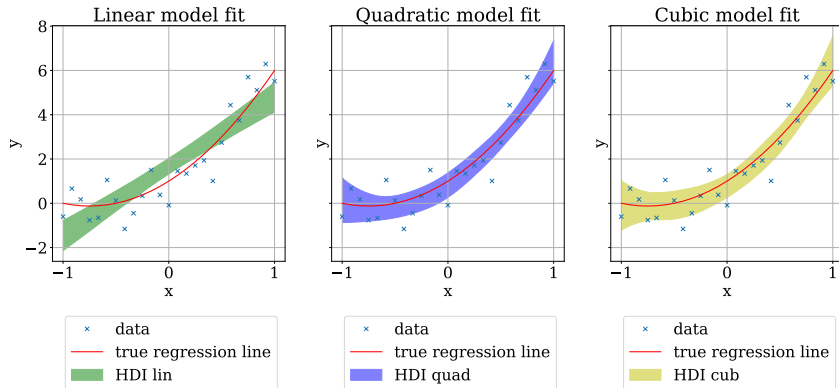


BAYESIAN model selection

Seminar physics760 – Computational Physics



DOMINIC SCHÜCHTER

✉ dschuechter@uni-bonn.de | 🌐 dschuechter

JAKOB KRAUSE

✉ krause@hiskp.uni-bonn.de | 🌐 krausejm

Tutor: ANDREAS WIRZBA

✉ a.wirzba@fz-juelich.de

19.03.2021

BAYES' Theorem – a quick reminder

$$\text{prob}(\boldsymbol{\theta}|y) = p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(y)}$$

with

- ▶ *posterior* $p(\boldsymbol{\theta}|y)$
- ▶ *likelihood* $p(y|\boldsymbol{\theta})$
- ▶ *prior* $p(\boldsymbol{\theta})$
- ▶ *marginal likelihood* $p(y) = \int_{-\infty}^{+\infty} d\boldsymbol{\theta} p(y|\boldsymbol{\theta}) p(\boldsymbol{\theta})$

This can be used for *model selection* (?)

1. Theory

Parameter estimation

Model comparison

2. Methods

Monte-Carlo-Sampling

SAVAGE-DICKEY-Density-Ratio (SDDR)

3. Examples

Coin Flip

Fitting a polynomial of unknown degree

4. Summary

1. Theory

Parameter estimation

Model comparison

2. Methods

Monte-Carlo-Sampling

SAVAGE-DICKEY-Density-Ratio (SDDR)

3. Examples

Coin Flip

Fitting a polynomial of unknown degree

4. Summary

JAN and MARIUS already talked about this, so here we only sketch the basics again

$$\text{marginal posterior } p(\theta_i|y, M) = \int p(\boldsymbol{\theta}|y, M) \prod_{j \neq i} d\theta_j \quad (1)$$

$$\begin{aligned} \text{best fit value(s)} \quad p(\theta|y, M) &= \max \Leftrightarrow \theta = \hat{\theta} \\ \langle \theta \rangle &= \int_{-\infty}^{\infty} d\theta p(\theta|y, M) \cdot \theta \end{aligned} \quad (2)$$

[sivia]

BAYES factor

$$p(M_i|y) = \frac{p(M_i) \cdot p(y|M_i)}{p(y)}. \quad (3)$$

$$\begin{aligned} O_{ij} &:= \frac{p(M_i|y)}{p(M_j|y)} \\ &\quad \text{posterior odds} \\ &= \underbrace{\frac{p(y|M_i)}{p(y|M_j)}}_{\text{BAYES Factor}} \cdot \underbrace{\frac{p(M_i)}{p(M_j)}}_{\text{prior odds}} \quad (4) \\ &= B_{ij} \cdot \frac{p(M_i)}{p(M_j)}. \end{aligned}$$

How do we turn BAYES' theorem into a tool for model comparison?

$ \ln B_{ij} $	Odds	Strength of evidence
< 1.0	$\lesssim 3 : 1$	Inconclusive
1.0	$\sim 3 : 1$	Weak evidence
2.5	$\sim 12 : 1$	Moderate evidence
5.0	$\sim 150 : 1$	Strong evidence

Table 1: Empirical scale for evaluating the strength of evidence when comparing two models M_i vs. M_j , adapted from [Trotta_2008]

[Trotta_2008]

BAYESIAN complexity

$$\mathcal{C}_b = -2 \int d\boldsymbol{\theta} p(\boldsymbol{\theta}|y, M) \log(\mathcal{L}(\boldsymbol{\theta})) + 2 \log(\mathcal{L}(\tilde{\boldsymbol{\theta}})), \quad (5)$$

$$\mathcal{C}_b = \overline{\chi^2(\boldsymbol{\theta})} - \chi^2(\tilde{\boldsymbol{\theta}}), \quad (6)$$

- ▶ describes how many model parameters the data is able to constrain
- ▶ is a useful tool for examining models with an increasing number of parameters

[kunz]

1. Theory

Parameter estimation

Model comparison

2. Methods

Monte-Carlo-Sampling

SAVAGE-DICKEY-Density-Ratio (SDDR)

3. Examples

Coin Flip

Fitting a polynomial of unknown degree

4. Summary

Benefits of Monte-Carlo-Sampling

$$\langle \boldsymbol{\theta} \rangle \approx \int p(\boldsymbol{\theta}|y) \boldsymbol{\theta} d\boldsymbol{\theta} = \frac{1}{N} \sum_{t=0}^{N-1} \boldsymbol{\theta}^{(t)}, \quad (7)$$

$$\langle f(\boldsymbol{\theta}) \rangle \approx \frac{1}{N} \sum_{t=0}^{N-1} f(\boldsymbol{\theta}^{(t)}). \quad (8)$$

also, *marginal posterior* distributions are obtained trivially by binning values of θ_i ignoring $\theta_{j \neq i}$ [**Trotta_2008**]



Figure 1: ArviZ [ArviZ] and PyMC3 [PyMC3]

But how exactly do we get samples $\boldsymbol{\theta}^{(t)}$?

Sequential Monte Carlo (SMC)

First let us introduce an auxiliary *temperature parameter* $\beta \in [0, 1]$ and write

$$p(\boldsymbol{\theta}|y)_\beta = \frac{p(y|\boldsymbol{\theta})^\beta \cdot p(\boldsymbol{\theta})}{Z_\beta} \quad (9)$$

with $Z_\beta = \int d\boldsymbol{\theta} p(y|\boldsymbol{\theta})^\beta \cdot p(\boldsymbol{\theta})$.

Main idea: gradually sample from simple distribution ($\beta = 0$) to complex/true distribution ($\beta = 1$) using METROPOLIS-HASTINGS

SMC then allows us to estimate the *marginal likelihood* as

$$\hat{p}(y) = \prod_i \frac{\widehat{Z_{\beta_i}}}{Z_{\beta_{i-1}}}. \quad (10)$$

[SMC_PEPE; PyMC3_SMC]

consider model M_j with free parameters ω, ψ and a submodel M_i with one free parameter ψ and fixed $\omega = \omega_\star$. Let us further assume separable priors

$$p(\omega, \psi | M_j) = p(\omega | M_j) p(\psi | M_i).$$

We can then write the BAYES factor as

Savage Dickey-Density-Ratio

$$B_{ij} = B_{ji}^{-1} = \frac{p(\omega | y, M_j)}{p(\omega | M_j)} \bigg|_{\omega=\omega_\star} \quad (\text{SDDR}). \quad (11)$$

[trotta]

1. Theory

Parameter estimation

Model comparison

2. Methods

Monte-Carlo-Sampling

SAVAGE-DICKEY-Density-Ratio (SDDR)

3. Examples

Coin Flip

Fitting a polynomial of unknown degree

4. Summary

- ▶ remember the flipping of a biased coin from the lecture
- ▶ we want to verify that the coin is biased by using the BAYES factor
- ▶ two models: M_1 – assumes a fair coin, M_2 – assumes biased coin

Posterior of the coin flip problem

$$p(\theta|y, M_i) = \frac{p(y|\theta, M_i) \cdot p(\theta|M_i)}{p(y|M_i)} \quad (12)$$

to get $p(y|M_i) = \int_{-\infty}^{+\infty} d\theta p(y|\theta, M_i) p(\theta|M_i)$ we need to specify a *prior* $p(\theta|M_i)$ and a *likelihood* $p(y|\theta, M_i)$

Choosing a prior

$$f(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$
$$:= \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}.$$

[PyMC3_BF]

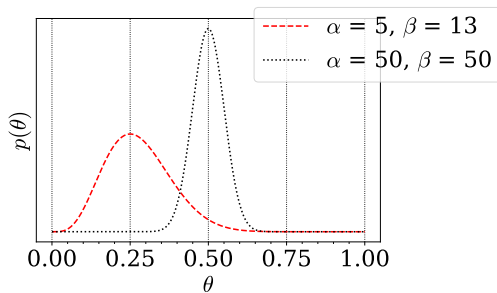


Figure 2: Beta-distribution as *prior* $p(\theta|M_i)$ for the two different models

Choosing a likelihood

Since we can assume i.i.d. outcomes of the coin flip a natural choice is a *Binomial distribution*. If we observe k heads out of N coin throws ($y = (N, k)$)

$$p(y|\theta, M_i) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}.$$

We simulated data for $N = 50$ and the biased coin with $p(H) = 0.25$.

Finally computing the BAYES factor

$$p(y|M_i) \propto \int_0^1 d\theta \frac{1}{B(\alpha, \beta)} \cdot \theta^{\alpha+k-1} \cdot (1-\theta)^{N-k+\beta-1} = \frac{B(\alpha+k, \beta+N-k)}{B(\alpha, \beta)}$$
$$\Rightarrow B_{21} = B_{12}^{-1} = \frac{B(\alpha_2+k, \beta_2+N-k) \cdot B(\alpha_1, \beta_1)}{B(\alpha_1+k, \beta_1+N-k) \cdot B(\alpha_2, \beta_2)} = 9.5839$$

Can we reproduce this numerically?

[PyMC3_BF]

let us obtain 2000 samples from the posterior distribution using the same *likelihood* and *priors* via the SMC algorithm provided by PyMC3

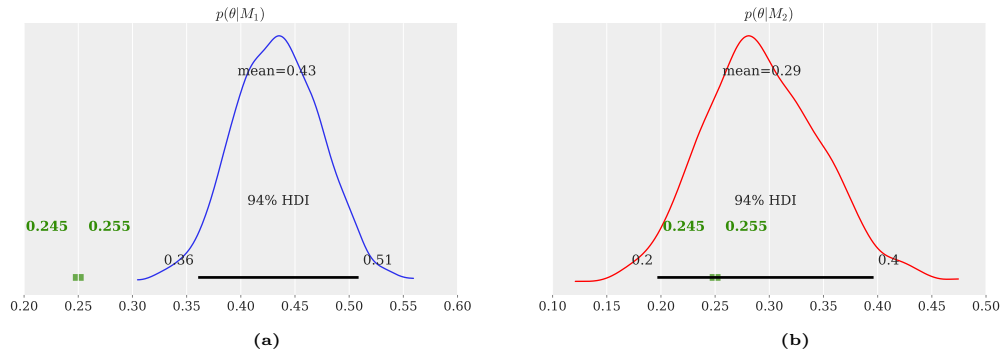


Figure 3: The *marginal posterior* for $\alpha = \beta = 50$ (3a) and $\alpha = 5, \beta = 13$ (3b) of 2000 samples. HDI means highest density interval. The highlighted green intervals denote the expected value.

We find $B_{21} = B_{12}^{-1} = 9.5829 \pm 0.4719$ (noice! ✓)

Fitting a polynomial of unknown degree

We wish to determine the true model underlying the generated data depicted in the figure below

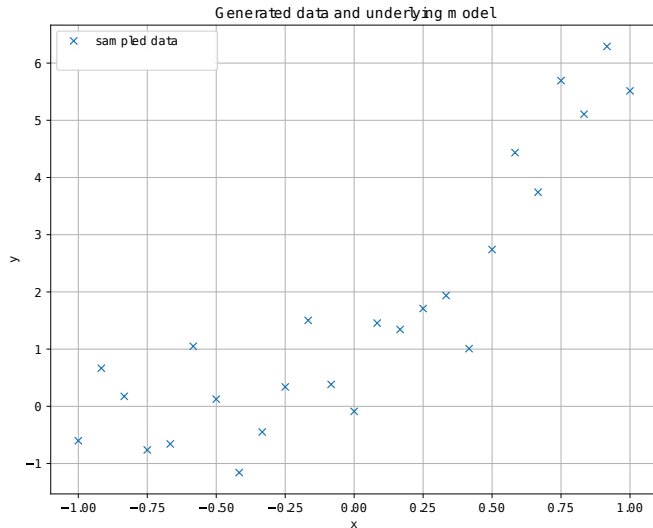


Figure 4: $N = 25$ datapoints distributed with a gaussian noise of $\sigma = 0.7$. Linear? Quadratic? Cubic?

we want to find the correct model by using

- ▶ BAYES factor
- ▶ SDDR (as sanity check)
- ▶ BAYESIAN complexity

We will tackle this problem numerically by sampling from the *posterior* \rightarrow we need to assign *priors* and *likelihoods* again

a suitable choice for *prior* and *likelihood* are normal distributions, since the noise is Gaussian [sivia].

Choosing a prior

The *priors* for the fit-parameters a_α are each described by a normal distribution with $\mu_{\text{prior}} = 0$ and $\sigma_{\text{prior}} = 2$

Choosing a likelihood

$$p(y|\boldsymbol{\theta}, M_i) = \prod_{k=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(f(y_k; \boldsymbol{\theta}) - y_k)^2}{2\sigma^2} \right].$$

where $f(y_k; \boldsymbol{\theta})$ is the fit function $f_i(x) = \sum_{\alpha=0}^i a_\alpha x^\alpha$, with $i = 1, 2, 3$

Fitting a polynomial of unknown degree

Now let's generate 2000 samples following the *posterior*

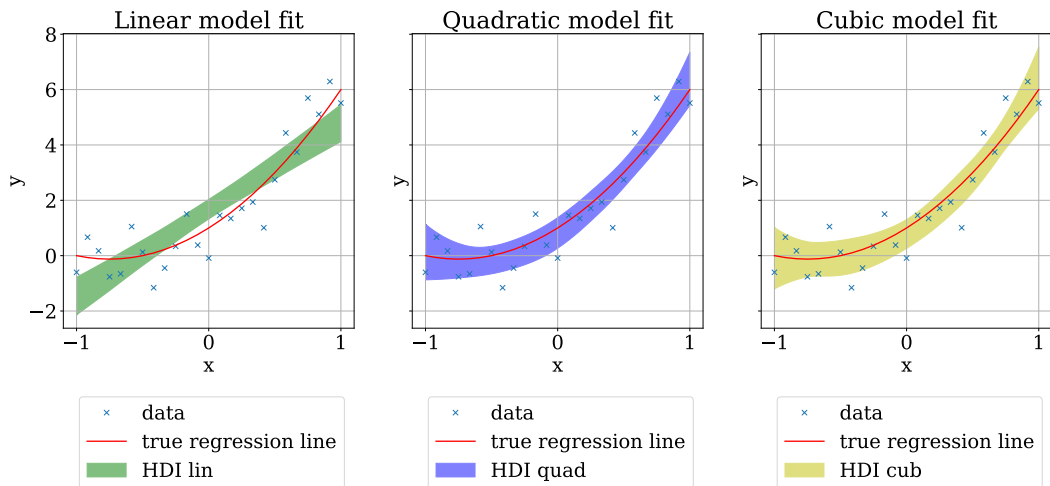


Figure 5: Result of parameter estimation with SMC. The data was generated with $\sigma = 0.7$

Fitting a polynomial of unknown degree

Comparison M_1 vs. M_2	$\ln(B_{12}(\sigma = 0.7))$
square vs. linear	8.5507 ± 0.053
cubic vs. linear	7.6225 ± 0.094
square vs. cubic	0.9371 ± 0.1093

Table 2: Results of BAYES factor via SMC

Comparison M_1 vs. M_2	$\ln(B_{12}(\sigma = 0.7))$
square vs. linear	$> 2.4301 \pm 0.27613$
square vs. cubic	0.8091 ± 0.0265

Table 3: Results of BAYES factor via SDDR.

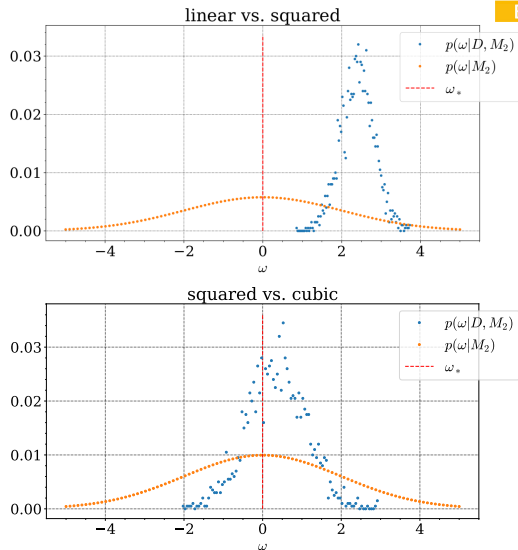


Figure 6: Computation of the SDDR ($\sigma = 0.7$)

Fitting a polynomial of unknown degree

What about the complexity?

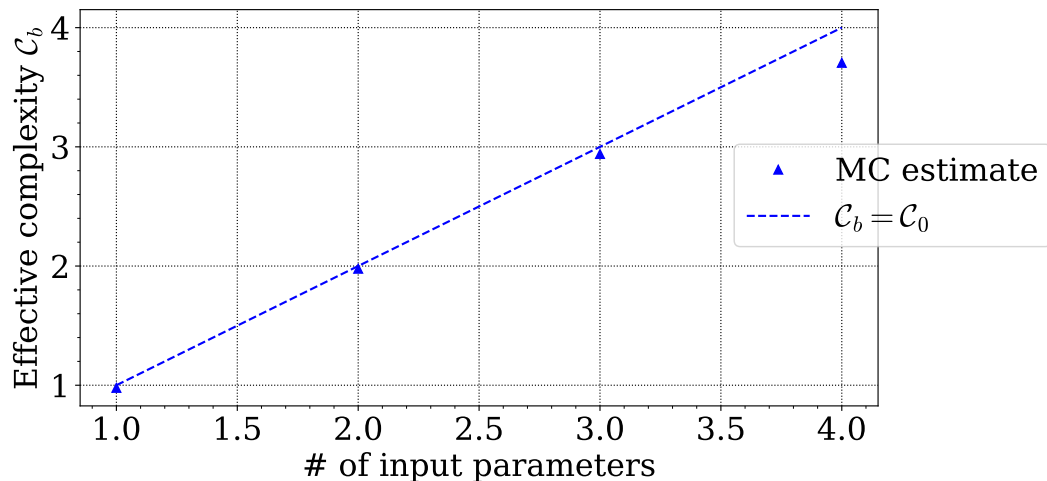
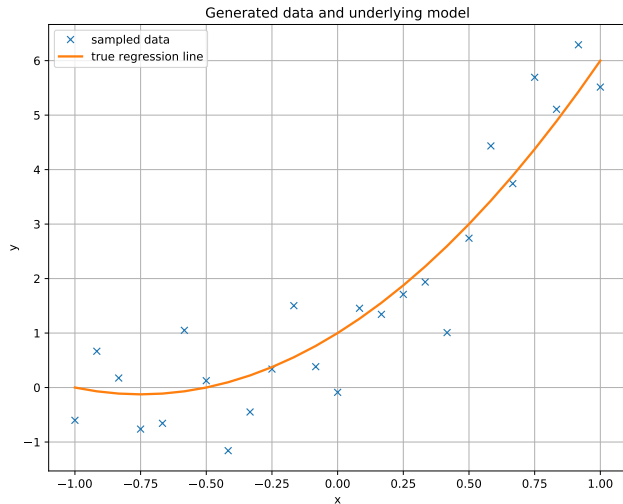


Figure 7: Numerical computation of the complexity \mathcal{C}_b , 3 parameters are supported.

Fitting a polynomial of unknown degree

And the true regression line is...



$$\begin{aligned} f(x; \theta) &= a \cdot x^2 + b \cdot x + c \\ &= 2 \cdot x^2 + 3 \cdot x + 1. \end{aligned}$$

(again, nice! ✓)

Figure 8: True regression line

1. Theory

Parameter estimation

Model comparison

2. Methods

Monte-Carlo-Sampling

SAVAGE-DICKEY-Density-Ratio (SDDR)

3. Examples

Coin Flip

Fitting a polynomial of unknown degree

4. Summary

Theory and methods

- ▶ BAYESIAN statistics provides the BAYES factor and BAYESIAN complexity as measures of *model comparison*
- ▶ Monte-Carlo-techniques can be used to compute both quantities

Examples

- ▶ We can say with weak to moderate confidence that the coin is biased
- ▶ We can say with weak to moderate evidence that the quadratic model is favoured over the others
- ▶ The BAYESIAN complexity diverges for $\dim \theta > 3$

Outlook

- ▶ Highly scalable
- ▶ physics, actually



GitHub.com

