$$\sum_x p(x) = 1 \;\text{ or }\; \int p(x)dx = 1$$



mean/expectation: variance:

$$E[x] = \sum_x xp(x) \qquad var[x] = \sum_x (x - E[x])^2 p(x)$$

$$E[x] = \int xp(x)dx \qquad var[x] = \int (x - E[x])^2 p(x)dx$$

Joint and Conditional Probability

Systems with more than one random variable

Example: rolling a pair of dice. x: 1st dice, y: is sum even or odd

Joint and Conditional Probability

Systems with more than one random variable

Example: rolling a pair of dice. *x*: 1st dice, *y*: is sum even or odd

Conditional Probability: Probability of subset of variables, if remaining variables known

Example: sum of dice: *y* = even, **if** 1st dice: *x* = 3

$$p\left(y = even \mid x = 3\right) = ?$$

Joint and Conditional Probability

Systems with more than one random variable

Example: rolling a pair of dice. *x*: 1st dice, *y*: is sum even or odd

Conditional Probability: Probability of subset of variables, if remaining variables known

Example: sum of dice: *y* = even, **if** 1st dice: *x* = 3

$$p\left(y = even \mid x = 3\right) = ?$$

If 1st dice 3, second dice must be odd (1, 3, 5)

$$p\left(y = even \mid x = 3\right) = \frac{1}{2}$$

Joint and Conditional Probability

Systems with more than one random variable

Example: rolling a pair of dice. *x*: 1st dice, *y*: is sum even or odd

Joint Probability: Probability of configuration involving all variables

Example: 1st dice: *x* = 3 **and** sum of dice: *y* = even

$$p\left(x=3, y=even\right)=?$$

Joint and Conditional Probability

Systems with more than one random variable

Example: rolling a pair of dice. *x*: 1st dice, *y*: is sum even or odd

Joint Probability: Probability of configuration involving all variables

Example: 1st dice: $x = 3$ **and** sum of dice: y = even

$$p\left(x = 3, y = even\right) = ?$$

$$p\left(x = 3\right) = \frac{1}{6}$$

Joint and Conditional Probability

Systems with more than one random variable

Example: rolling a pair of dice. *x*: 1st dice, *y*: is sum even or odd

Joint Probability: Probability of configuration involving all variables

Example: 1st dice: *x* = 3 **and** sum of dice: *y* = even

$$p\left(x=3, y=even\right)=?$$

$$p\left(x=3\right)=\frac{1}{6}$$

(second dice must be odd)

$$p\left(y=even \mid x=3\right)=\frac{1}{2}$$

Joint and Conditional Probability

Systems with more than one random variable

Example: rolling a pair of dice. *x*: 1st dice, *y*: is sum even or odd

Joint Probability: Probability of configuration involving all variables

Example: 1st dice: *x* = 3 **and** sum of dice: *y* = even

$$p\left(x=3, y=even\right)=?$$

(second dice must be odd)

$$p\left(x=3\right)=\frac{1}{6} \qquad p\left(y=even \mid x=3\right)=\frac{1}{2}$$

$$p\left(x=3, y=even\right)=p\left(x=3\right)\cdot p\left(y=even \mid x=3\right)=\frac{1}{6}\cdot\frac{1}{2}=\frac{1}{12}$$

Joint and Conditional Probability

Systems with more than one random variable

Example: rolling a pair of dice. *x*: 1st dice, *y*: is sum even or odd

Joint Probability: Probability of configuration involving all variables

Example: 1st dice: *x* = 3 **and** sum of dice: *y* = even

$$p\left(x=3, y=even\right)=?$$

(second dice must be odd)

$$p\left(x=3\right)=\frac{1}{6} \qquad p\left(y=even \mid x=3\right)=\frac{1}{2}$$

Product rule

$$p\left(x=3, y=even\right)=p\left(x=3\right)\cdot p\left(y=even \mid x=3\right)=\frac{1}{6}\cdot\frac{1}{2}=\frac{1}{12}$$

Notation

- Notation example (normal densities):
  - for scalars: $p(x) = N(x; \mu, \sigma^2)$ μ = mean; σ^2 = variance
  - for vectors: $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ Σ = covariance matrix = $E[\,(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\,]$

same thing, just expressed wrt. precision

- Notation example (normal densities):
  - for scalars: $\boxed{p(x) = N(x; \mu, \lambda^{-1})}$ μ = mean; $\lambda = 1/\sigma^2$ = precision
  - for vectors: $\boxed{p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \Lambda^{-1})}$ Λ = precision matrix

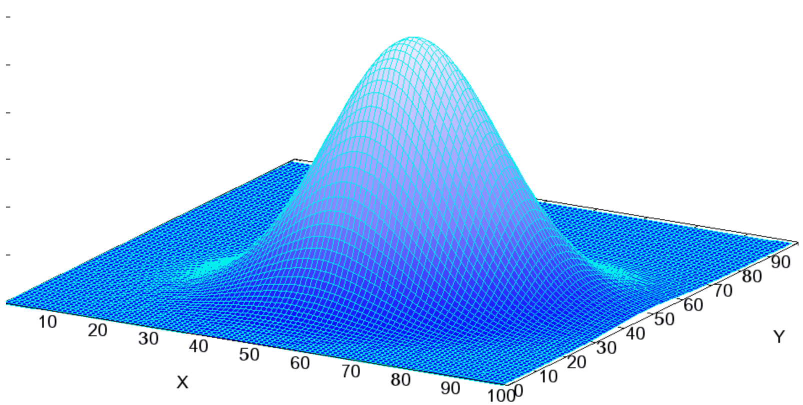
Example: Multivariate Gaussian/Normal

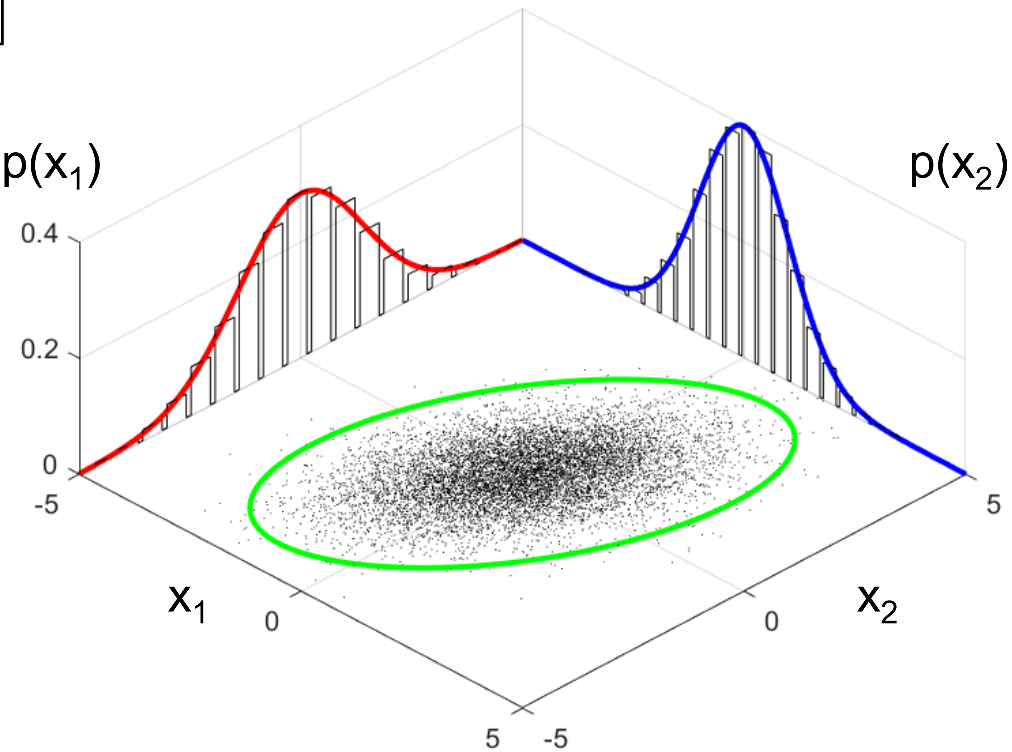
p-dimensional random vector: $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

PDF: $$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right)$$

covariance matrix: $$\boldsymbol{\Sigma} = \mathbf{E}\left[\left((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \right) \right]$$





Figures adapted from *Wikipedia*

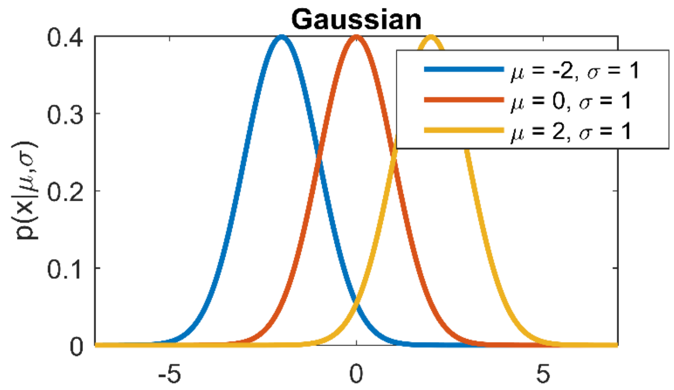
Example: Multivariate Gaussian/Normal

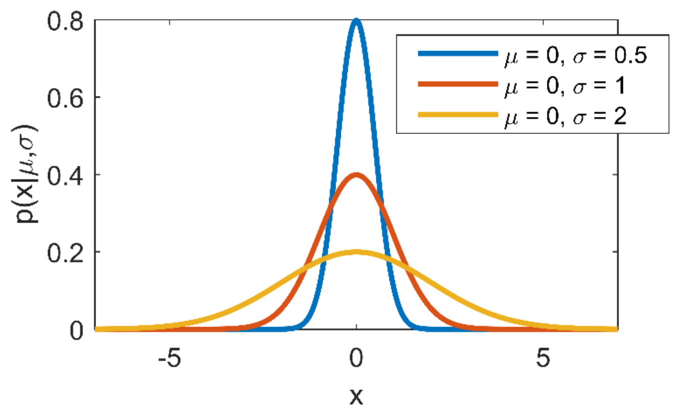
p-dimensional random vector: $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

PDF: $$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right)$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ \end{bmatrix}$$

covariance matrix: $$\boldsymbol{\Sigma} = \mathbf{E}\left[\left((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \right) \right]$$





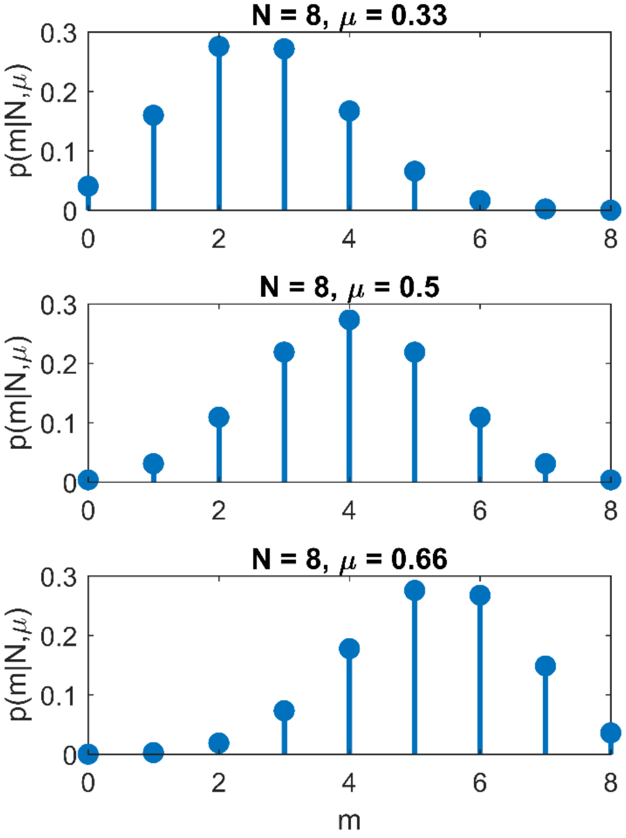
- For continuous and unconstrained random variable
- Models measurement errors
- Central limit theorem

Example: Binomial and Multinomial

- For discrete random variables
- Models number of outcomes/events
- Probability vector over events (mu) -> event counts (m)

$$p(m) = \binom{N}{m_1 m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\binom{N}{m_1 m_2 \cdots m_K} = \frac{N!}{m_1!\, m_2! \cdots m_K!}$$



Application: Bayesian model selection for group studies
(Stephan et al. (2009) *NeuroImage*)

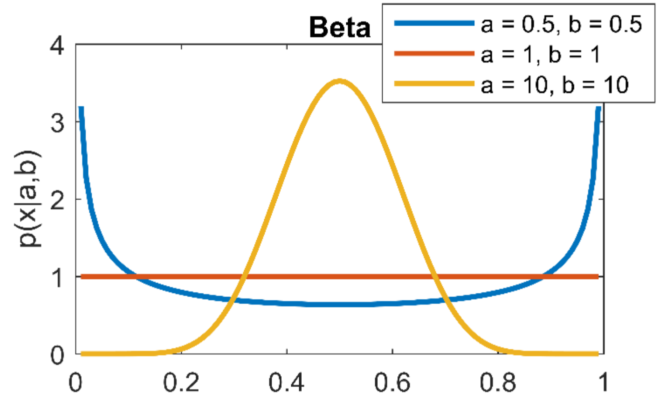
Example: Beta and Dirichlet

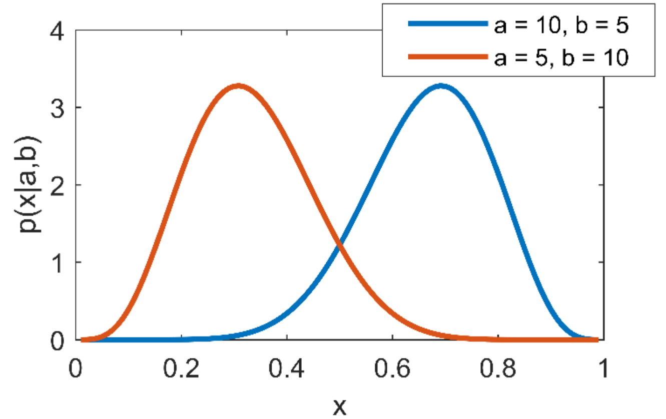
$\sum_{k=1}^{K} \mu_k = 1$ and $\mu_k \geq 0$

- For continuous random variables on unit simplex
- Models distribution over probability vectors
- (Observed) event counts (alpha) -> probability of events (mu)

$$p(\mu) = Dir(\mu|\alpha) = C(\alpha) \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$C(\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}$$





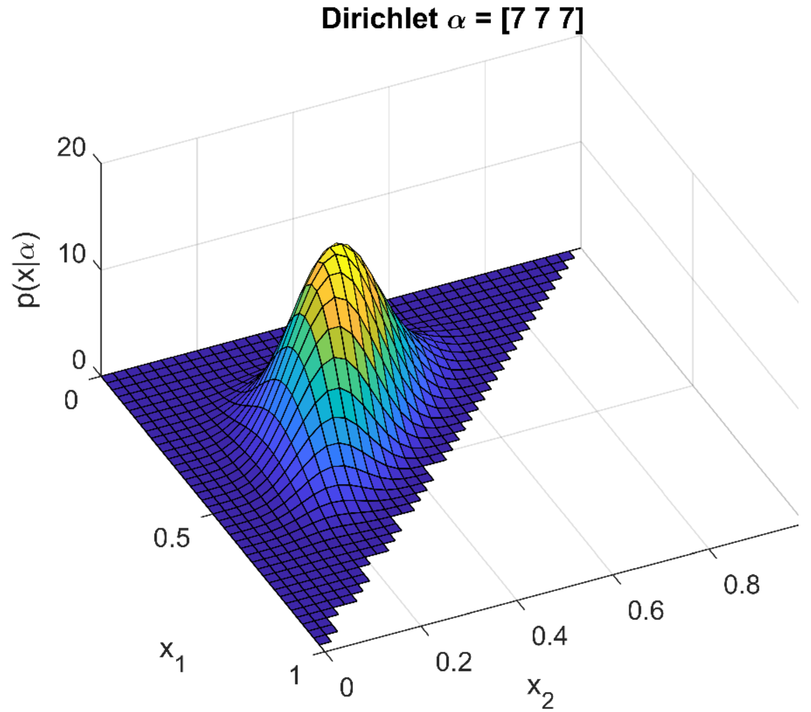
Example: Beta and Dirichlet

$\sum_{k=1}^{K} \mu_k = 1$ and $\mu_k \geq 0$

- For continuous random variables on unit simplex
- Models distribution over probability vectors
- (Observed) event counts (alpha) -> probability of events (mu)

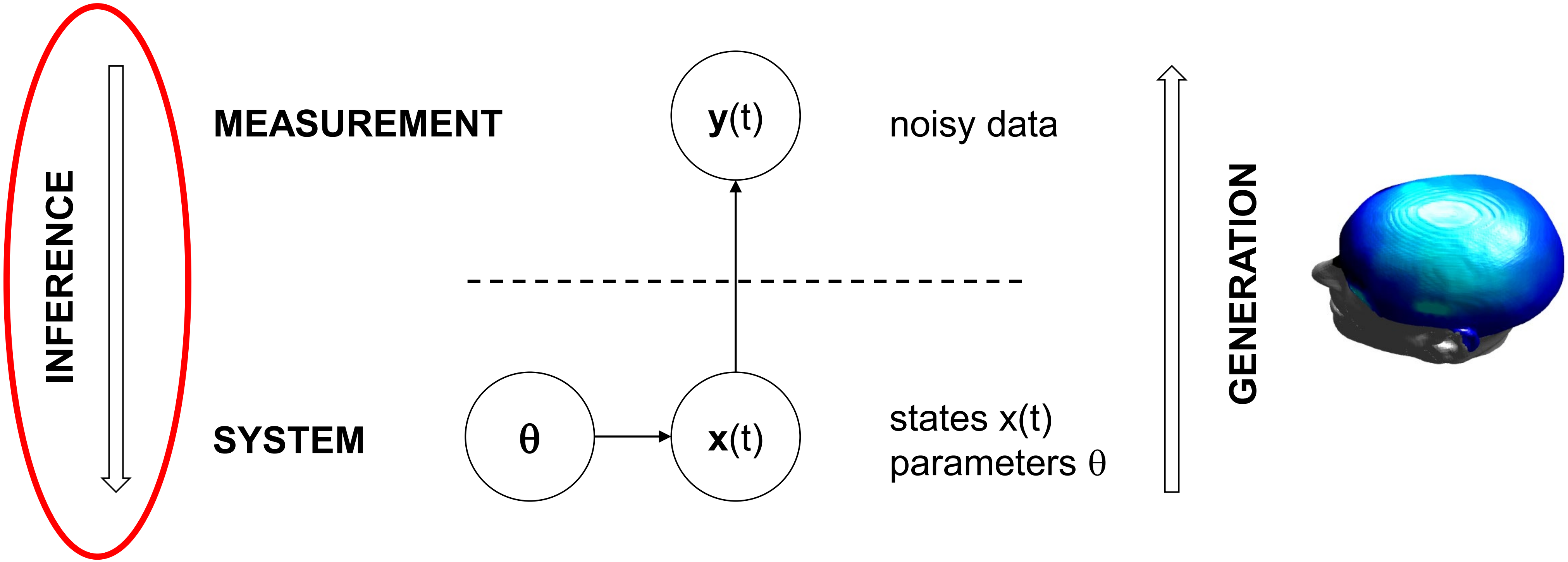
$$p(\mu) = Dir(\mu|\alpha) = C(\alpha) \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$C(\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}$$



Application: Bayesian model selection for group studies
(Stephan et al. (2009) *NeuroImage*)

Model Fitting



The Likelihood Function

Under a given model, the **likelihood** is defined as:

The probability of a given dataset *y* as a function of the model parameters *θ*

$$L(\theta) = p(y \mid \theta)$$

The likelihood encodes information about the **forward model**.

The Likelihood Function

Under a given model, the **likelihood** is defined as:

The probability of a given dataset *y* as a function of the model parameters *θ*

$$L(\theta) = p(y \mid \theta)$$

The likelihood encodes information about the **forward model**.

Note:

In practice one typically uses the logarithm of the likelihood

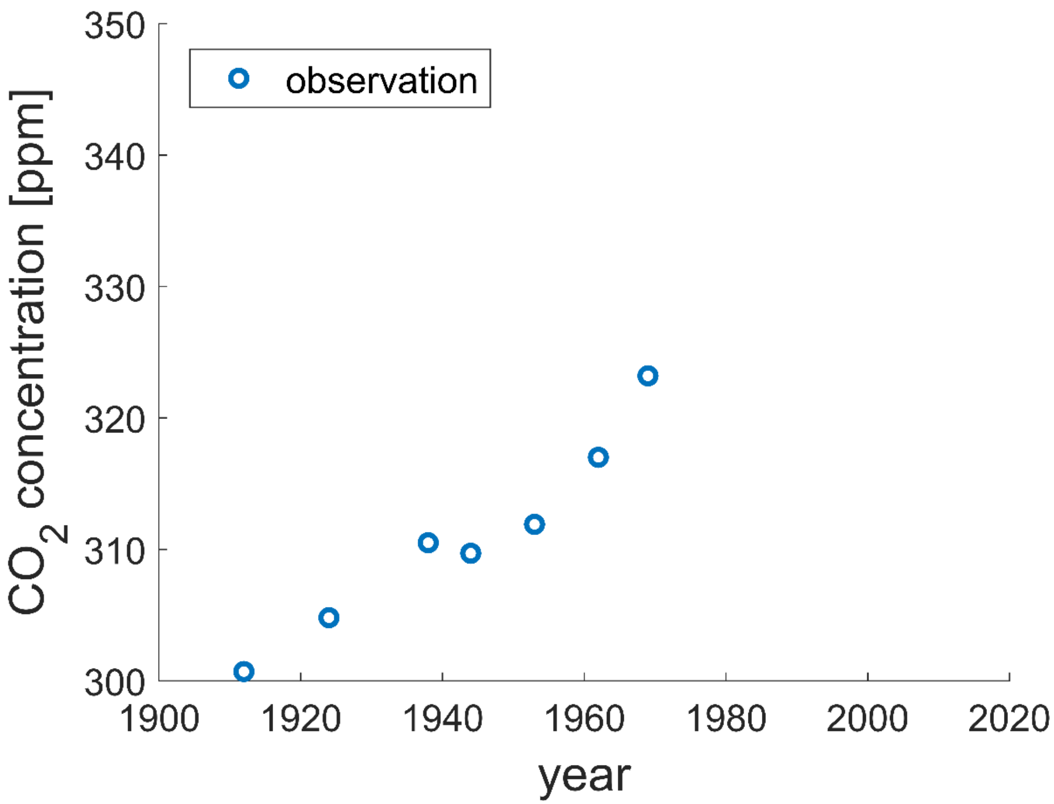
$$llh(\theta) = \log p(y \mid \theta)$$

The likelihood is a function of *θ* and is un-normalized.

$$\int L(\theta) d\theta \neq 1$$

Example

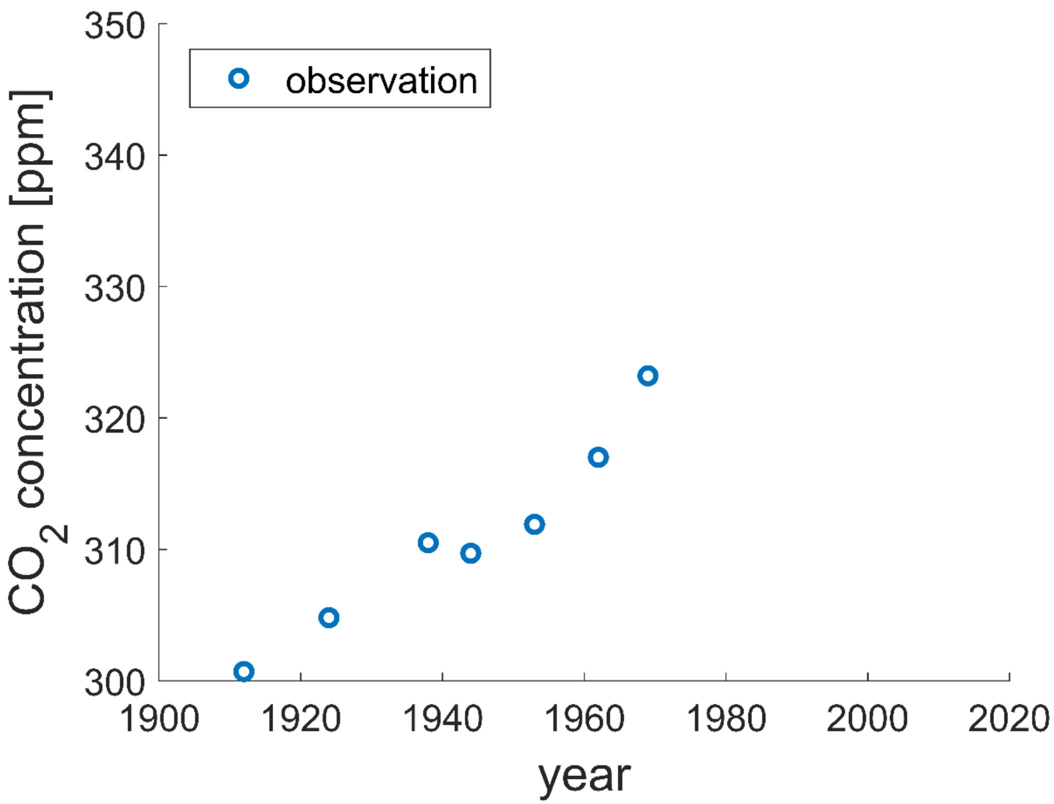
Data: Atmospheric CO_2 concentration



Data from: Etheridge et al. (1998)

Example

Data: Atmospheric CO_2 concentration



Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0, 1)$$

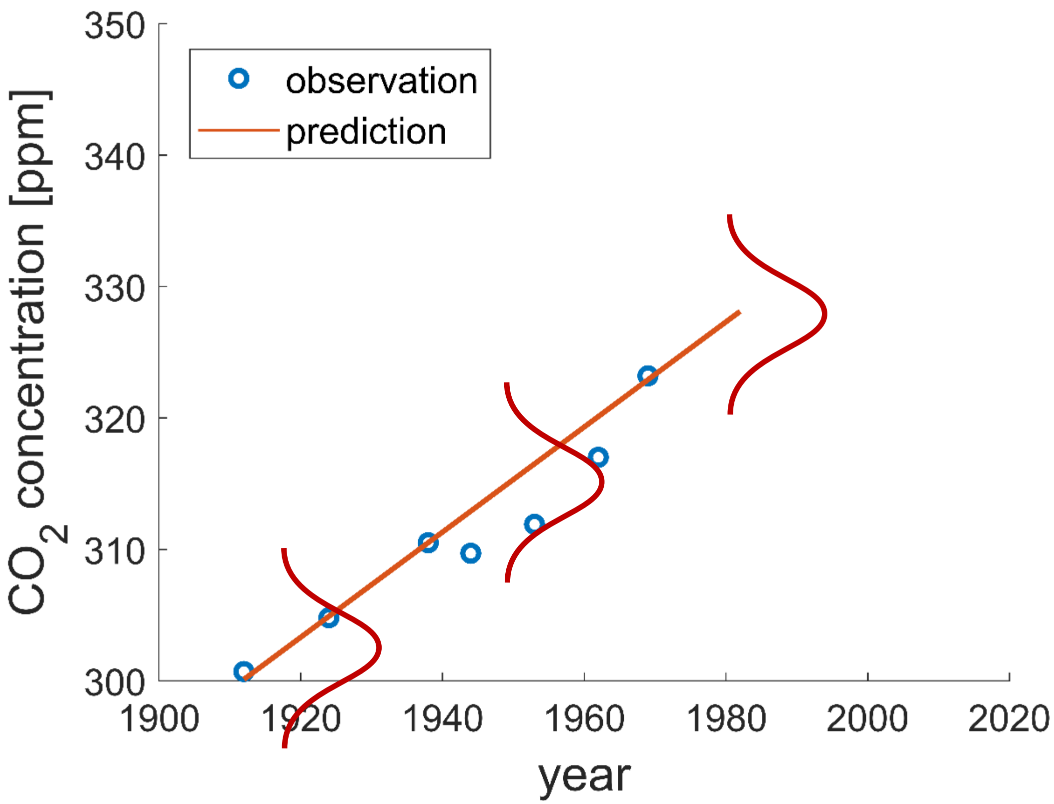
Likelihood:

$$\begin{aligned} L(\theta) &= p(y \mid \theta) \\ &= N(y \mid \theta_1 t + \theta_0, 1) \end{aligned}$$

Data from: Etheridge et al. (1998)

Example

Data: Atmospheric CO_2 concentration



Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0, 1)$$

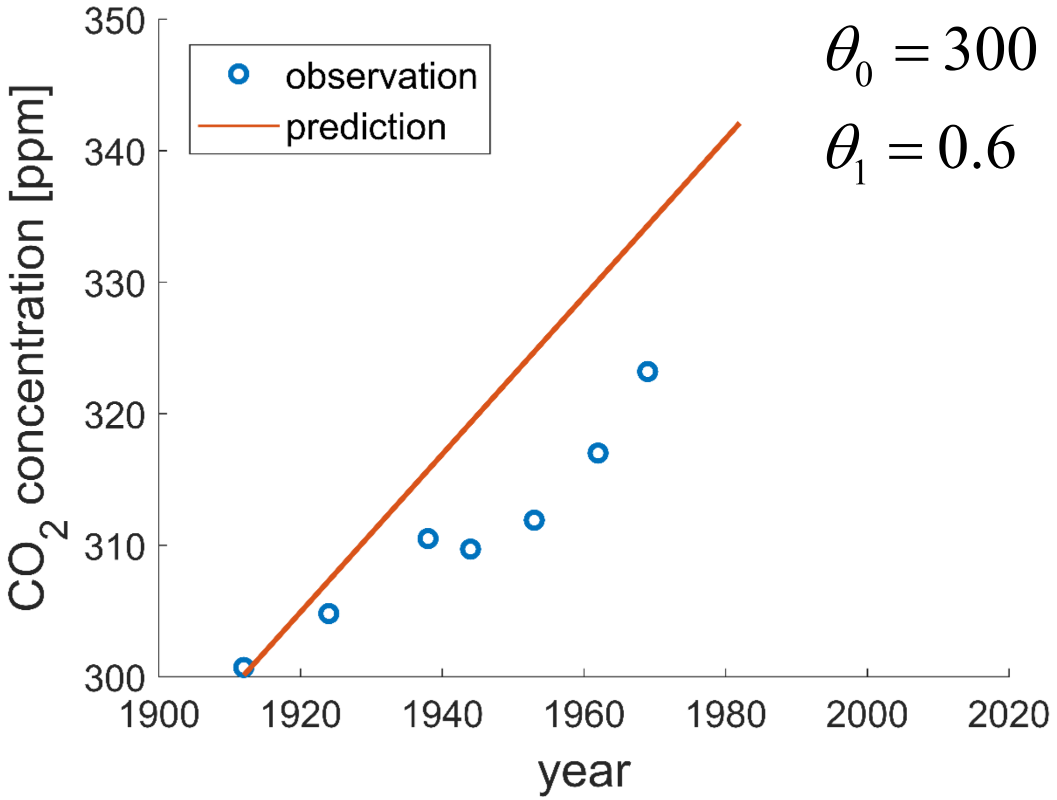
Likelihood:

$$L(\theta) = p(y \mid \theta)$$
$$= N(y \mid \theta_1 t + \theta_0, 1)$$

Data from: Etheridge et al. (1998)

Example

Data: Atmospheric CO_2 concentration

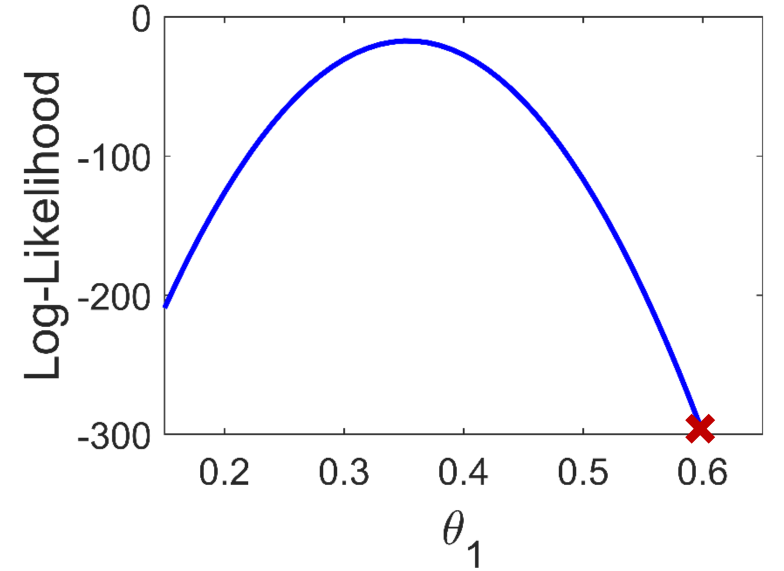


$$\theta_0 = 300$$
$$\theta_1 = 0.6$$

Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$
$$p(e) = N(e \mid 0,1)$$

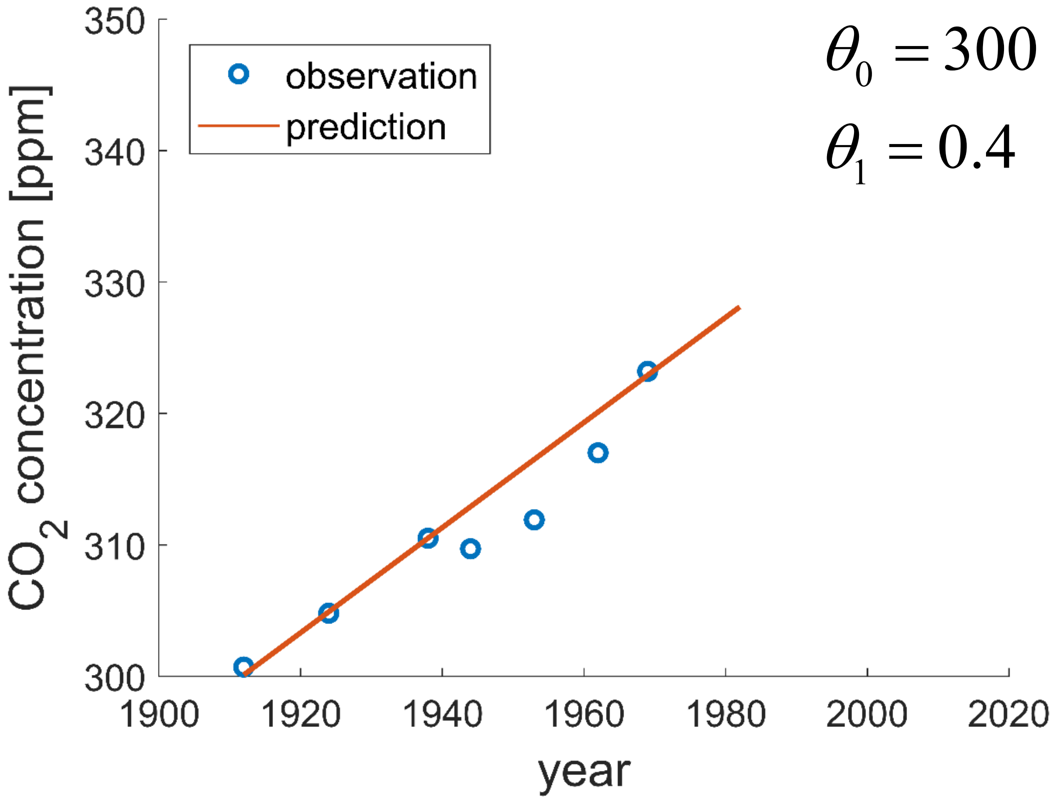
Likelihood:



Data from: Etheridge et al. (1998)

Example

Data: Atmospheric CO_2 concentration

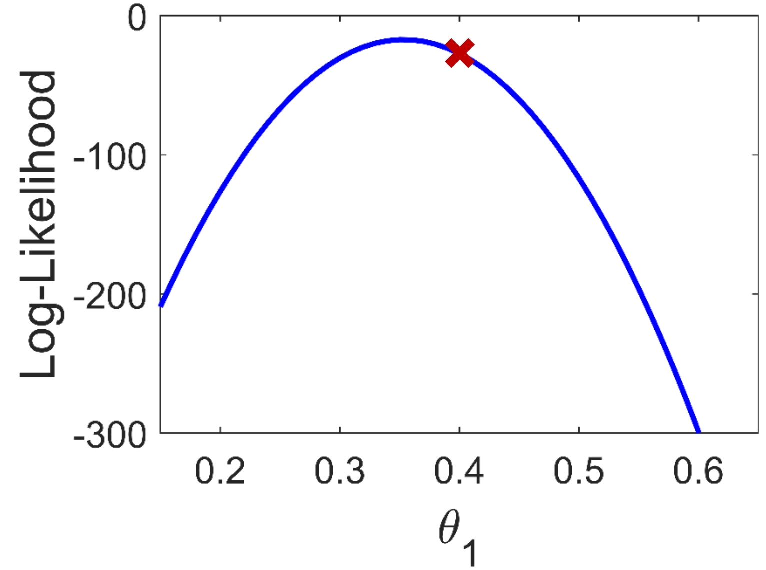


$$\theta_0 = 300$$
$$\theta_1 = 0.4$$

Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$
$$p(e) = N(e \mid 0,1)$$

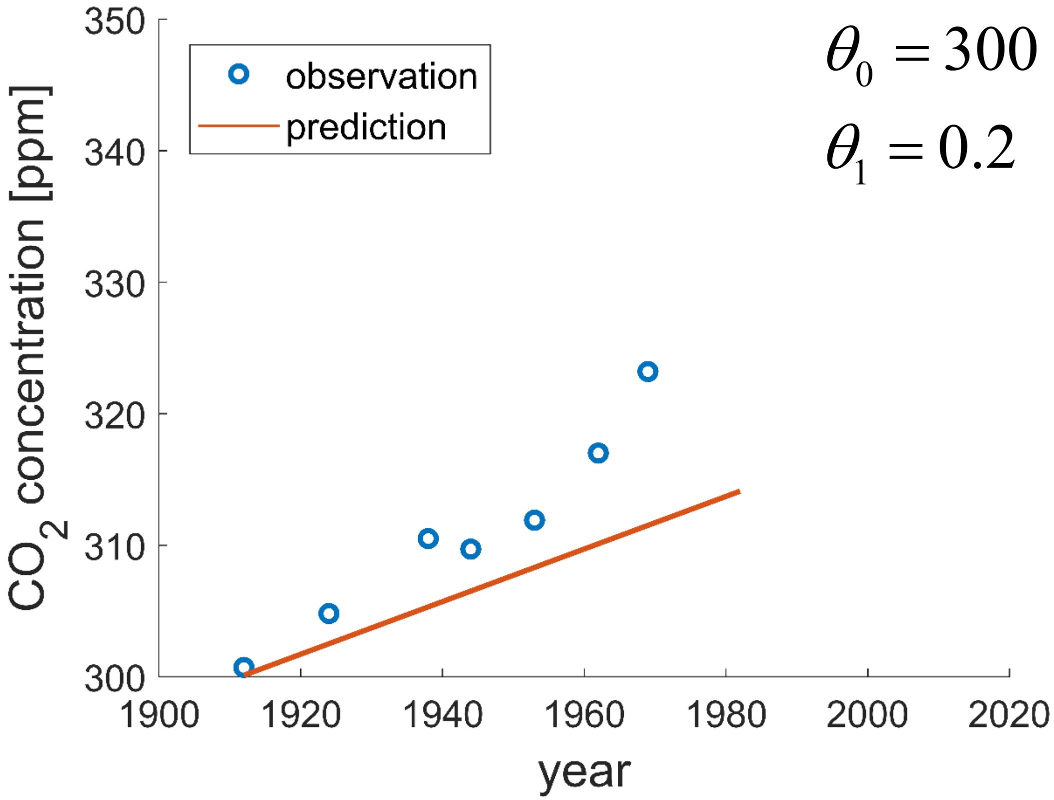
Likelihood:



Data from: Etheridge et al. (1998)

Example

Data: Atmospheric CO_2 concentration

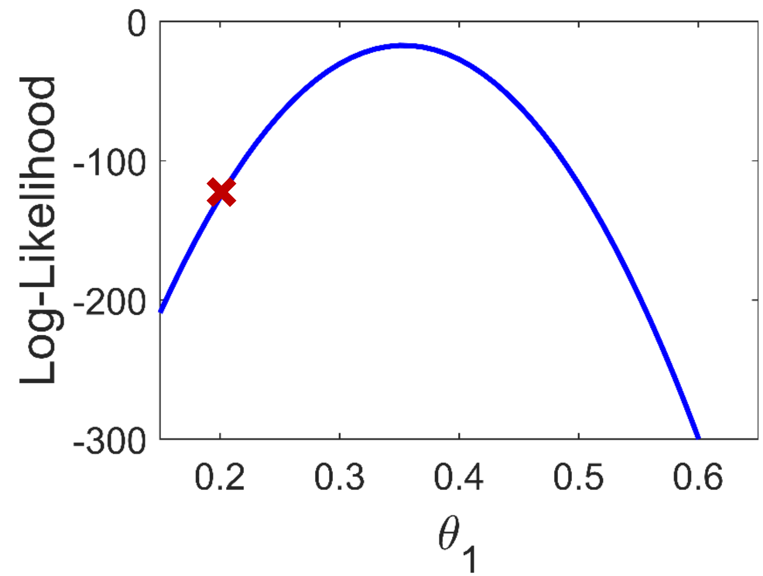


$$\theta_0 = 300$$
$$\theta_1 = 0.2$$

Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$
$$p(e) = N(e \mid 0,1)$$

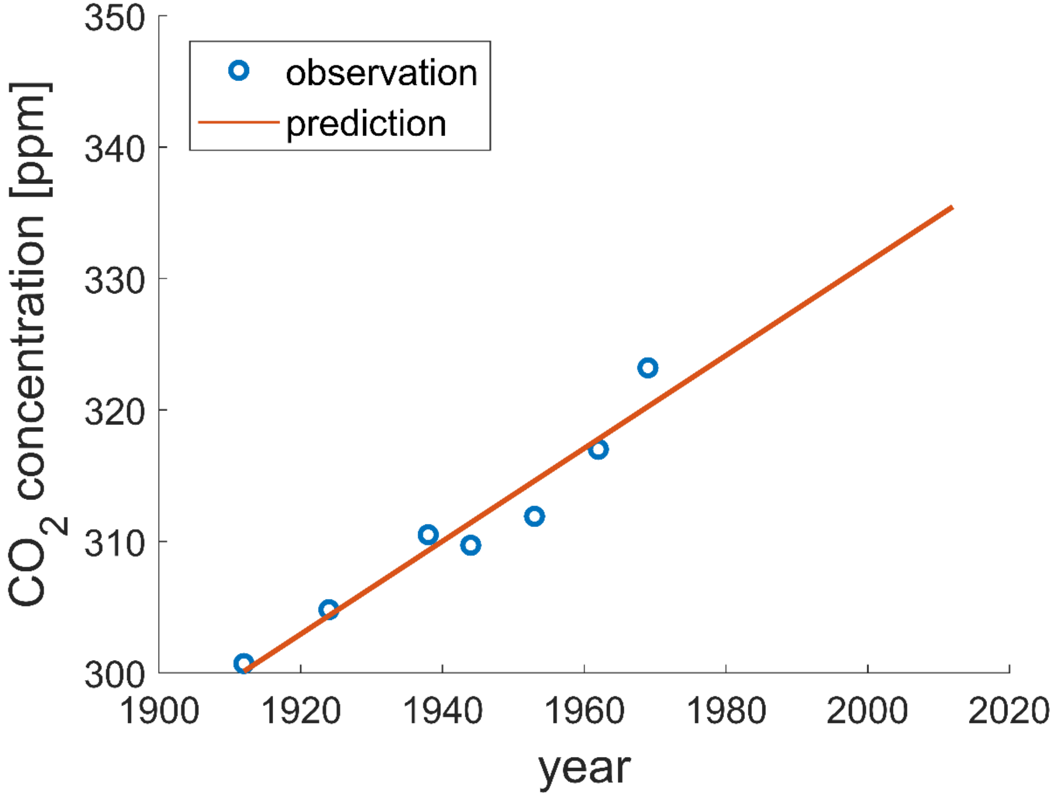
Likelihood:



Data from: Etheridge et al. (1998)

Maximum Likelihood

Data: Atmospheric CO_2 concentration

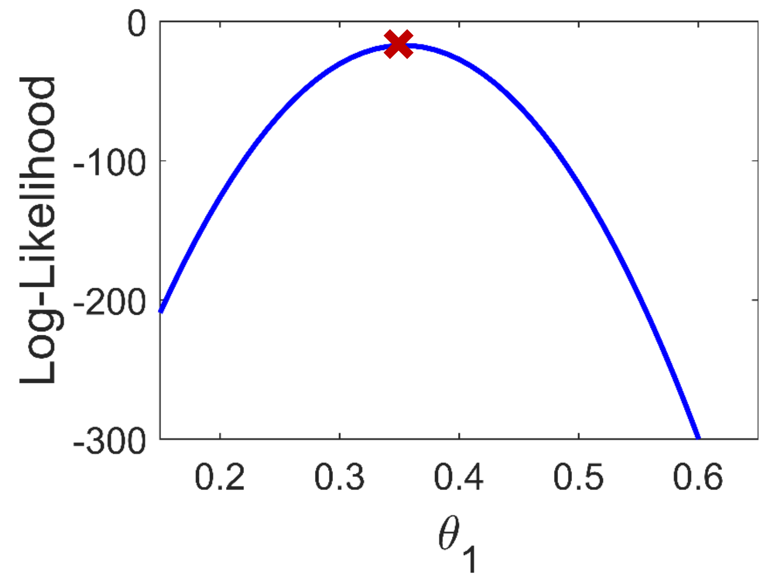


Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0, 1)$$

Likelihood:



Data from: Etheridge et al. (1998)

Maximum Likelihood (example Matlab code)

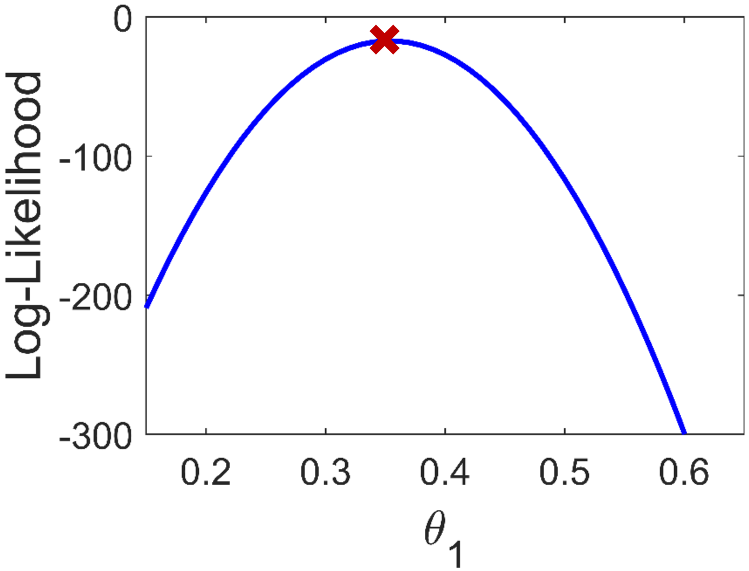
```
%% maximum likelihood estimation
function [ estimate ] = ml(y, x, P)

X = power(x,0:P);

% log-likelihood function
llh = @(y,X,sigma,theta) -sum((y-
X*theta).^2)/2/sigma^2-
numel(y)*log(2*pi*sigma^2)/2;

% inital esimate
est0 = [zeros(P+1,1);1];

% maximize log-likelihood with respect to
% model parameters (including the
% variance) using fminsearch
% (don't use fminsearch in practice)
estimate = fminsearch(@(est)llh(y, X,
est(1:P+1),est(end)), est0);

end
```

Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0,1)$$

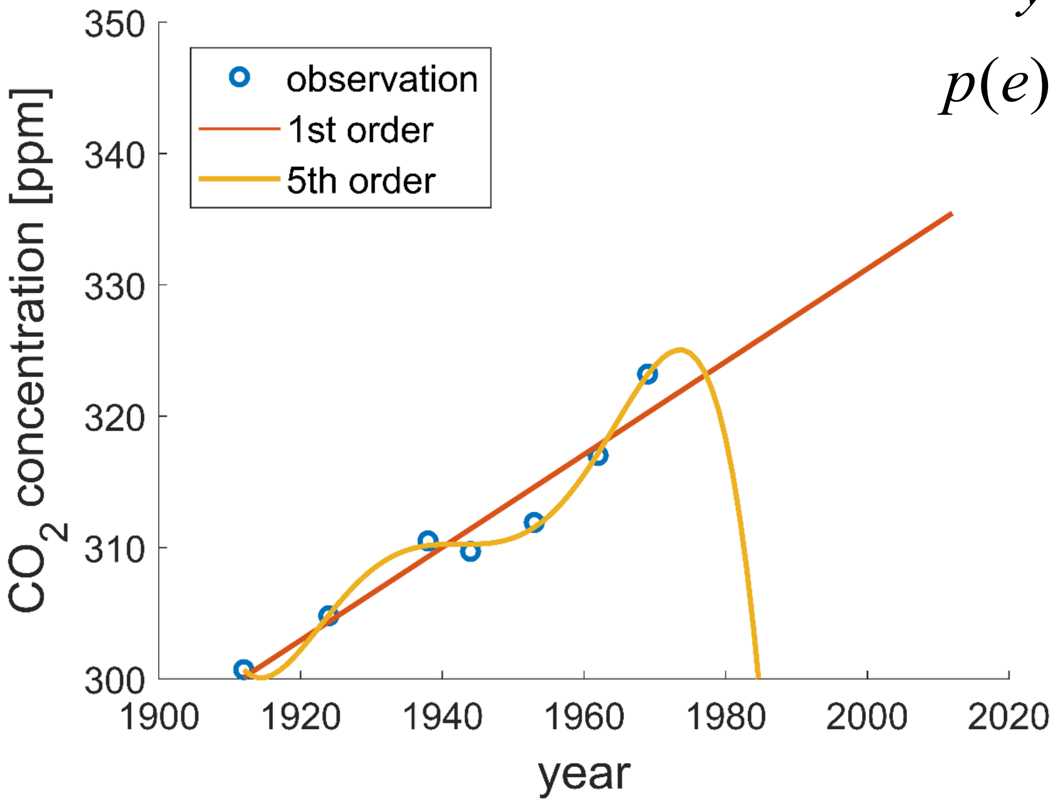
Likelihood:



Data from: Etheridge et al. (1998)

Overfitting

Data: Atmospheric CO_2 concentration

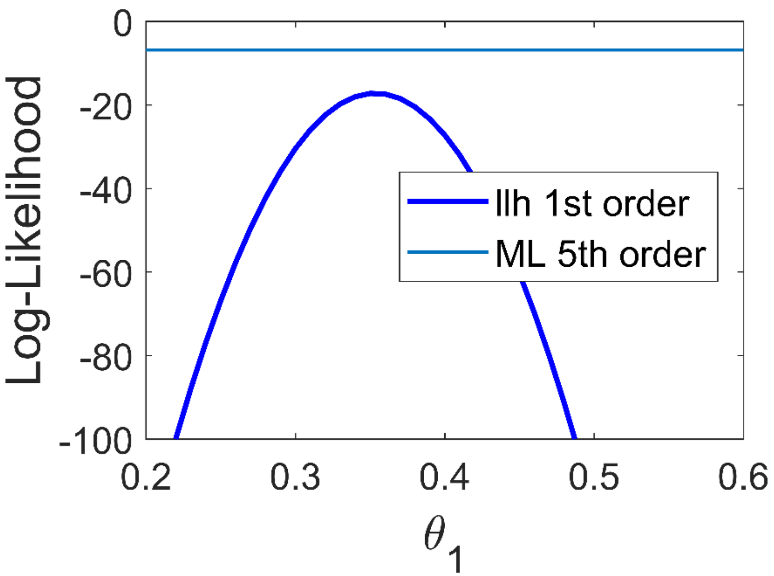


Model: 5^{th} order polynomial

$$y = \theta_5 t^5 + \theta_4 t^4 + \theta_3 t^3 + \theta_2 t^2 + \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0, 1)$$

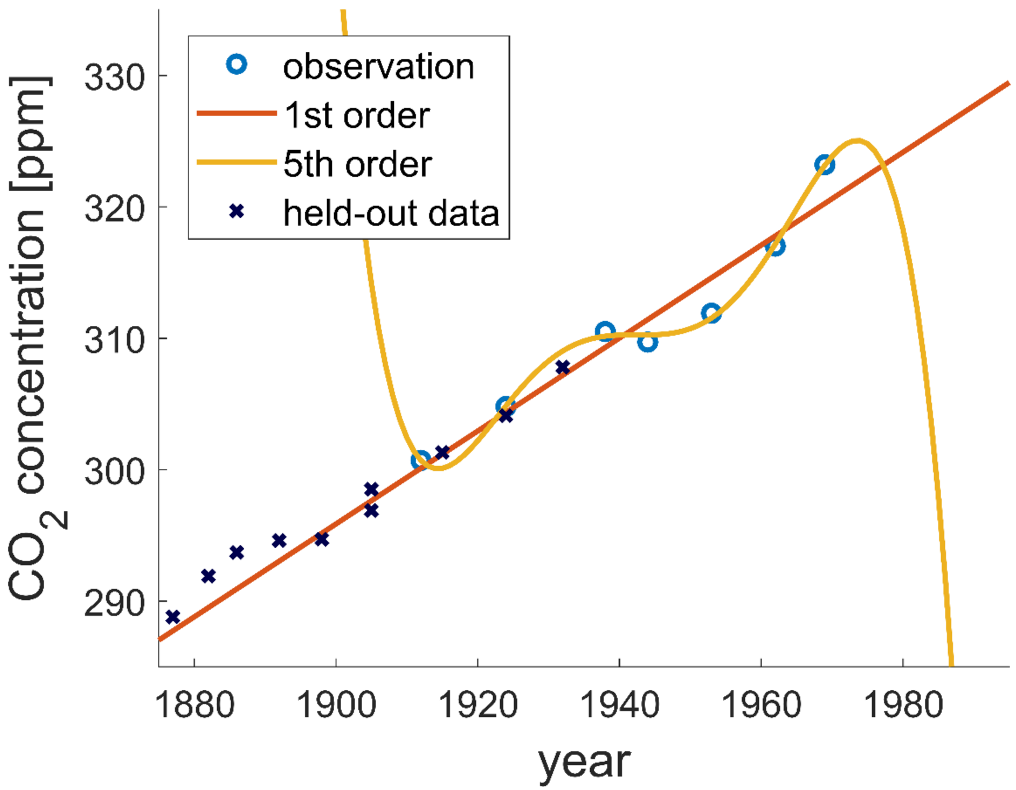
Likelihood:



Data from: Etheridge et al. (1998)

Held-out Data

Data: Atmospheric CO_2 concentration

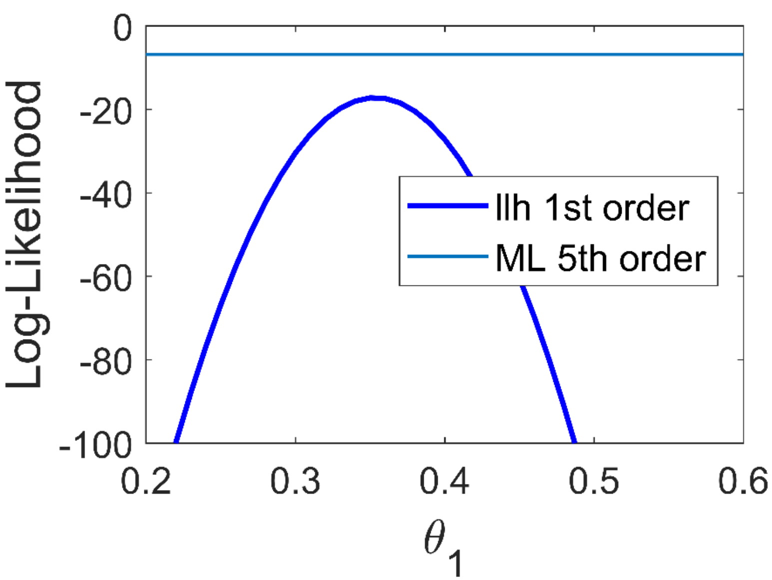


Log-likelihood on held-out data:

1st order: -18.7

5th order: -4.3x10^5

Likelihood:



Data from: Etheridge et al. (1998)

Further Reading

- Bishop: *Pattern Recognition and Machine Learning*
  - chapters 1 and 2, appendix B
- MacKay: *Information Theory, Inference, and Learning Algorithms*
  - *pages: 3 – 64, chapter 23*
  - *http://www.inference.org.uk/itprnn/book.pdf*
- Gelman: *Bayesian Data Analysis*
  - *appendix A*

Thank you

Many thanks to Klaas E. Stephan for the introductory slide!