A Short Introduction to Bayesian Inference

Olivier Le Maître[1]
with Colleagues & Friends
Francesco Rizzi and Omar Knio

[1] LIMSI, CNRS
UPR-3251, Orsay, France
https://perso.limsi.fr/olm/

UTOPIÆ Uncertainty Treatment and Optimisation in Aerospace Engineering
Handling the unknown at the edge of tomorrow

PhD course on UQ - DTU

Background: Bayes' theorem

- A set of observations is used to update (refine) some *a priori* knowledge about a certain hypothesis.
- Suppose that we have a set of data ($\{d^i\}_{i=1}^N$) and we assume a certain model to represent it. Let H be the set of parameters (i.e. our hypotheses) defining (parametrizing) our model.

Bayes' theorem

$$\pi(H|\{d^i\}_{i=1}^N) \propto \mathcal{L}(\{d^i\}_{i=1}^N|H)\, \mathcal{P}(H)$$

- ◇ $\mathcal{P}(H)$ is the prior of H.
- ◇ $\mathcal{L}(\{d^i\}_{i=1}^N|H)$ is the likelihood.
- ◇ $\pi(H|\{d^i\}_{i=1}^N)$ is the posterior probability.

- Interpretation: a process of continuously updating the current state of knowledge in view of new observations.

Limsi

Background: Bayes' theorem

- The likelihood $\mathcal{L}(\{d^i\}_{i=1}^{N}|H)$ represents the probability of obtaining the data given the hypotheses H.
- The prior $\mathcal{P}(H)$ represents the information that we have about the parameters before the observations are taken into consideration.
- The choice of the prior is a key step in the inference process and should be based, when possible, on some *a priori* knowledge about the parameters.
- In general, we distinguish between informative (e.g. a Gaussian with known mean and variance), and non-informative priors (e.g. a uniform distribution where we only need the upper and lower bounds).
- Let's look at an example.

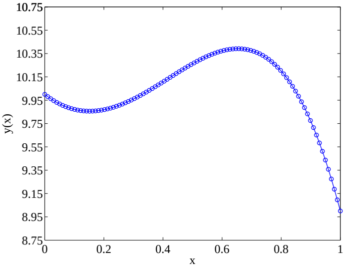
Example

- Suppose that we have the following polynomial:

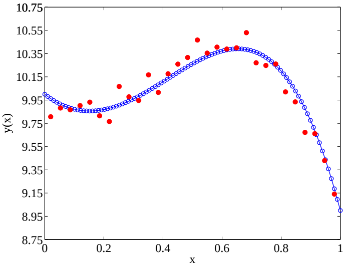
"True" polynomial

$$y(x) = 10 - 2x + 7.5x^2 - 3.3x^3 - 3.2x^4$$

where $x \in (0, 1)$.



- We perturb the "true" curve at N coordinates $\{x_i\}_{i=1}^{N}$ with a Gaussian noise with mean zero and variance 0.01, i.e. $\mathcal{N}(0, 0.01)$.
- This yields a set of noise observations, $(\{x_i, d_i\}_{i=1}^{N})$.
- For this example we have $N = 30$. (We will discuss the effect of the number of observations)



Limsi

Example

- Objective: given the data $\mathbf{d} = \{d_i\}_{i=1}^{N}$, can we recover the original polynomial?
- We need to define a model (i.e. the hypothesis) to describe the data.
- Our model is a polynomial of certain order p:

$$M(x) = \sum_{k=0}^{p} c_k x^k \quad (1)$$

- It follows that our set of hypothesis is:

$$H = \{c_0, c_1, c_2, \ldots, c_p\} \quad (2)$$

Bayes' theorem

$$\pi(\{c_k\}_{k=0}^{p} | \{d_i\}_{i=1}^{N}) \propto \mathcal{L}(\{d_i\}_{i=1}^{N} | \{c_k\}_{k=0}^{p})\, \mathcal{P}(\{c_k\}_{k=0}^{p})$$

- We now need to define the likelihood and priors.

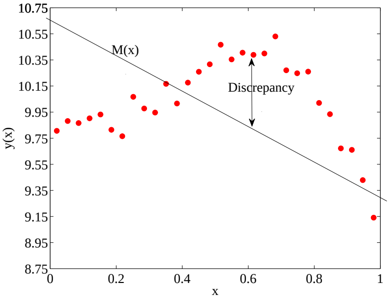
Limsi

Likelihood

- To formulate the likelihood we assume the following relationship:

$$d_i = M(x_i) + \epsilon_i \ ,$$

where ϵ_i is a random variable which represents the discrepancy between the i-th observation, d_i, and the model evaluated at the i-th coordinate, $M(x_i)$.



- Assuming N independent realizations and $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, ..., N$, the likelihood can be written as

$$\mathcal{L} \equiv p(\{d_i\}_{i=1}^N | \{c_k\}_{k=0}^p) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(d_i - M(x_i))^2}{2\sigma^2}\right)$$

- Objective: jointly infer σ^2 and $\{c_k\}_{k=0}^p$.

Limsi

Prior selection

- The choice of a prior should be based, when possible, on some a priori knowledge about the parameters.
- We have $p + 2$ unknowns, i.e. the $(p + 1)$ coefficients $\{c_k\}_{k=0}^{p}$ and the variance σ^2.
- For each c_k, since we have limited information and for the purpose of this exercise, we choose a uniform distribution

$$\mathcal{P}(c_k) = \begin{cases} \frac{1}{400} & \text{for } -200 < c_k \leq 200, \\ 0 & \text{otherwise ,} \end{cases}$$

- In theory, these bounds can be made arbitrarily large.
- We know that σ^2 cannot be negative: this information is what we defined as *a priori* knowledge about a parameter. We assume a Jeffreys prior:

$$\mathcal{P}(\sigma^2) = \begin{cases} \frac{1}{\sigma^2} & \text{for } \sigma^2 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Posterior

Final form of the joint posterior

$$\pi(\{c_k\}_{k=0}^{p}, \sigma^2 | \{d_i\}_{i=1}^{N}) \propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(d_i - M(x_i))^2}{2\sigma^2}\right)\right] \mathcal{P}(\sigma^2) \prod_{j=0}^{p} \mathcal{P}(c_j)$$

- The problem now reduces to simulate (sample) this posterior.
- We are dealing with a $(p+2)$-dimensional probability distribution.
- For high-dimensional cases, which are also the only interesting ones, use Markov chain Monte Carlo (MCMC) methods.
- MCMC: class of algorithms suitable to sample high-dimensional probability distributions.
- Must pay attention to mixing ability, convergence...
- Important feature: the quality of the sample improves as a function of the number of steps.

Posterior sampling

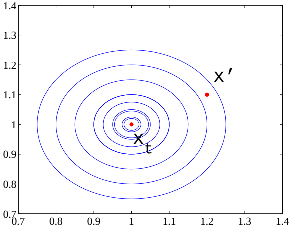
- Basic idea: the algorithm generates a Markov chain, i.e. at a certain time t, the state x_t depends only on the previous one x_{t-1}.

---

1 Suppose the current value of the chain is x_t. We draw a candidate, x', from a Gaussian centered at the current state and with a given covariance matrix: $x' \sim N(x_t, \beta^2 I)$.

2 Calculate the follwing ratio:

$$r = \frac{\pi(x')}{\pi(x_t)}$$



2 Draw a sample $\alpha \sim U(0,1)$.

3 The new state x_{t+1} is chosen according to the following rule:

$$x_{t+1} = \begin{cases} x' & \text{if} \quad \alpha < r, \qquad \text{ACCEPTED}, \\ x_t & \text{if} \quad \textit{otherwise}, \quad \text{REJECTED}. \end{cases}$$

4 Repeat loop...

- The parameter β must be tuned to have a well-mixing chain and must be fixed once at the beginning. In general, the objective is to have an average acceptance ratio between 0.2 and 0.5.

Example 1

Zeroth-order model

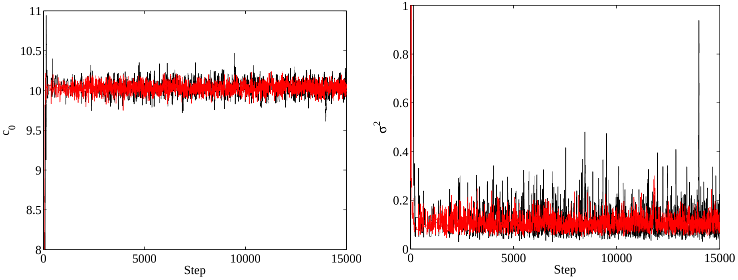
- Suppose that we infer a zeroth-order polynomial:

$$M(x) = c_0$$

- We know that this is far from the true model defined before, which was a fourth-order polynomial.

Two-dimensional joint posterior

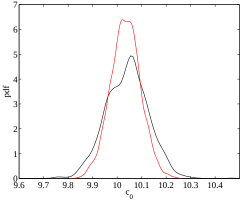
$$\pi(c_0, \sigma^2 | \{d_i\}_{i=1}^N) \propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(d_i - c_0)^2}{2\sigma^2} \right) \right] \mathcal{P}(\sigma^2)\, \mathcal{P}(c_0)$$



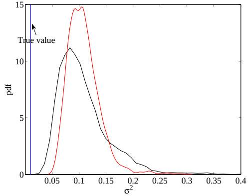
(a) Chain for c_0. (b) Chain for σ^2.

Posterior distributions

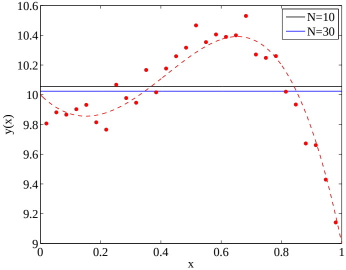
- Chain samples can be used to estimate the marginalized posteriors of the parameters via KDE.



(c) Posterior for c_0.



(d) Posterior for σ^2.



(e) Compare with true.

This approach only allows us to infer the mean value.

Inference for higher-dimensional case

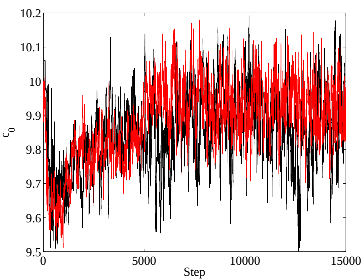
fourth-order model

- Suppose that we infer a fourth-order polynomial:

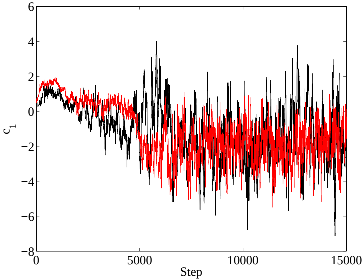
$$M(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4$$

Six-dimensional joint posterior

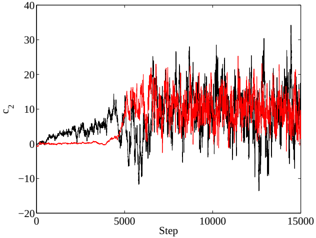
$$\pi(\{c_k\}_{k=0}^{4}, \sigma^2 | \{d_i\}_{i=1}^{N}) \propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(d_i - M(x_i))^2}{2\sigma^2} \right) \right] \mathcal{P}(\sigma^2) \prod_{j=0}^{p} \mathcal{P}(c_j)$$



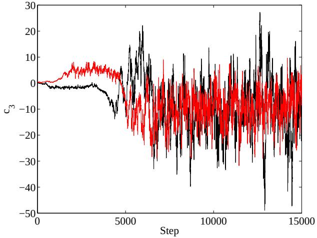
(f) Chain for c_0.



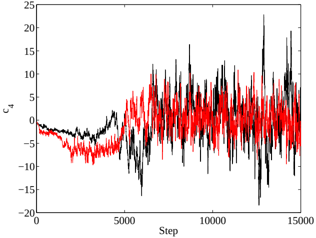
(g) Chain for c_1.



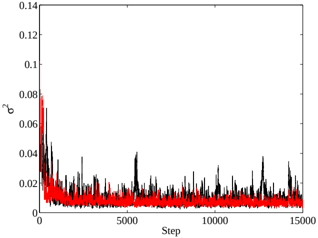
(h) Chain for c_2.



(i) Chain for c_3.



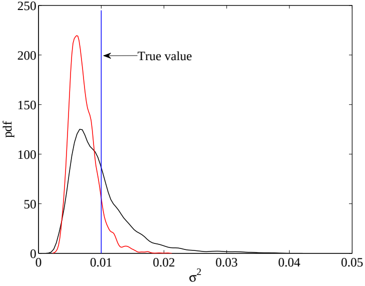
(j) Chain for c_4.



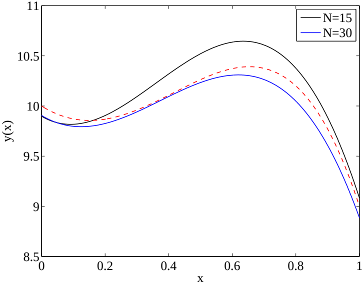
(k) Chain for σ^2.

Closing remarks

- Results based on the MAP estimates of the coefficients.
- Note: increasing the order of the polynomial yields a lower value of the variance because the model is getting closer to the true curve.



(l) Posterior for σ^2.



(m) Comparison between true and inferred curve.