The Bayesian approach to statistics: Basics

For Bayesians, all prior knowledge (or lack of) about unknown parameters should be described by a probability density.

Back to the biased coin

The Bayesian statistician may assume that his **lack of knowledge** (or **prior belief**) about θ **before** she/he has seen the data, should be represented by a prior distribution. Take eg

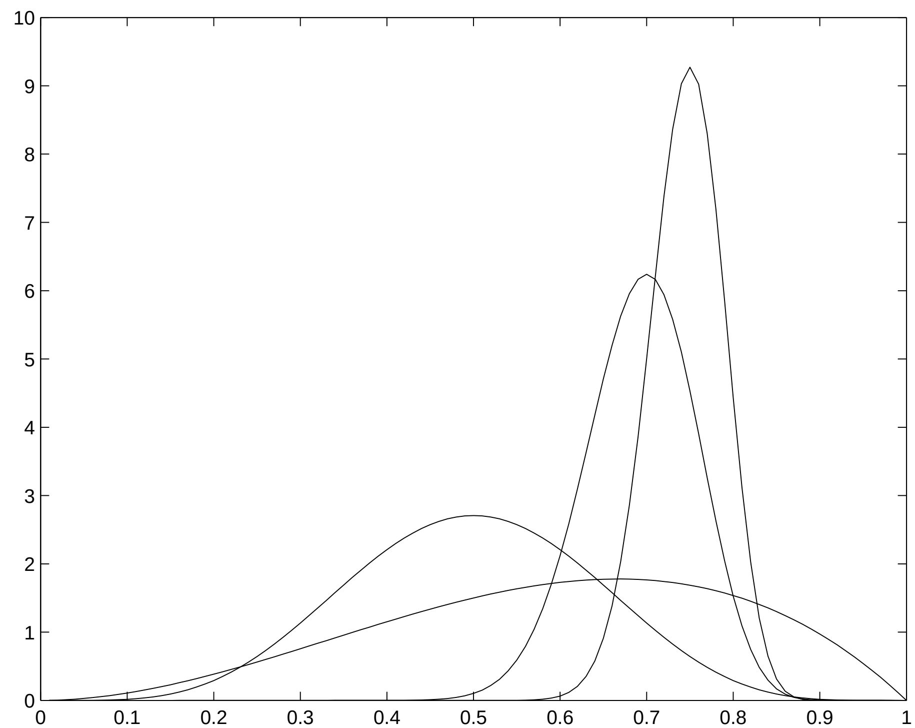
$$p(\theta) = 1 \qquad \text{for } 0 \leq \theta \leq 1 \ .$$

The information **from the data** is described by the likelihood $P(D|\theta)$. Using **Bayes rule**, we compute the **posterior distribution** which gives our belief about θ **after** seeing the data

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

with the **evidence**

$$P(D) = \int_0^1 P(D|\theta)\, p(\theta)\, d\theta \ .$$



Posterior density of θ for the biased coin for $n = 3, 10, 50, 100$. The true value under which the data were generated was $\theta = 0.7$.

Estimators:

A reasonable estimate for the unknown parameter could be the **MAP value** for θ, ie the value which has the **Ma**ximum **Posterior** probability (density). For our choice of prior, this coincides with the ML value.

Another estimator is the the **posterior** mean of θ which is given by

$$\widehat{\theta}_{pm} = \int_0^1 \theta \, p(\theta|D) \, d\theta = \frac{n_1 + 1}{n + 2}$$

$\widehat{\theta}_{pm}$ minimises the **loss function**

$$L_2(\widehat{\theta}) = \int \left(\widehat{\theta} - \theta\right)^2 p(\theta|D) \, d\theta$$

For large n, we see that the posterior mean $\widehat{\theta}_{pm} \to \widehat{\theta}_{ML}$ and the **posterior variance** $\to 0$.

In general, the **Bayes optimal prediction** for the unknown distribution is the **predictive distribution**

$$p(x|D) = \int_{-\infty}^{\infty} p(x|\theta)p(\theta|D)d\theta$$

Properties of Bayes procedures

- Implements prior knowledge

- Regularises problem if small amount of data

- Simple approach to model selection, error bars

- Conceptually simple but often computationally hard

- Could be sensitive to wrong priors, but we can learn priors too!

Bayes for Gaussian densities: 1-D

We assume that σ^2 is known but μ is unknown. Use a (conjugate) prior

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\, e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

This yields the posterior density

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{p(D)} = \frac{p(\mu)}{p(D)} \prod_i \left\{ \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right\} = \frac{1}{\sqrt{2\pi\sigma_n^2}}\, e^{-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}}$$

with

$$\begin{aligned} \mu_n &= \tfrac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\overline{x} + \tfrac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0, \\ \frac{1}{\sigma_n^2} &= \tfrac{n}{\sigma^2} + \tfrac{1}{\sigma_0^2}, \end{aligned}$$

where $\overline{x}$ is the sample mean $\sum_i x_i/n$.

Conjugate priors

For exponential families, conjugate priors allow for simple computations:

$$p(\boldsymbol{\theta}|\boldsymbol{\tau}, n_0) \propto \exp\left[\boldsymbol{\psi}(\boldsymbol{\theta}) \cdot \boldsymbol{\tau} + n_0 g(\boldsymbol{\theta})\right]$$

In this case, the posterior will be of the same form:

$$p(\boldsymbol{\theta}|D\boldsymbol{\tau}, n_0) \propto \exp\left[\boldsymbol{\psi}(\boldsymbol{\theta}) \cdot (\sum_{i=1}^{n} \boldsymbol{\phi}(x_i) + \boldsymbol{\tau}) + (n + n_0) g(\boldsymbol{\theta})\right]$$

We simply replace $n_0 \to n_0 + n$ and $\boldsymbol{\tau} \to \sum_{i=1}^{n} \boldsymbol{\phi}(x_i) + \boldsymbol{\tau}$

Bayes Model selection

If we have a variety of models $\mathcal{M}_1$, $\mathcal{M}_2$, ... with different priors on parameters $p(\theta_1|\mathcal{M}_1)$, $p(\theta_2|\mathcal{M}_2)$, etc, the optimal thing would be a prior over models $P(\mathcal{M})$ and mix them all together. One may then calculate the posterior probability of a model

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

and vote for the most likely one. For equal priors $P(\mathcal{M})$ we choose the model with the largest **evidence** $\int P(D|\theta, \mathcal{M})p(\theta|\mathcal{M})d\theta$.

Example: Bayesian polynomial regression

Assume data generated as $y_i = f(x_i) + \nu_i$ for $i = 1, \dots, N$, with $f(\cdot)$ unknown, ν_i i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.

Class of models: polynomials

$$f_{\mathbf{w}}(x) = \sum_{j=0}^{K} w_j x^j$$

allowing for different orders K. The **likelihood** is

$$p(D|\mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\sum_{i=1}^{N} \frac{(y_i - f_{\mathbf{w}}(x_i))^2}{2\sigma^2}\right]$$

Prior distribution on weights $p(\mathbf{w}) = \frac{1}{(2\pi\sigma_0^2)^{(K+1)/2}} \exp\left[-\frac{\sum_{j=0}^{K} w_j^2}{2\sigma_0^2}\right]$

Posterior density of the parameters $\mathbf{w}$ is given by

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)}$$

which is a multivariate Gaussian. The *evidence* of the data:

$$p(D) = \int p(D|\mathbf{w})\, p(\mathbf{w}) d\mathbf{w}$$

The posterior density is a multivariate Gaussian density with mean

$$E[\mathbf{w}|D] = \left(\frac{\sigma^2}{\sigma_0^2}\mathbf{I}_{K+1} + \mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{y} \tag{8}$$

where the matrix elements of $\mathbf{X}$ are given by $X_{lk} = x_l^k$.

We can show that the evidence of the data is given by:

$$\ln p(D) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} \ , \tag{9}$$

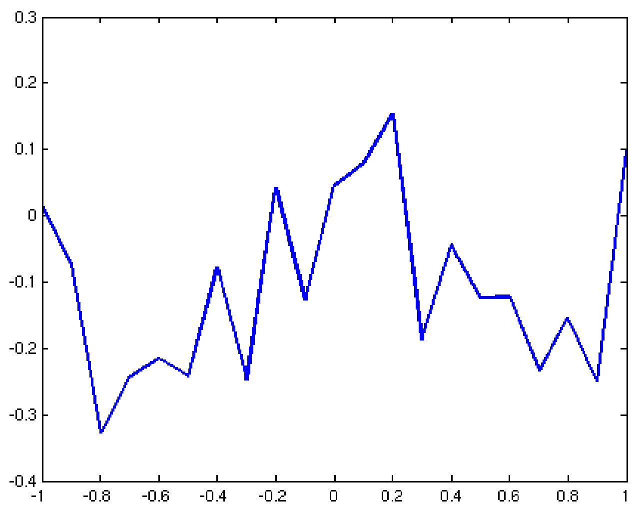
where

$$\mathbf{\Sigma} = \sigma_0^2\mathbf{X}\mathbf{X}^T + \sigma^2\mathbf{I}_N \tag{10}$$

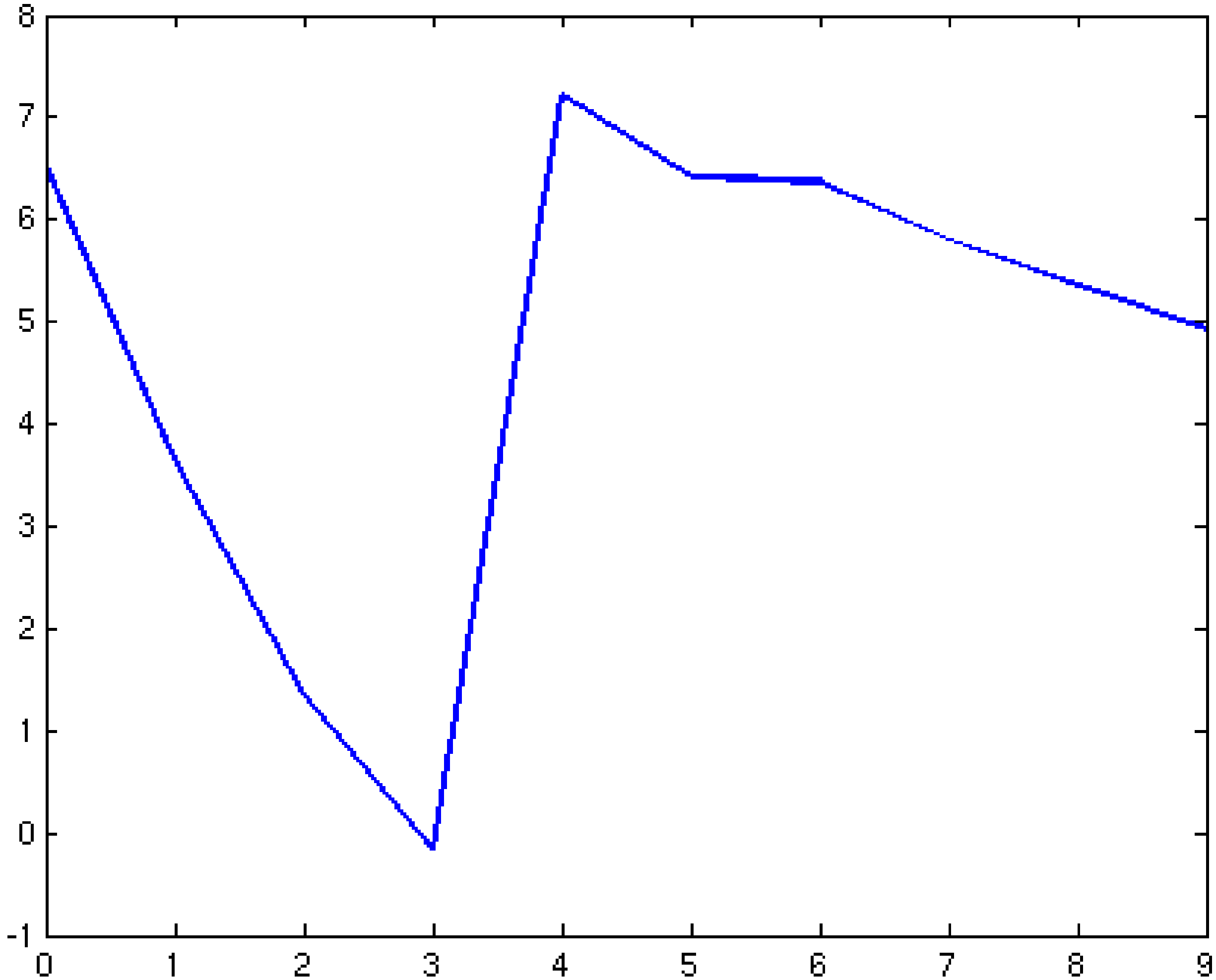
Experiment: $N = 21$ data-points y_i, equally spaced inputs x_i, with true $f(x) = x^4 - x^2$ and $\sigma^2 = 0.01$ in the interval $[-1, 1]$.

prior distribution with variance $\sigma_0^2 = 1$.

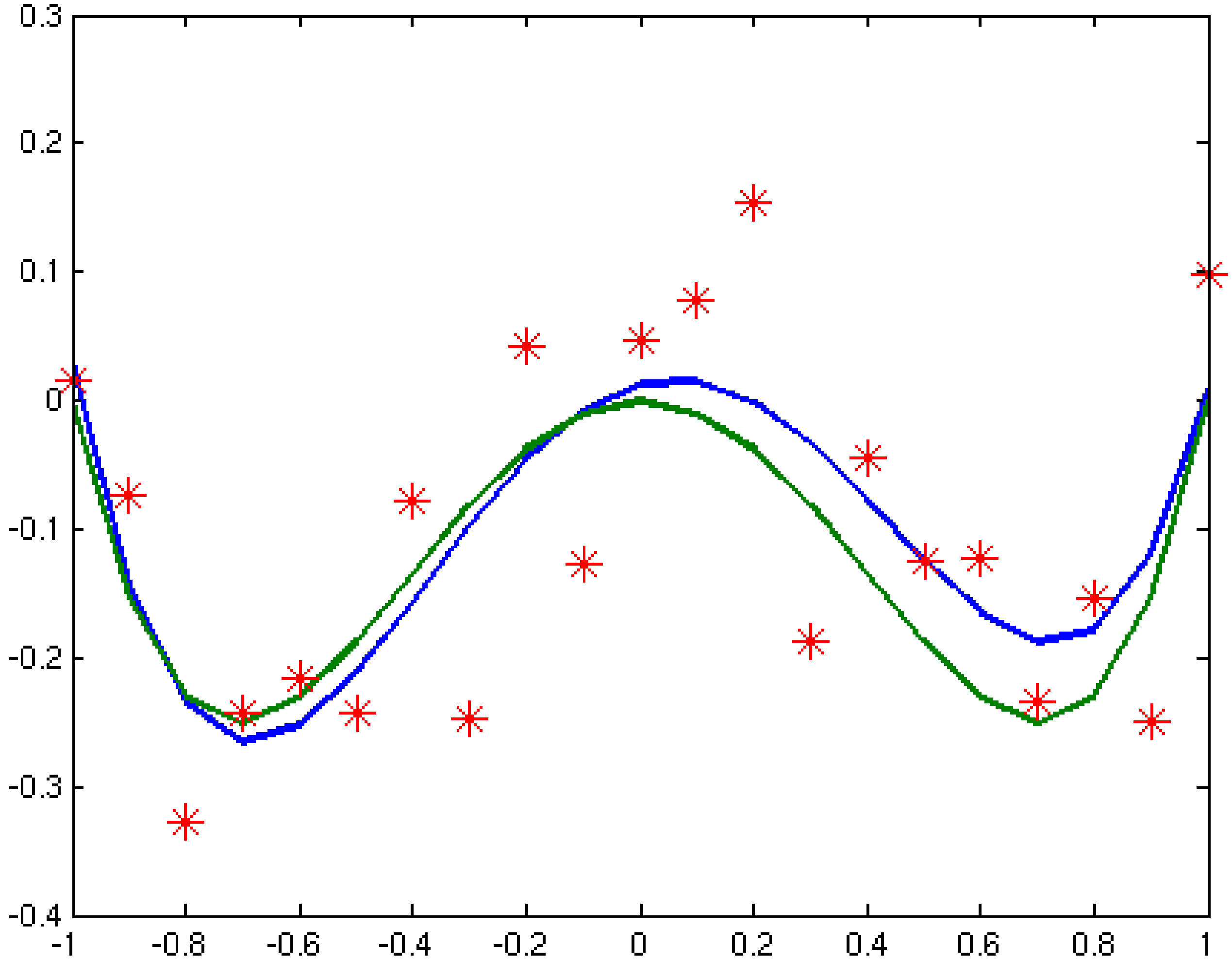
Typical observations



Log-evidence as function of K

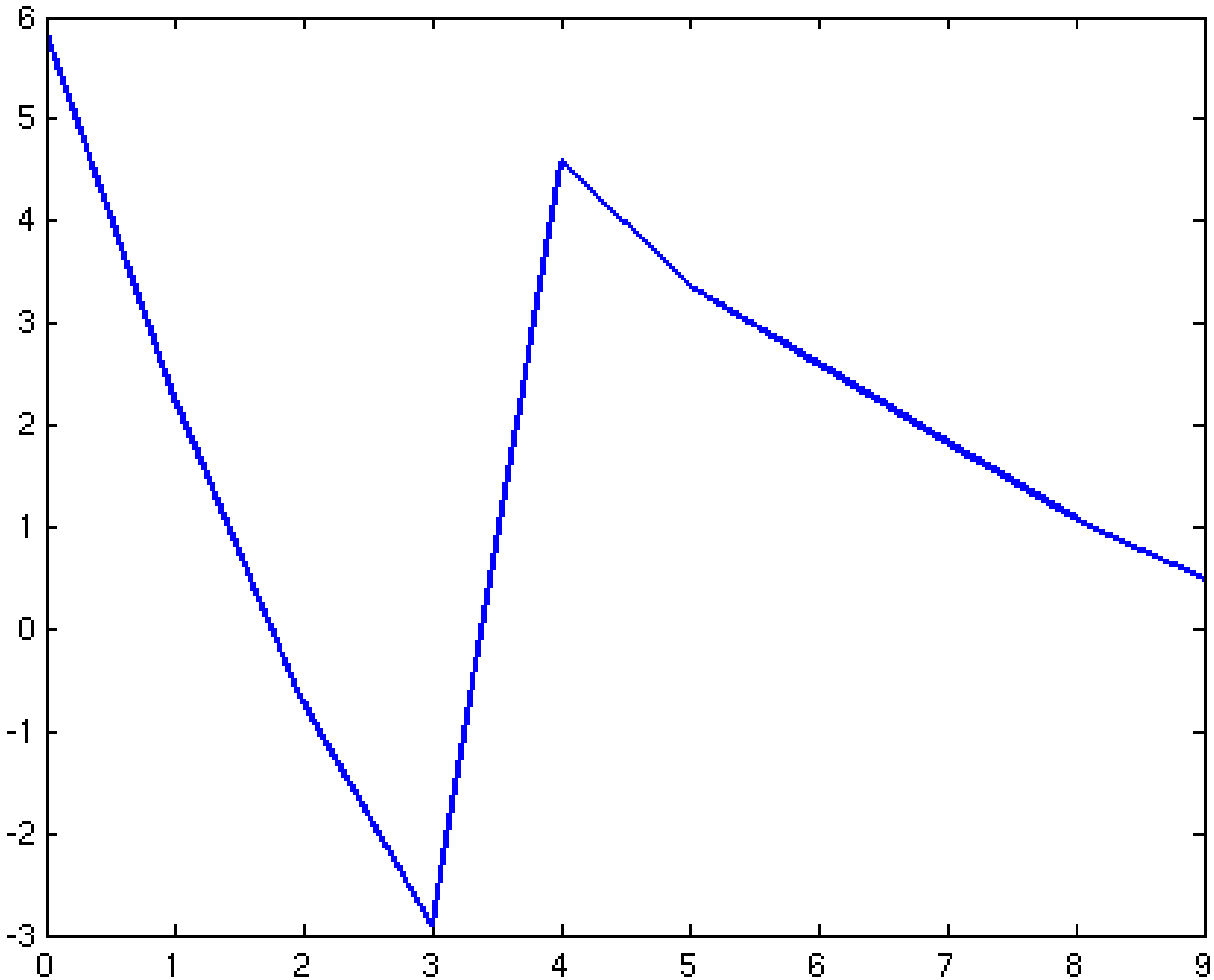


Reconstruction using posterior mean $E[\mathbf{w}|D] = \int d\mathbf{w}\; p(\mathbf{w}|D)\; f_{\mathbf{w}}(x)$



The same, but now with a different prior $\sigma_0 = 2$

Log-evidence as function of K



Reconstruction using posterior mean $E[\mathbf{w}|D] = \int d\mathbf{w}\ p(\mathbf{w}|D)\ f_{\mathbf{w}}(x)$