Friedrich-Alexander-Universität
Technische Fakultät

Machine Learning
Data Analytics

Machine Learning for Time Series (MLTS)

Lecture 3: Bayesian Inference and Gaussian Processes

Dr. Dario Zanca

Machine Learning and Data Analytics (MaD) Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg

31.10.2024

Topics overview

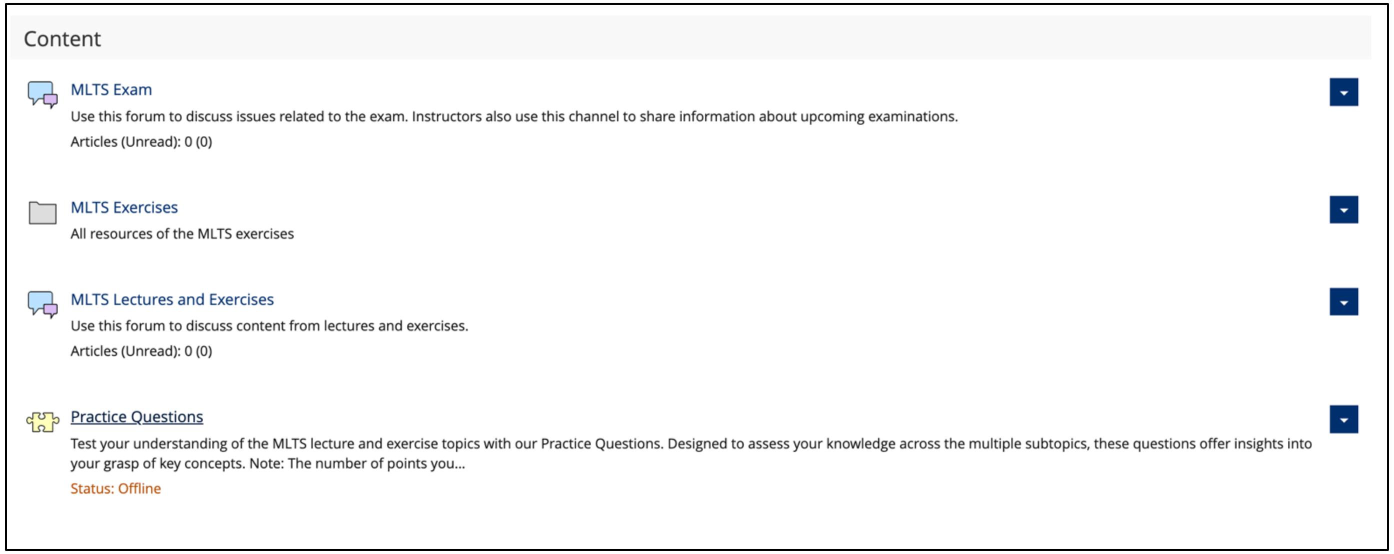
1. Time series fundamentals and definitions (Part 1)
2. Time series fundamentals and definitions (Part 2)
3. Bayesian Inference and Gaussian Processes
4. State space models (Kalman Filters)
5. State space models (Particle Filters)
6. Autoregressive models
7. Data mining on time series
8. Deep Learning (DL) for Time Series (Introduction to DL)
9. DL – Convolutional models (CNNs)
10. DL – Recurrent models (RNNs and LSTMs)
11. DL – Attention-based models (Transformers)
12. DL – From BERT to ChatGPT
13. DL – New Trends in Time Series processing
14. Time series in the real world

Topics overview

1. Time series fundamentals and definitions (Part 1)
2. Time series fundamentals and definitions (Part 2)
3. Bayesian Inference and Gaussian Processes
4. State space models (Kalman Filters)
5. State space models (Particle Filters)
6. Autoregressive models
7. Data mining on time series
8. Deep Learning (DL) for Time Series (Introduction to DL)
9. DL – Convolutional models (CNNs)
10. DL – Recurrent models (RNNs and LSTMs)
11. DL – Attention-based models (Transformers)
12. DL – From BERT to ChatGPT
13. DL – New Trends in Time Series processing
14. Time series in the real world

Course StudON

StudOn 2024-2025: https://www.studon.fau.de/crs5911979.html



Content

MLTS Exam

Use this forum to discuss issues related to the exam. Instructors also use this channel to share information about upcoming examinations.

Articles (Unread): 0 (0)

MLTS Exercises

All resources of the MLTS exercises

MLTS Lectures and Exercises

Use this forum to discuss content from lectures and exercises.

Articles (Unread): 0 (0)

Practice Questions

Test your understanding of the MLTS lecture and exercise topics with our Practice Questions. Designed to assess your knowledge across the multiple subtopics, these questions offer insights into your grasp of key concepts. Note: The number of points you...

Status: Offline

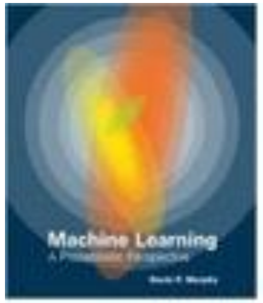
If you are entitled to take or re-take the exam, you can request access to this year's material at this "MATERIALS ONLY" StudON group:

https://www.studon.fau.de/crs6083795_join.html

References

Machine learning: A Probabilistic Perspective,

by Kevin Murphy (2012)



Machine Learning

A Probabilistic Perspective

Kevin P. Murphy

In this lecture…

1. *Bayes‘ Theorem*
2. *Bayesian model selection*
3. *Prior distributions*
4. *Linear regression (Bayesian treatment)*
5. *Gaussian Process*
6. *Recap*

Bayesian Inference

Bayes’ Theorem

Bayes’ Theorem

Formulation

The **Bayes’ Theorem** was formulated by the English philosopher **Thomas Bayes** (1701 – 1761), whose notes were edited and published posthumously by Richard Price.

Bayes' theorem is stated mathematically as the following equation:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}$$

where A and B are events, and $P(B) \neq 0$.

Portrait from: Terence O'Donnell, *History of Life Insurance in Its Formative Years* (Chicago: American Conservation Co:, 1936), p. 335

Bayes’ Theorem

Formulation

Posterior probability

The probability of event A occurring given that B is true

Likelihood

The probability of event B occurring given that A is true

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Marginal probability

The probability of observing B without any given conditions

Prior probability

The probability of observing A without any given conditions

Portrait from: Terence O'Donnell, *History of Life Insurance in Its Formative Years* (Chicago: American Conservation Co:, 1936), p. 335

Bayes’ Theorem

An example. Iterative application of the Bayes’ theorem.

We know that:

- Disease chance: 1%
- Test accuracy: 95%

$\boldsymbol{A}$: Having the disease
$\boldsymbol{B}$: Testing positive to the disease

$\mathrm{P}(\boldsymbol{B}|\boldsymbol{A})$: Test sensitivity (Likelihood)
$\mathrm{P}(\boldsymbol{B})$: Prob. of a positive test (Marginal)
$\mathrm{P}(\boldsymbol{A})$: Disease chance (Prior)

First positive test

$$\mathrm{P}(A|B) = \frac{.95 \times 0.01}{0.95 \cdot 0.01 + (1-0.95)(1-0.01)} = 0.161$$

Second positive test

$$\mathrm{P}(A|B) = \frac{.95 \times 0.161}{0.95 \cdot 0.161 + (1-0.95)(1-0.161)} = 0.785$$

Third positive test

$$\mathrm{P}(A|B) = \frac{.95 \times 0.785}{0.95 \cdot 0.785 + (1-0.95)(1-0.785)} = 0.986$$

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\ \mathrm{P}(A)}{\mathrm{P}(B)}$$

Portrait from: Terence O'Donnell, *History of Life Insurance in Its Formative Years* (Chicago: American Conservation Co:, 1936), p. 335

Bayesian modelling

Example of a linear model

Let $\mathcal{D}$ denote the **observed data**,

$$\mathcal{D} \ = \left\{x^{(n)}, y^{(n)}\right\}$$

with $x^{(n)} \in \ \mathcal{R}$ represents the input, and $y^{(n)} \in \ \mathcal{R}$ represents the output (labels).

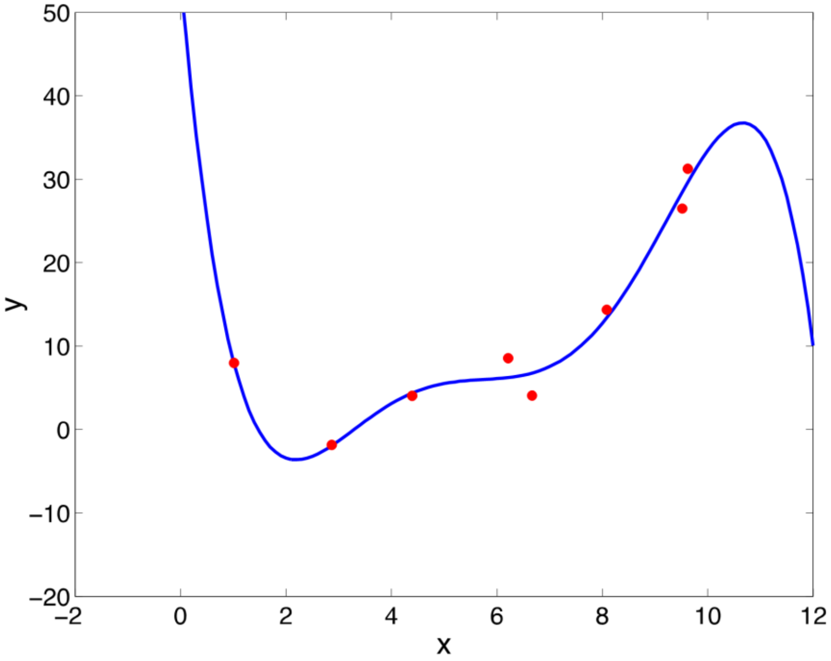
The **model** is defined as

$$y^{(n)} = \omega_0 + \omega_1 x^{(n)} + \omega_2 x^{(n)} \ldots + \omega_2 x^{(n)} + \epsilon$$

where data noise is Gaussian distributed, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

We denote with θ the **unknown parameters**,

$$\theta \ = (\omega_0, \ldots, \omega_m, \sigma)$$



Bayesian modelling

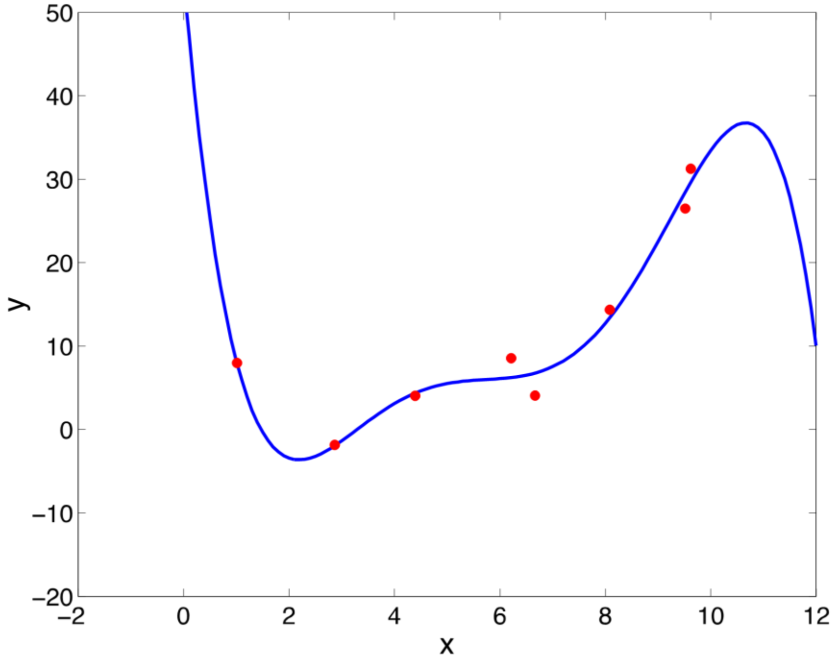
Example of a linear model

Data: $\mathcal{D} = \{x^{(n)}, y^{(n)}\}$

Model: $y^{(n)} = \omega_0 + \omega_1 x^{(n)} + \omega_2 x^{(n)} \dots + \omega_2 x^{(n)} + \epsilon$

Unknown parameters: $\theta = (\omega_0, \dots, \omega_m, \sigma)$

Goal: To infer $\boldsymbol{\theta}$ from the data and to predict future outputs $p(y|x, \boldsymbol{\theta}, \boldsymbol{\mathcal{D}})$



Bayesian modelling

$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta})$: prior probability of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$: posterior of $\boldsymbol{\theta}$ given $\boldsymbol{\mathcal{D}}$

$p(\boldsymbol{\mathcal{D}})$: marginal probability of $\boldsymbol{\mathcal{D}}$

$p(y|x, \boldsymbol{\mathcal{D}})$: predictive distribution

Bayesian modelling

$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta})$: prior probability of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$: posterior of $\boldsymbol{\theta}$**,** given $\boldsymbol{\mathcal{D}}$

$p(\boldsymbol{\mathcal{D}})$: marginal probability of $\boldsymbol{\mathcal{D}}$

$p(y|x, \boldsymbol{\mathcal{D}})$: predictive distribution

Bayes' rule:

$$p(\theta|\boldsymbol{\mathcal{D}}) = \frac{p(\boldsymbol{\mathcal{D}}|\theta)p(\theta)}{P(\boldsymbol{\mathcal{D}})}$$

Prediction of a new point:

$$p(y|x, \boldsymbol{\mathcal{D}}) = \int p(y|\theta, x, \boldsymbol{\mathcal{D}})p(\theta|\boldsymbol{\mathcal{D}})\, d\theta$$

Bayesian modelling

Observations

- In contrast to the maximum likelihood estimation (MLE), in Bayesian learning **we average over possible parameter settings** rather than optimizing over parameter space.
- Bayesian inference gives us a **systematic way to express our uncertainty** about future predictions. Prediction is not just a point estimate (as for MLE) but has a probability form that expresses the uncertainty about the predictions.

Bayesian Inference

Bayesian Model Selection

hould not be
multiplic
Frater

The principle of Occam’s Razor

The **principle of Occam’s razor** in its original formulation states that:

> "Entia non sunt multiplicanda praeter necessitatem"

Picture from: https://www.britannica.com/topic/Occams-razor

The principle of Occam’s Razor

The **principle of Occam’s razor** in its original formulation states that:

"Entia non sunt multiplicanda praeter necessitatem"

(In English, “Entities should not be multiplied unnecessarily”)

Many scientists have adopted or reformulated the Occam's Razor principle, which is often cited in stronger forms, as in the following statement:

- "If you have two theories that both explain the observed facts, then you should use the simplest until more evidence comes along"
- “One should pick the simplest model that adequately explains the data”

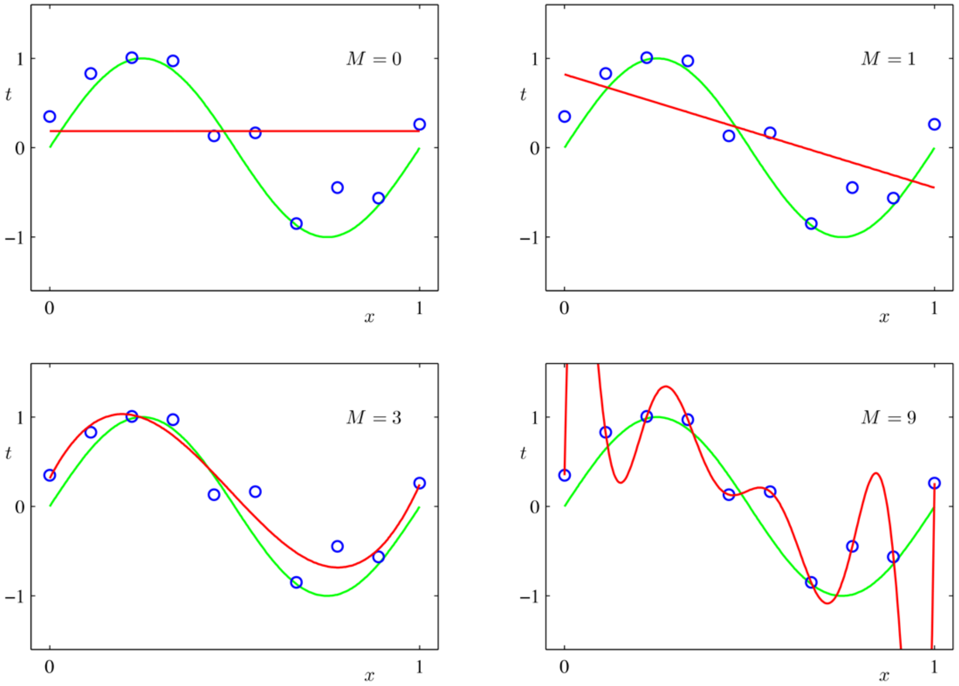
Picture from: https://www.britannica.com/topic/Occams-razor

The model Selection problem

We could perform K-fold cross-validation (CV) to estimate the generalization error of all candidates model.

However, it requires fitting each candidate model K times!

→ A more efficient approach is given by Bayesian modelling



Which of the above models represents data the best?

Occam’s razor and Bayesian model selection

We can compare different models using the marginal likelihood:

$$p(\boldsymbol{\mathcal{D}}|M_i) = \int p(\boldsymbol{\mathcal{D}}|\theta, M_i)p(\theta|M_i)\, d\theta$$

- ➢ Model classes that are **too simple** are unlikely to generate the data set.
- ➢ Model classes that are **too complex** can generate many possible data sets, so again, they are unlikely to generate that particular data set $\boldsymbol{\mathcal{D}}$.

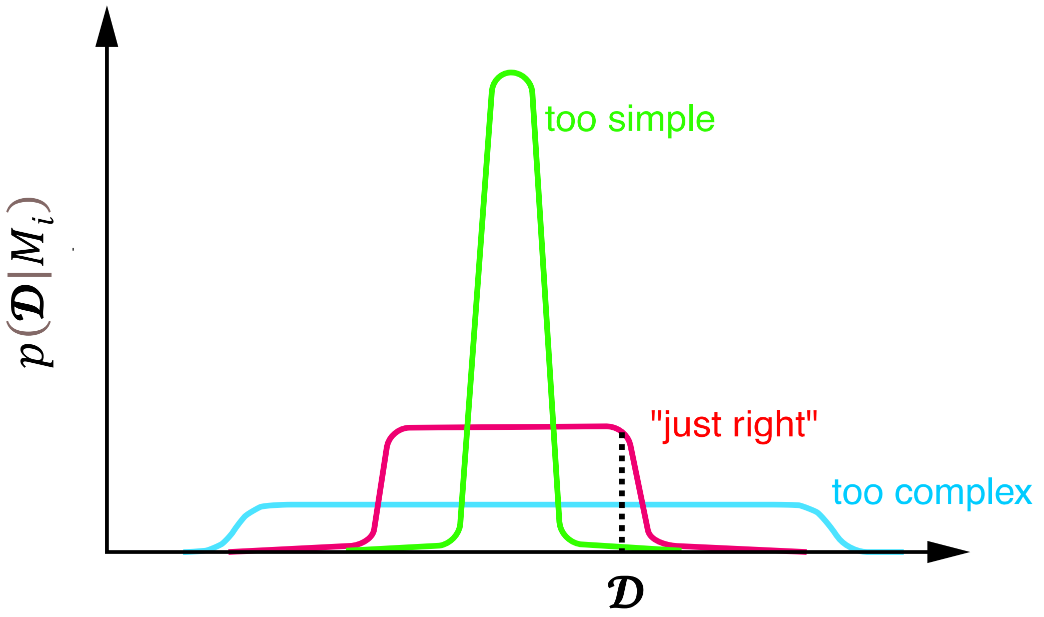


Image from: Rasmussen, C., & Ghahramani, Z. (2000). Occam's razor. Advances in neural information processing systems, 13.

Occam’s razor and Bayesian model selection

To understand the Bayesian Occam’s razor, we notice that:

$$\int_{\mathcal{D}} p(\mathcal{D}|M_i) = 1$$

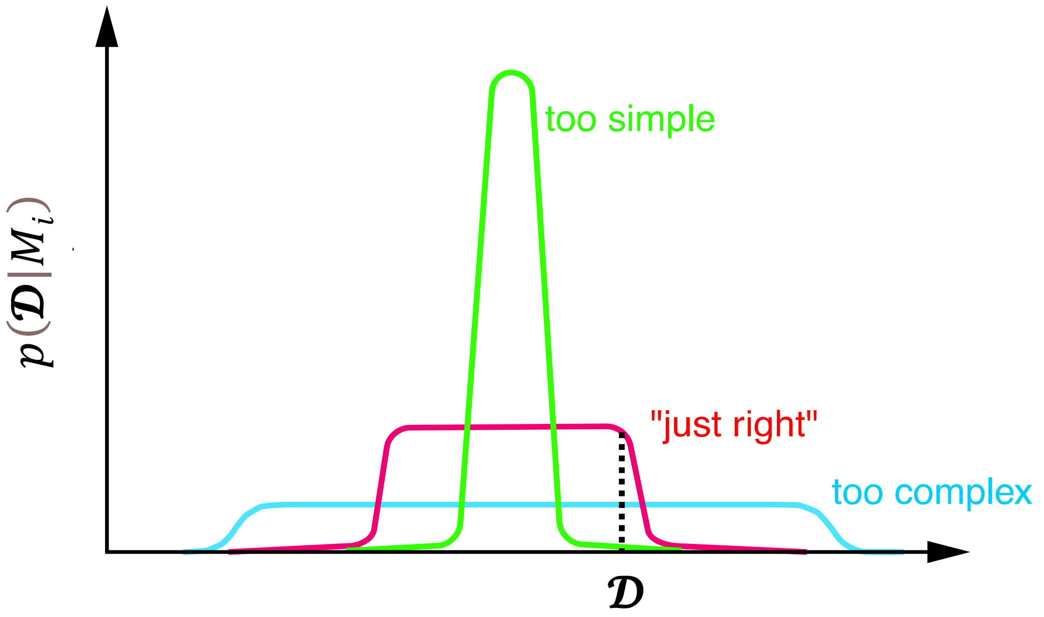


Image from: Rasmussen, C., & Ghahramani, Z. (2000). Occam's razor. Advances in neural information processing systems, 13.

Occam's razor and Bayesian model selection

To understand the Bayesian Occam's razor, we notice that:

$$\int_{\mathcal{D}} p(\mathcal{D}|M_i) = 1$$

Intuitively, complex models which can predict *many* datasets, must spread their probability mass
→ They don't attribute large probability for any given data set as simpler models.

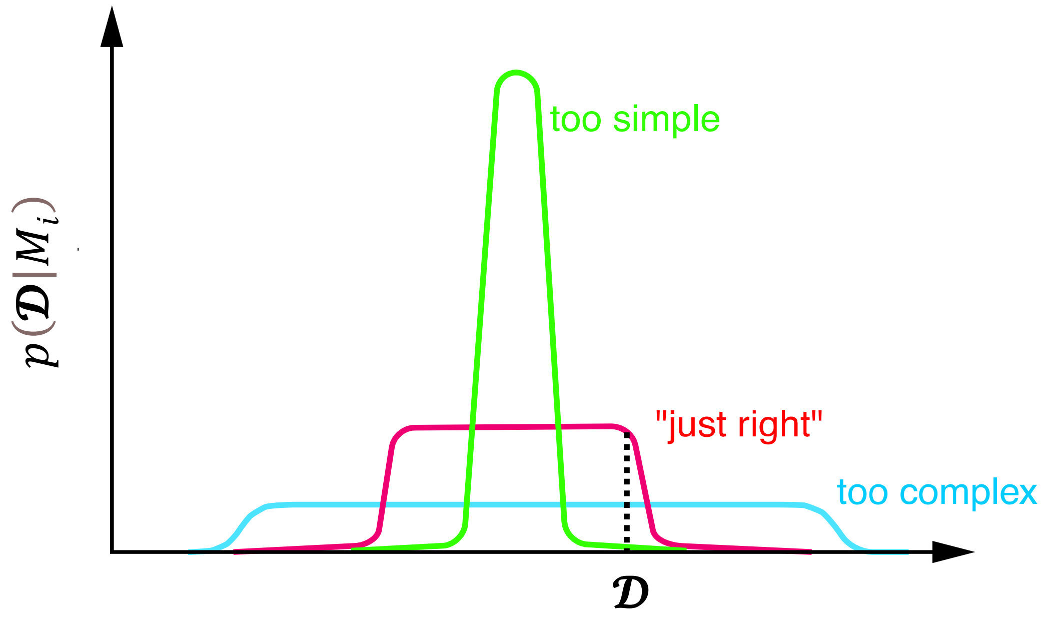


Image from: Rasmussen, C., & Ghahramani, Z. (2000). Occam's razor. Advances in neural information processing systems, 13.

Bayesian Occam’s razor

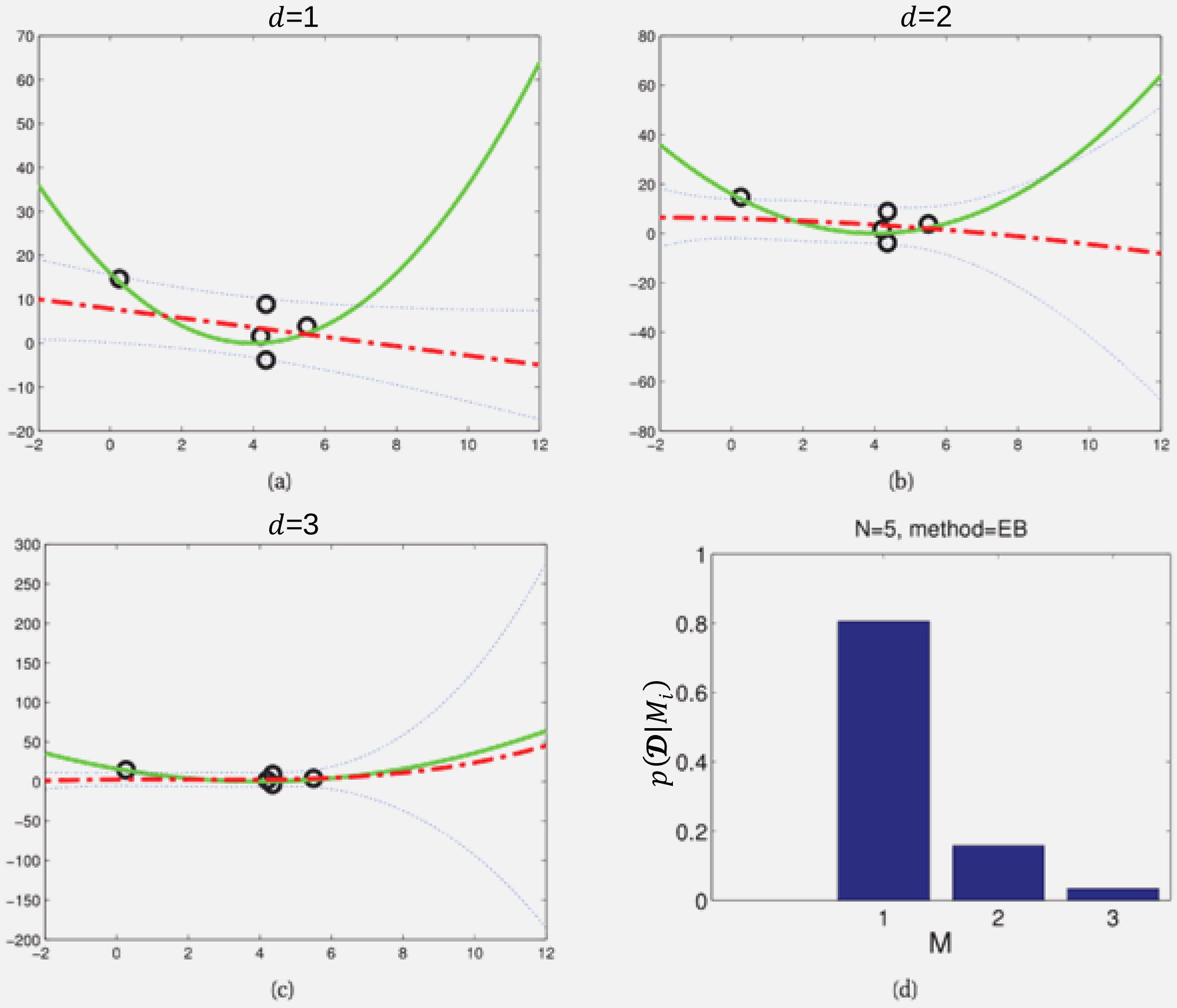
A concrete example

We plot polynomials of degrees 1, 2 and 3 fit to N=5 data points using (empirical) Bayes.

━ True function

- · - Prediction

⋯ $\pm\sigma$ around the mean



There is not enough data to justify a complex model, so the best model is d = 1.

Bayesian Occam’s razor

A concrete example

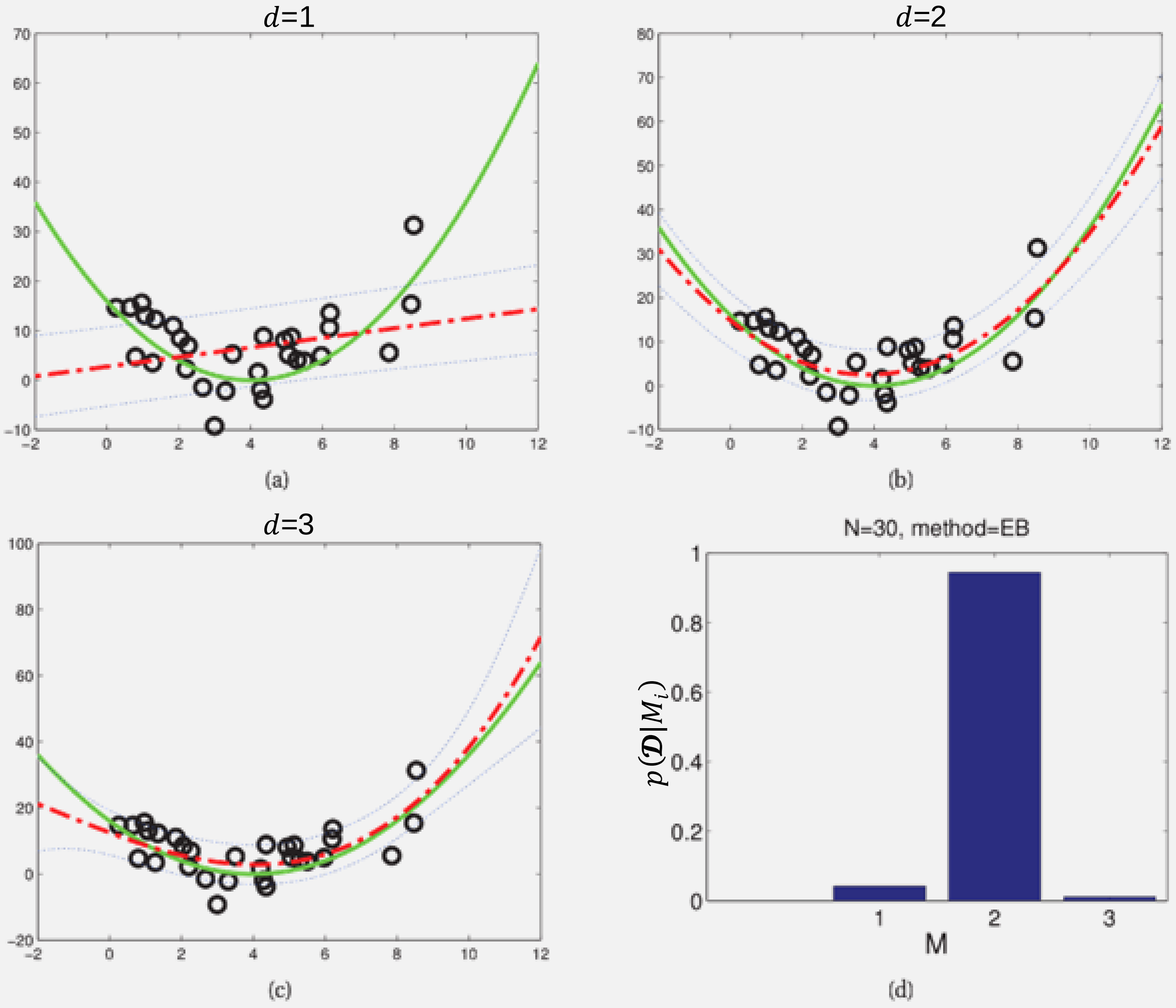
We plot polynomials of degrees 1, 2 and 3 fit to N=30 data points using (empirical) Bayes.

— True function

- · - Prediction

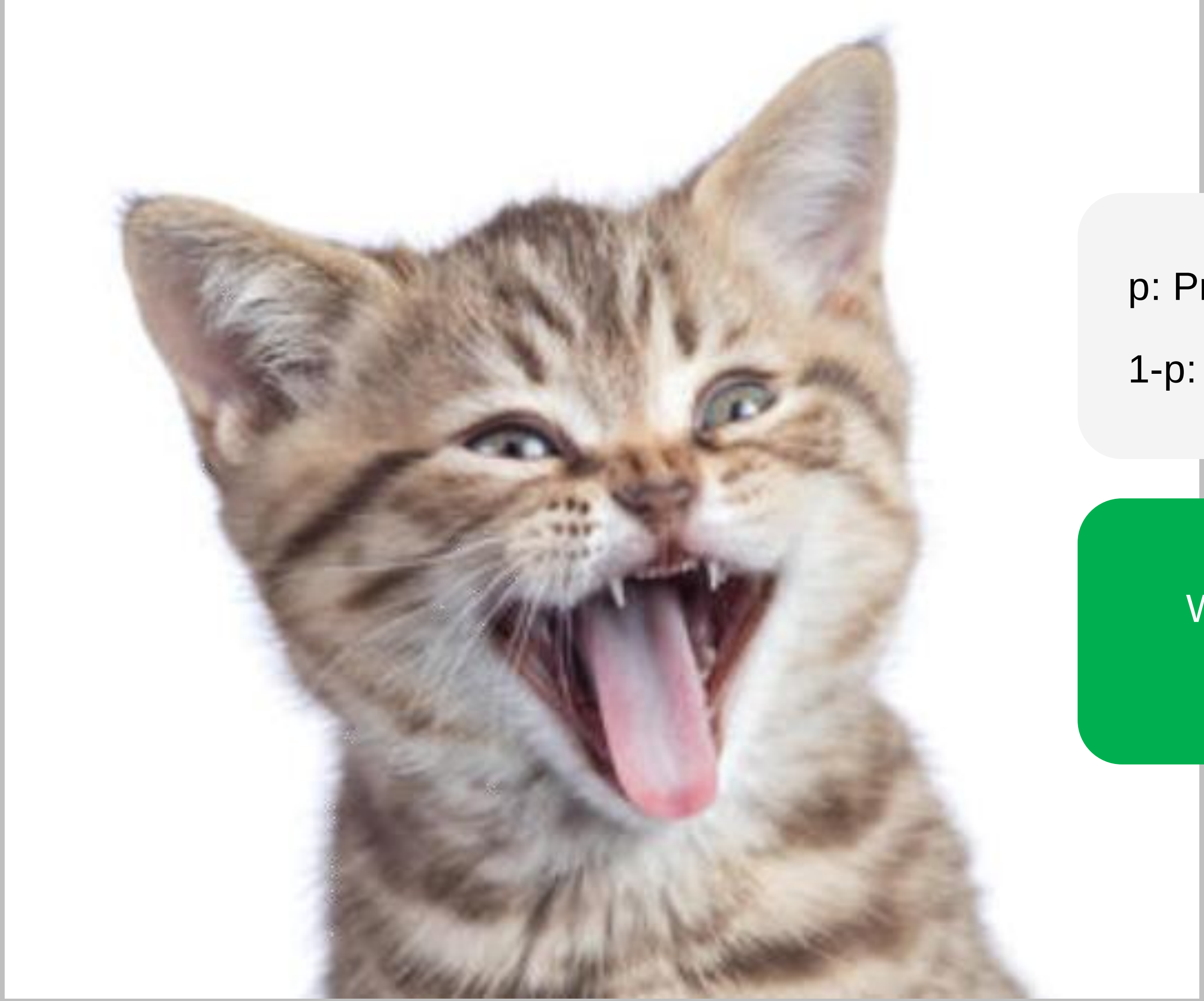
⋯ $\pm\sigma$ around the mean

When more data is available, d = 2 is the right model



Bayesian Inference

Prior Distributions



p: Prob. purring

1-p: Prob. grumpy

What is the best guess for the probability p?

How can I update my belief on p?

Prior distributions

The importance of priors in Bayesian Inference

$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta})$: prior probability of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$: posterior of $\boldsymbol{\theta}$**,** given $\boldsymbol{\mathcal{D}}$

$p(\boldsymbol{\mathcal{D}})$: marginal probability of $\boldsymbol{\mathcal{D}}$

$p(y|x, \boldsymbol{\mathcal{D}})$: predictive distribution

Bayes' rule:

$$p(\theta|\boldsymbol{\mathcal{D}}) = \frac{p(\boldsymbol{\mathcal{D}}|\theta)\ \ p(\theta)}{P(\boldsymbol{\mathcal{D}})}$$

Prior distributions

The importance of priors in Bayesian Inference

$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta})$: prior probability of $\boldsymbol{\theta}$

$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$: posterior of $\boldsymbol{\theta}$**,** given $\boldsymbol{\mathcal{D}}$

$p(\boldsymbol{\mathcal{D}})$: marginal probability of $\boldsymbol{\mathcal{D}}$

$p(y|x, \boldsymbol{\mathcal{D}})$: predictive distribution

Bayes' rule:

$$p(\theta|\boldsymbol{\mathcal{D}}) = \frac{p(\boldsymbol{\mathcal{D}}|\theta)\ \boxed{p(\theta)}}{P(\boldsymbol{\mathcal{D}})}$$

A **prior probability distribution** of an uncertain quantity is the probability distribution that would express one's belief, before some evidence is taken into account.

→ For example, a prior could represent the distribution of votes coming from an opinion poll, prior to the election.

Priors: Subjective vs. Objective

A **subjective prior** expresses the modeler's subjective belief.

- ➢ We formulate our (subjective) assumptions about modeling the data in terms of priors
- ➢ We have to work hard to understand the system under study in order to formulate our assumptions

An **objective prior** constrain prior beliefs to be "uninformative" about the parameters.

- ➢ The objective Bayes view is that formulating our assumptions is too difficult, especially in complex models

If we don't have strong beliefs about what θ should be, it is common to use an "uninformative" priors → **"Let the data speak for itself!"**

Priors: Informative vs. Uninformative

An **informative prior** expresses a specific information about a variable.

- ➢ For example, a reasonable informative prior about the temperature at noon tomorrow could be given by a normal distribution with expeced value equal to today’s noon temperature and variance equal to the daily variance of the temperature.

An **uninformative prior** is designed to express vague or general information about a variable.

- ➢ For example, when tossing a coin, we assign the probability of 0.5 to both heads and tails.

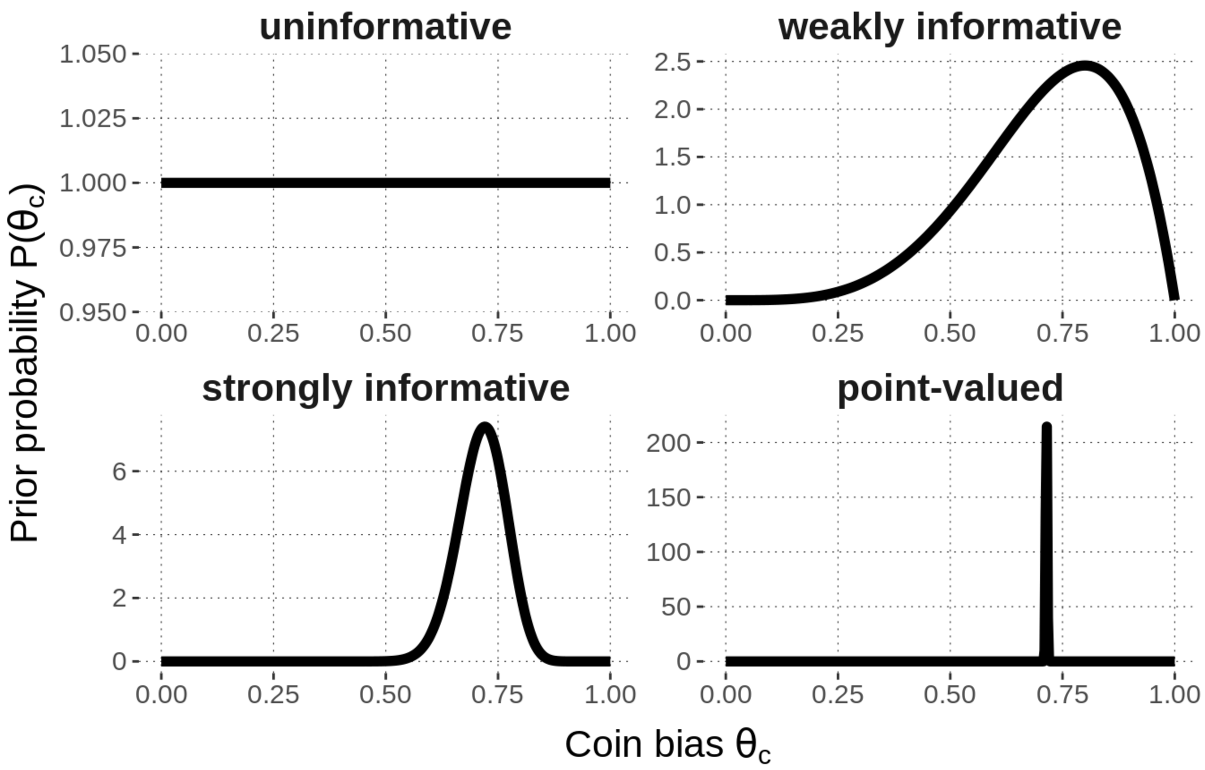


Image from: https://michael-franke.github.io/intro-data-analysis/Chap-03-03-models-parameters-priors.html

Conjugate Prior

A computational “shortcut”

A prior $p(\theta)$ is a **conjugate prior** for a particular likelihood $p(y|\theta)$ if the resulting posterior $p(\theta|y)$ has the same algebraic form.

Conjugate priors are widely used because they provide advantages:

- they usually allow us to derive a closed-form expression for the posterior distribution;
- they are easy to interpret,

Note: Conjugate priors simplify the computation, but are often not flexible enough to encode our prior knowledge → We can also use mixture of conjugate priors.

Conjugate Prior

An example

Likelihood (Binomial): $p(\mathcal{D}|\boldsymbol{\theta}) = Bin(\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$

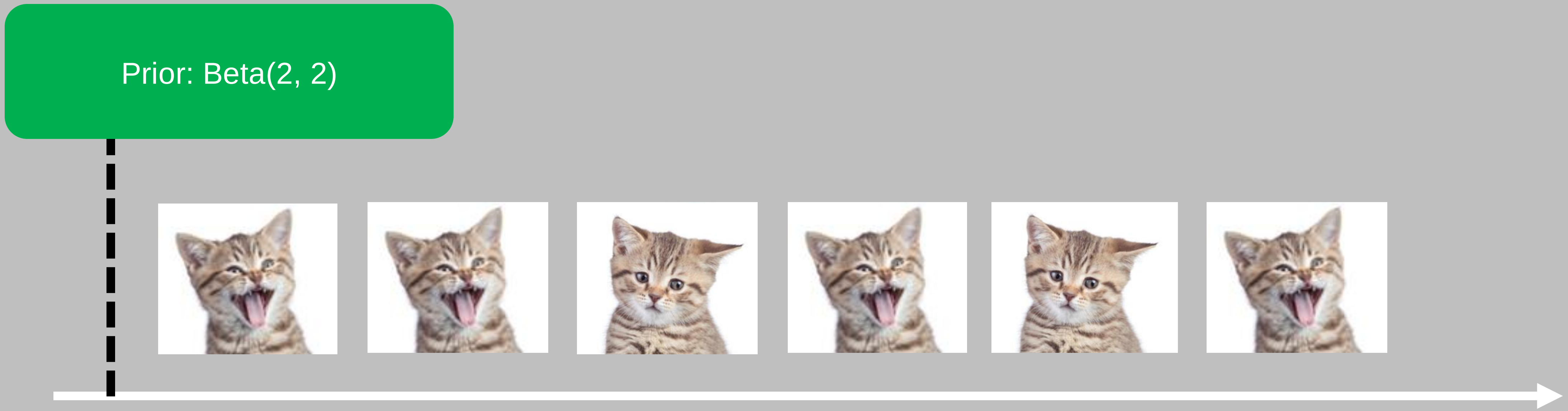
Prior (Beta): $p(\boldsymbol{\theta}) = Beta(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, where $B(\alpha,\beta) = \int t^{\alpha-1}\,(1-t)^{\beta-1}\quad dt$

We plug them into the Bayes‘ formula to derive the posterior distribution:

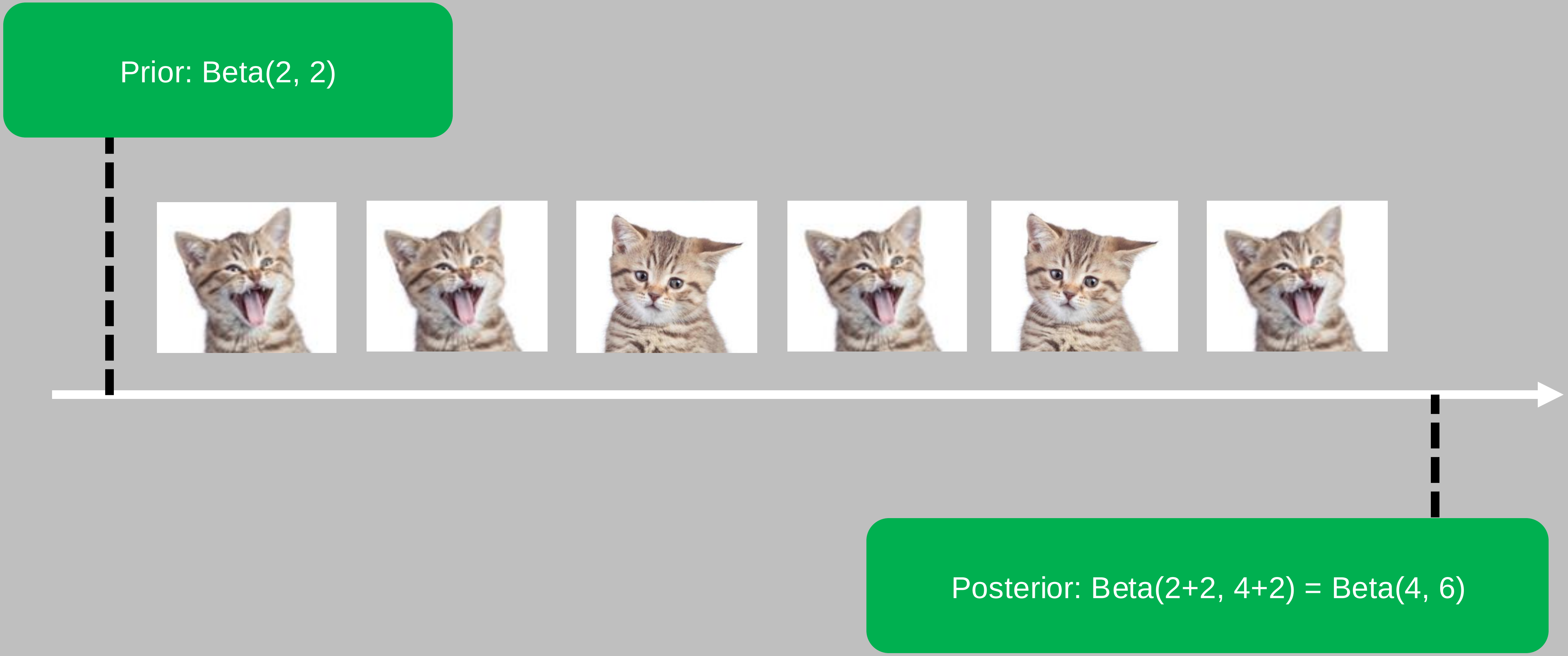
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\ p(\theta)}{\int p(\mathcal{D}|\theta)\ \ p(\theta)d\theta}$$

$$= \frac{\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\ \ \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\ \ \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\ d\theta}$$

$$= \frac{\theta^{x+\alpha-1}\,(1-\theta)^{n-x+\beta-1}}{\int\theta^{x+\alpha-1}\,(1-\theta)^{n-x+\beta-1}\quad d\theta} = Beta(x+\alpha, n-x+\beta)$$

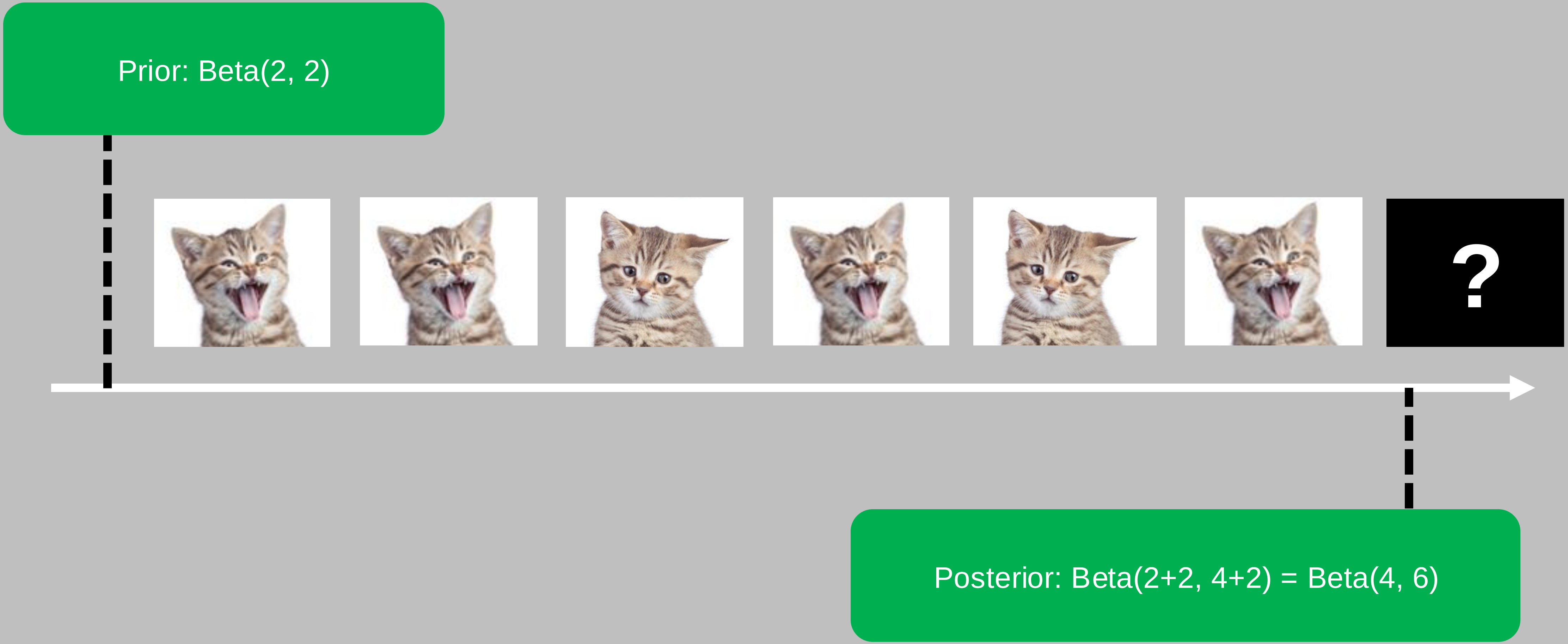


Prior: Beta(2, 2)



Prior: Beta(2, 2)

Posterior: Beta(2+2, 4+2) = Beta(4, 6)



Prior: Beta(2, 2)

Posterior: Beta(2+2, 4+2) = Beta(4, 6)

Example inspired by: https://towardsdatascience.com/understanding-conjugate-priors-21b2824cddae

Bayesian Inference

Linear Regression (Bayesian treatment)

Linear Regression (Bayesian treatment)

Given the observed data $\mathcal{D} \ = \left\{x^{(n)}, y^{(n)}\right\}$, we assume to know the noise variance σ^2.

We would like to compute the posterior over the parameters, i.e,

$$p(w|\mathcal{D}, \sigma^2).$$

(We assume throughout a Gaussian likelihood model).

In linear regression **the likelihood is given by:**

$$p(y|X, w, \mu, \sigma^2) = \ \mathcal{N}(y|\mu + Xw,\ \ \sigma^2 I_N)$$

$$\propto \exp(-\frac{1}{2\sigma^2}(y \ - \ \mu \ - Xw)^T(y \ - \ \mu \ - Xw))$$

where μ is an offset term.

Linear Regression (Bayesian treatment)

The conjugate prior of a Gaussian likelihood is also Gaussian*, which we will denote by

$$p(w) = \mathcal{N}(w|w_0, V_0).$$

Using the Bayes rule for Gaussian*, the posterior is given by

$$p(w|X, y, \sigma^2) \;\propto \mathcal{N}(w|w_0, V_0)\, \mathcal{N}(\mathrm{y}|\mathrm{Xw}, \sigma^2 I_N) = \mathcal{N}(w|w_N, V_N)$$

where

$$w_N = \; V_N V_0^{-1} w_0 + \frac{1}{\sigma^2} V_N X^T y$$

$$V_N = \; \sigma^2(\sigma^2 V_0^{-1} + \; X^T X)^{-1}$$

* See: Murphy K., „Machine Learning: A Probabilistic Perspective“ (2012)

Linear Regression (Bayesian treatment)

The posterior predictive distribution at a test point x is given by *

$$p(y|x, \boldsymbol{\mathcal{D}}, \sigma^2) = \int \mathcal{N}(y|x^T w, \sigma^2) \mathcal{N}(w|w_N, V_N) dw$$

$$= \mathcal{N}(y|w_N^T x, \sigma_N^2(x))$$

where $\sigma_N^2(x) = \sigma^2 + x^T V_N x$.

The variance in this prediction depends on the variance of the observation noise, σ^2, and the variance in the parameters, $\mathrm{V_N}$.

Gaussian processes

From Bayesian linear regression to Gaussian Processes

Prior on parameters

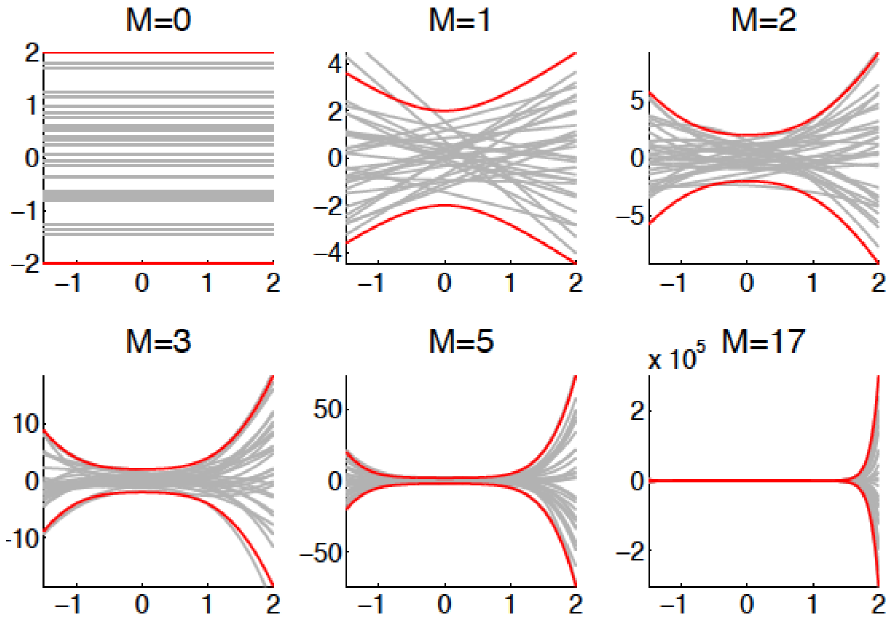
A model M is the result of the choice of:

- A **model structure**
- The model's parameters

In the example:

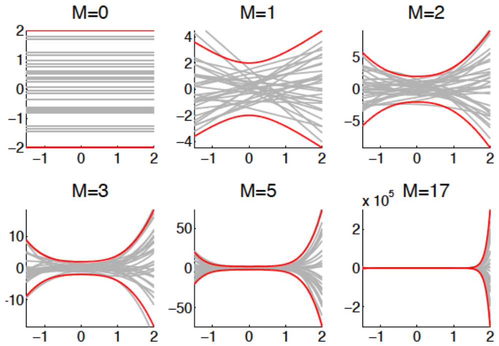
$$f_w(x) = \sum_{m=0}^{M} \omega_m \Phi_m(x), \quad \text{with } \Phi_m(x) = x^m$$

We have defined a prior distribution over functions but **in an indirect way**



Priors on functions

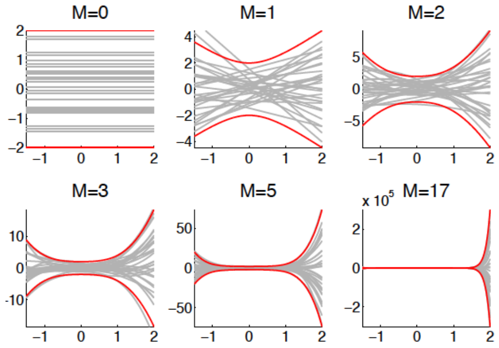
Models with priors on the weights *indirectly* specify priors over functions.

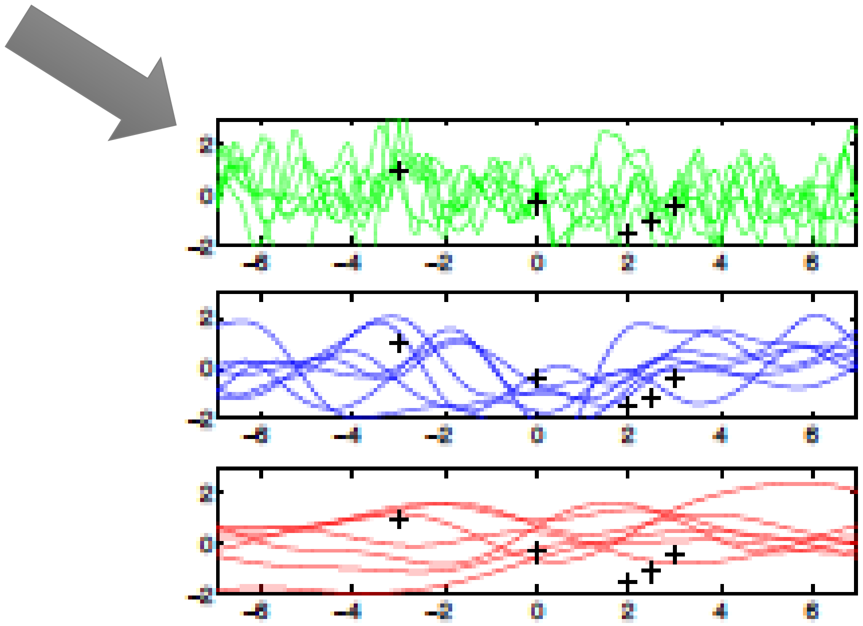


Priors on functions

Models with priors on the weights *indirectly* specify priors over functions.

- **What about specifying priors on functions directly?**
- **What does a probability density over functions even look like?**

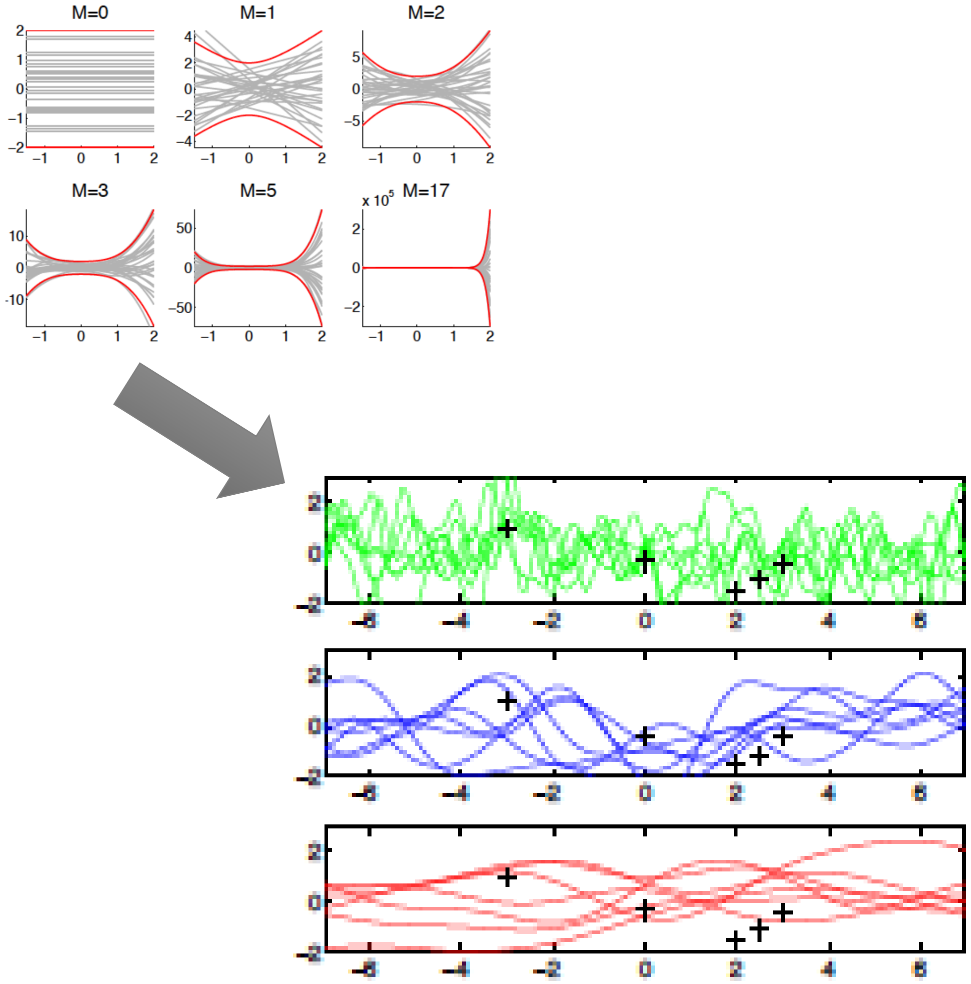




Priors on functions

Models with priors on the weights *indirectly* specify priors over functions.

- **What about specifying priors on functions directly?**
- **What does a probability density over functions even look like?**



Why move beyond Bayesian Linear Regression?

BLR Limitations: Bayesian Linear Regression (BLR) works well with linear assumptions or predefined basis functions, but struggles with complex, non-linear patterns.

→ A **Gaussian Process (GPs) is a non-parametric Bayesian model** that can capture complex, non-linear relationships without specifying a fixed function form.

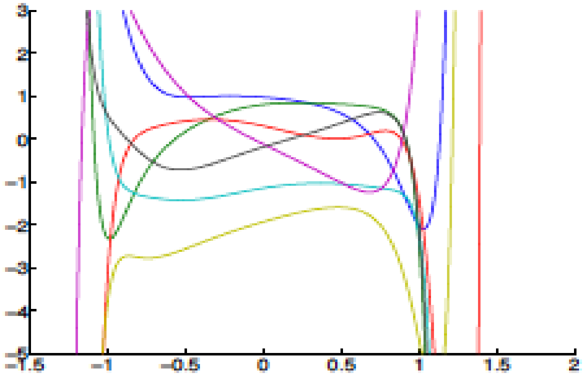
Key Difference:

- Bayesian linear regression uses fixed basis functions.
- GPs treat the function itself as a random variable with uncertainty.

Prior on parameters

The Bayes rule can be written as:

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



Samples from the prior

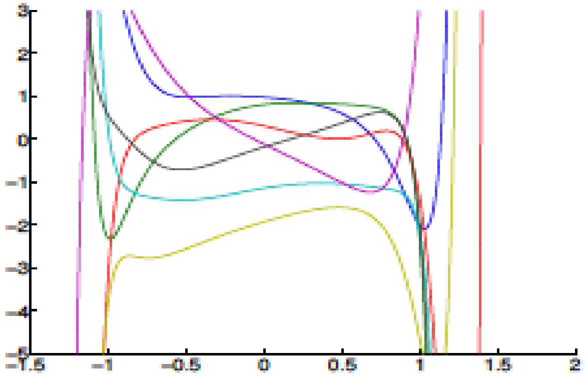
Prior on parameters

The Bayes rule can be written as:

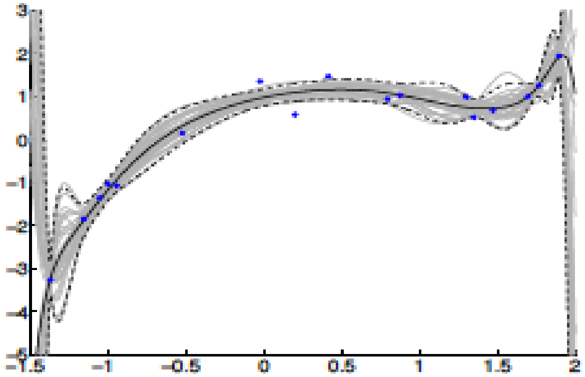
$$p(f|y) = \frac{p(y|f)p(f)}{p(y)}$$

We keep the functions which are “closer” to the data

→ Notion of **closeness** is given by the likelihood $p(y|f)$

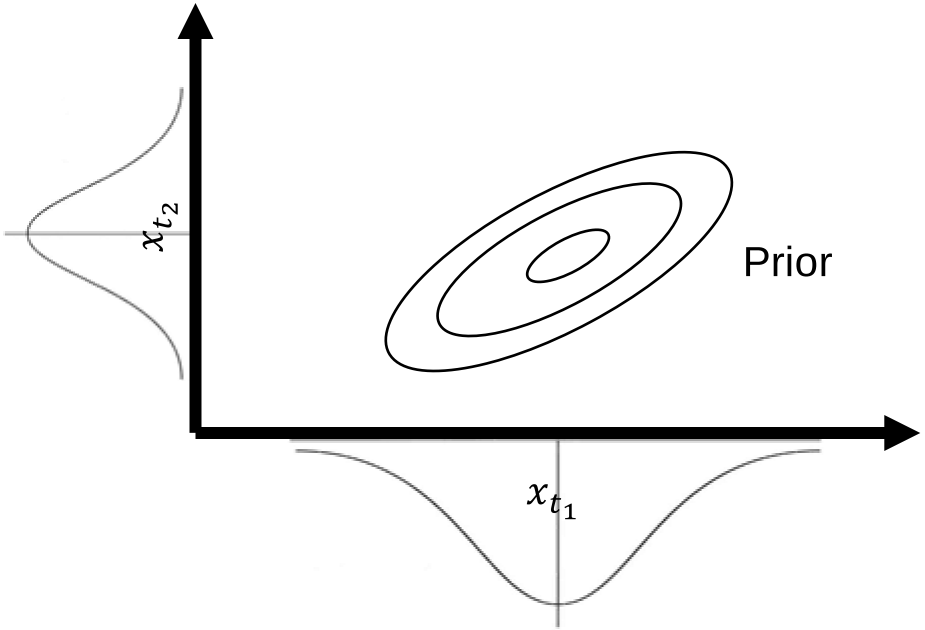


Samples from the prior

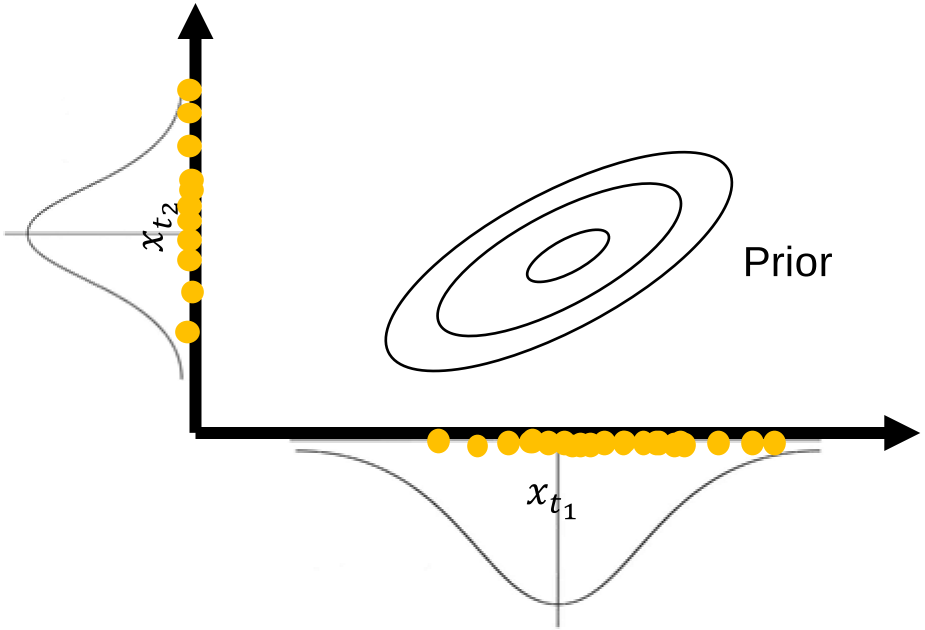


Samples from the posterior

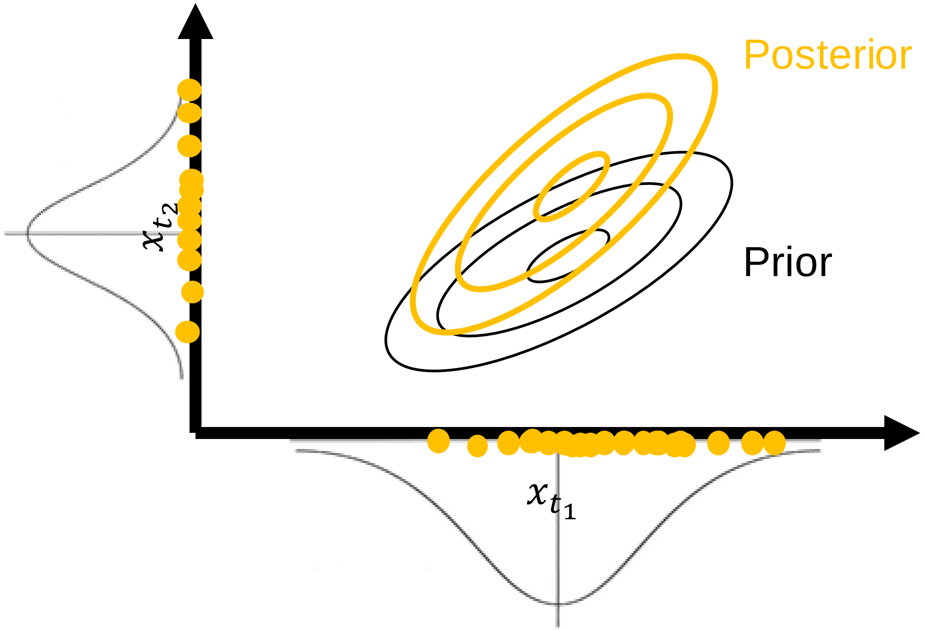
Towards Gaussian Processes



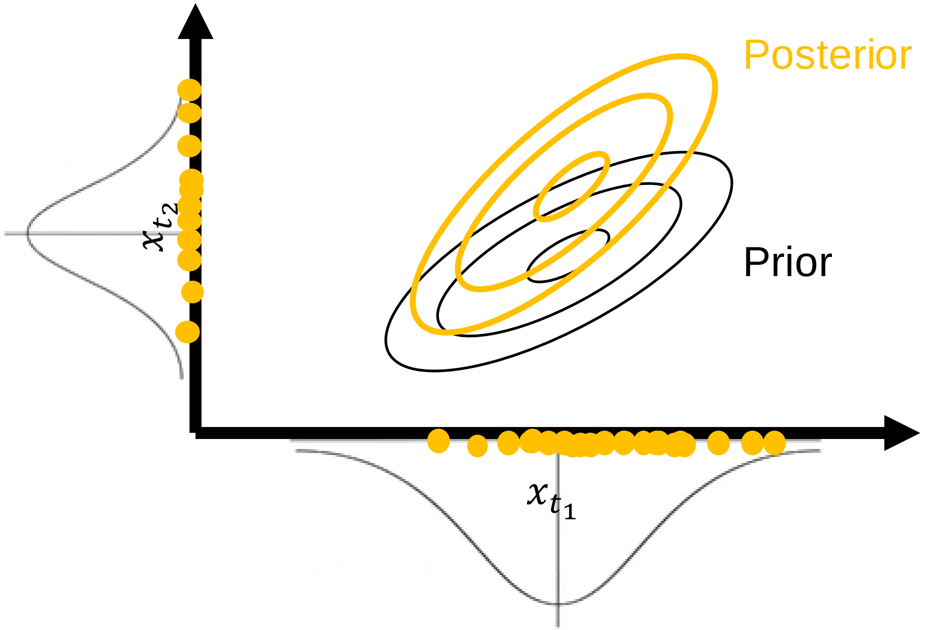
Towards Gaussian Processes



Towards Gaussian Processes

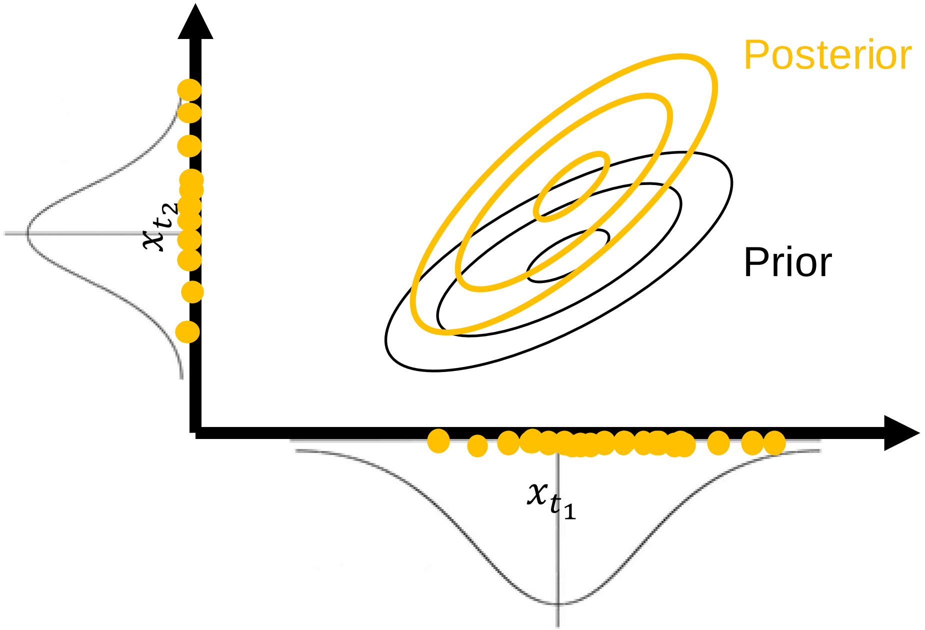


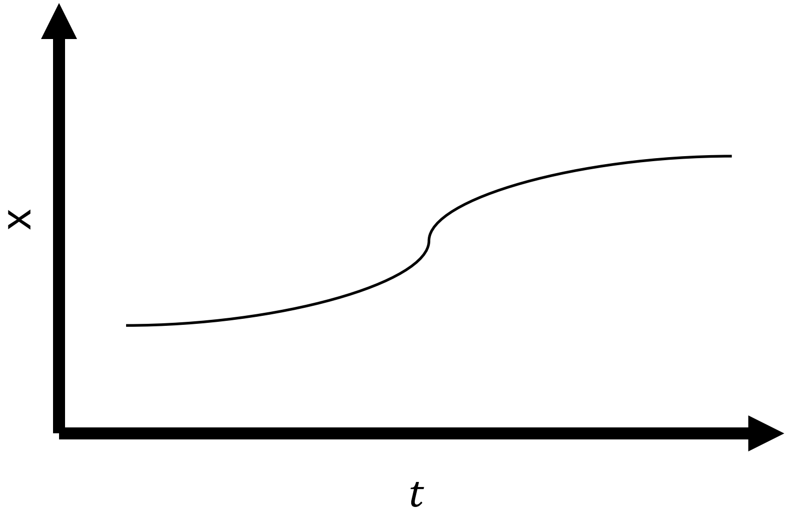
Towards Gaussian Processes



For **multivariate Gaussian distributions** we look at groups of real-valued variables.

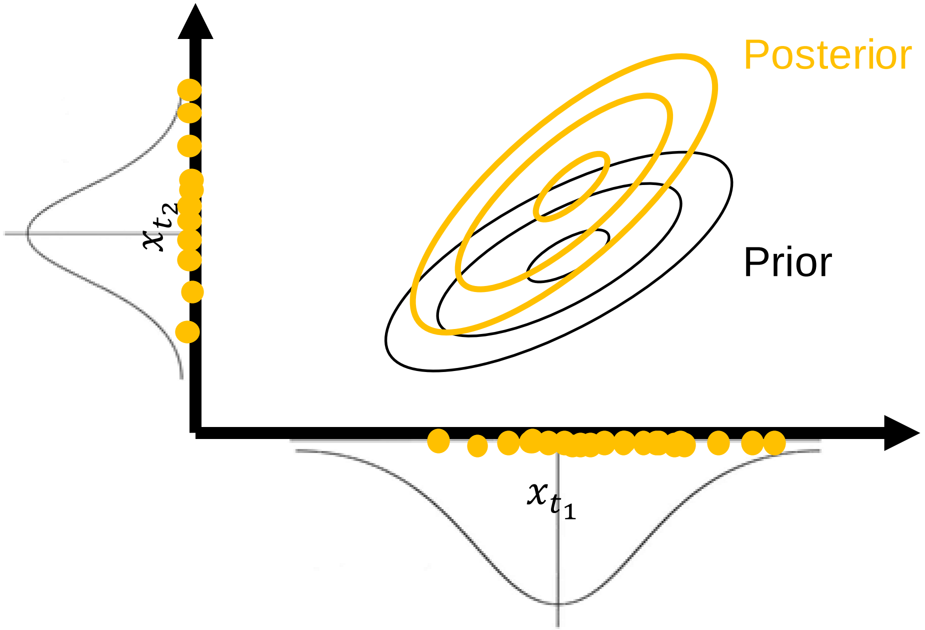
Towards Gaussian Processes

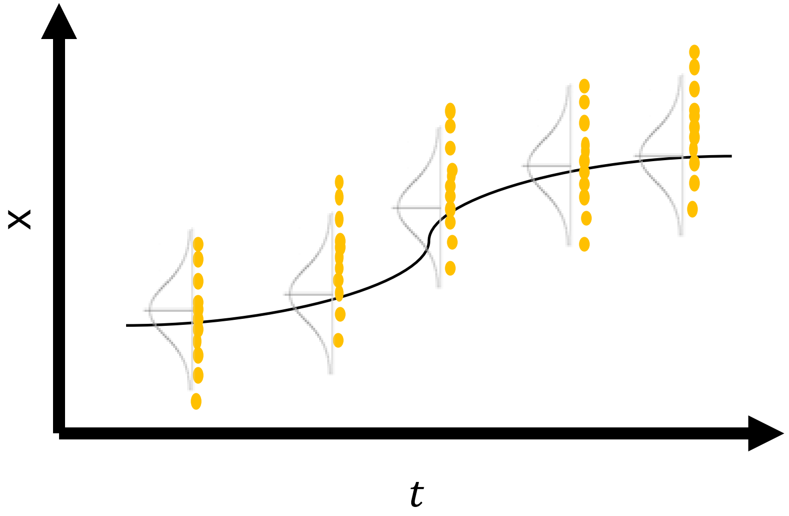




For **multivariate Gaussian distributions** we look at groups of real-valued variables.

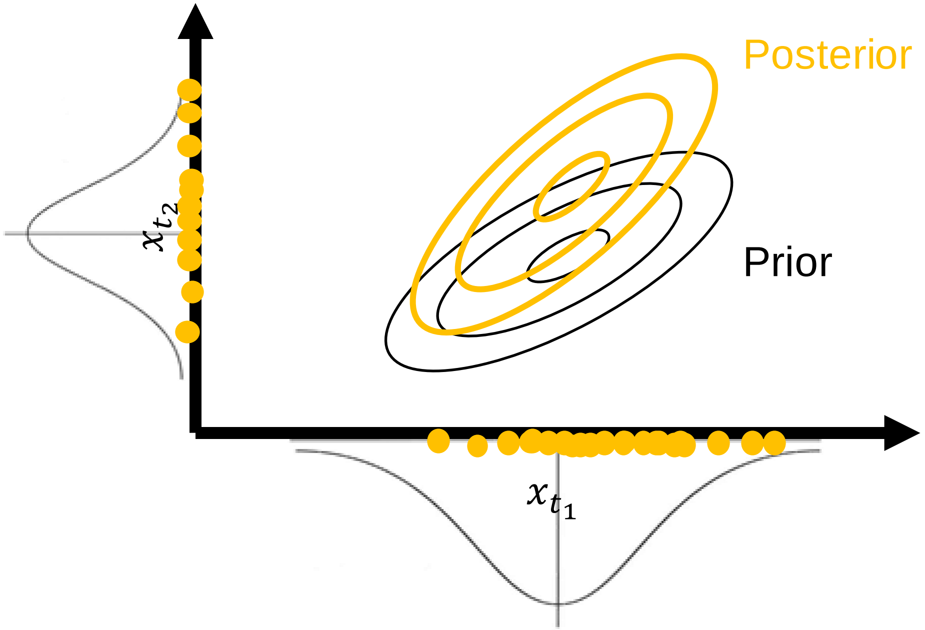
Towards Gaussian Processes

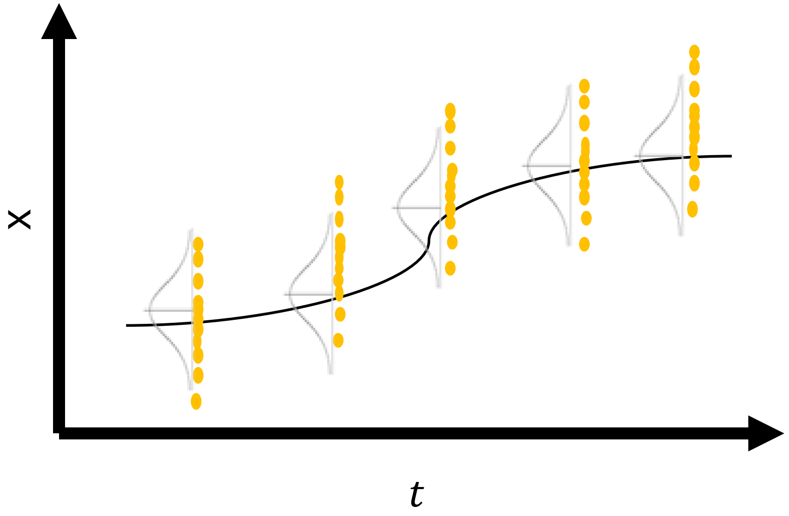




For **multivariate Gaussian distributions** we look at groups of real-valued variables.

Towards Gaussian Processes





For **multivariate Gaussian distributions** we look at groups of real-valued variables.

For **Gaussian processes** we look at very many random variables with Gaussian distribution.

→ GP are functions of (potentially infinite) number of real-valued variables.

Gaussian Process (GP)

Definition:

A function $\boldsymbol{f}$ is a Gaussian process if $\boldsymbol{f}(\boldsymbol{t}) = [\boldsymbol{f}(\boldsymbol{t_1}), \dots, \boldsymbol{f}(\boldsymbol{t_N})]^{\boldsymbol{T}}$ has multivariate distribution for each $\boldsymbol{t}$ $= [\boldsymbol{t_1}, \dots, \boldsymbol{t_N}]^{\boldsymbol{T}}$.

For any subset of t: $f(t) \sim N(\mu(t), \Sigma(t, t'))$

Gaussian Process (GP)

Notice: here we use t for time, but in general we can have a $x \in \mathbb{R}^d$.

The **mean function** is defined as

$$\mu: \mathbb{R} \to \mathbb{R} \quad \text{(or, } \mathbb{R}^d \to \mathbb{R}\text{)}$$

- Often, we subtract the mean from the data to have $\mu(t) = 0, \quad \forall\, t$

The **covariance function** is defined as $\Sigma\colon\ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$; positive semidefinite matrix.

- Often for GP we refer to a **kernel function** $k(t, t')$.

Gaussian Process (GP)

Notice: here we use t for time, but in general we can have a $x \in \mathbb{R}^d$.

The **mean function** is defined as

$$\mu: \mathbb{R} \rightarrow \mathbb{R} \quad \text{(or, } \mathbb{R}^d \rightarrow \mathbb{R}\text{)}$$

- Often, we subtract the mean from the data to have $\mu(t) = 0, \quad \forall\, t$

The **covariance function** is defined as $\Sigma\colon\ \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$; positive semidefinite matrix.

- Often for GP we refer to a **kernel function** $k(t, t')$.

We can, then, rewrite the Gaussian process as:

- $\boldsymbol{f(t) \sim \mathcal{N}(\mu(t), k(t, t'))}$ or $\boldsymbol{f(t) \sim \mathcal{N}(0, k(t, t'))}$

Gaussian Process (GP)

Notice: here we use t for time, but in general we can have a $x \in \mathbb{R}^d$.

The **mean function** is defined as

$$\boxed{\mu: \mathbb{R} \to \mathbb{R}} \text{ (or, } \mathbb{R}^d \to \mathbb{R}\text{)}$$

- Often, we subtract the mean from the data to have $\mu(t) = 0, \quad \forall\, t$

The **covariance function** is defined as $\Sigma\colon\ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$; positive semidefinite matrix.

- Often for GP we refer to a **kernel function** $k(t, t')$.

A GP is defined by its mean and kernel function, so we can write:

$$f \sim GP(\mu, k)$$

We can, then, rewrite the Gaussian process as:

- $\boldsymbol{f(t) \sim \mathcal{N}(\mu(t), k(t, t'))}$ or $\boldsymbol{f(t) \sim \mathcal{N}(0, k(t, t'))}$

Bayesian Inference and Gaussian Processes

Recap

Recap

- Bayes‘ theorem
  - Posterior, likelihood, prior, marginal
- Bayesian model selection
  - Occam‘s razor
- Prior distribution
  - Informative/uninformative
  - Conjugate priors
- Linear regression with Bayes
  - Uncertainty
- Gaussian processes
  - Basic definition