Exploring Monte-Carlo-integration techniques in Bayesian model selection

Jakob Krause* and Dominic Schüchter†
(Dated: March 2, 2021)

An article usually includes an abstract, a concise summary of the work covered at length in the main body of the article.

Usage: Secondary publications and information retrieval purposes.

Structure: You may use the `description` environment to structure your abstract; use the optional argument of the `\item` command to give the category of each item.

I. INTRODUCTION

In physics, one is often faced with the problem of *Model Selection* for a given data set. That means finding a mathematical description of the data, which on the one hand sufficiently characterizes the data structure and on the other hand satisfies the expected dependencies. This is by no means a trivial task; one has to understand the underlying physical model beforehand to not make the mistake of choosing a too complicated model although it may seemingly fit the data. At the same time too trivial assumptions can also lead in the wrong direction. Compactly this problem can be formulated in the following way (adapted from [7, Chap. 4]):

Alice has a theory; Bob also has a theory, but with an adjustable parameter λ. Whose theory should we prefer on the basis of data D?

Bayesian inference provides quantitative measures for this model-selection problem, e.g. the Bayes-factor and the Bayes-complexity, these have among others been successfully used in astronomy as can be found in [5, 9, 10] respectively. In this paper we will investigate two simulated example problems and apply various measures of bayesian model selection as to find out the true underlying model which was used to generate the simulated data. We will focus on the numerical evaluation of such problems, especially Monte-Carlo techniques.

II. THEORY

In the following we will give a short introduction into Bayes theorem and the underlying concepts of parameter estimation and model selection.

A. Bayes' Theorem

The fundamental equation of Bayesian statistics – for a dataset y and parameters $\boldsymbol{\theta}$ – is given by Bayes Theorem.

$$\text{prob}(\boldsymbol{\theta}|y) = p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(y)} \tag{1}$$

Where $p(\boldsymbol{\theta}|y)$ is the posterior probability for parameters $\boldsymbol{\theta}$ given the data y, $p(y|\boldsymbol{\theta})$ is the likelihood that the data fits a model with parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$ is the prior probability of $\boldsymbol{\theta}$ and $p(y) = \int_{-\infty}^{+\infty} \mathrm{d}\boldsymbol{\theta} p(y|\boldsymbol{\theta}) p(\boldsymbol{\theta})$ is the marginal likelihood which acts as a normalization. In the case of parameter selection, the normalization can often be neglected since it is only a constant. In the case of the model comparison it is a crucial quantity, as we will discuss in section (II C) [7, Chap. 2].

B. Parameter estimation

Assume we want to find the best parameters for a given dataset y and model M. Eq. (1) then can be written as

$$p(\boldsymbol{\theta}|y, M) = \frac{p(y|\boldsymbol{\theta}, M) \cdot p(\boldsymbol{\theta}|M)}{p(y|M)}. \tag{2}$$

Evaluating this equation will give probability density functions (PDFs)

$$p(\theta_i|y, M) = \int p(\boldsymbol{\theta}|y, M) \prod_{j \neq i} \mathrm{d}\theta_j \tag{3}$$

for each parameter θ_i. Because of equation (3) $p(\theta|D, M)$ is also called *marginal posterior*. From this one can find the best fit value either by finding the value of θ for which the marginal posterior is maximized or by calculating the mean with respect to the marginal posterior.

$$\begin{aligned} p(\theta|y, M) = \max &\Leftrightarrow \theta = \hat{\theta} \\ \langle\theta\rangle &= \int_{-\infty}^{\infty} \mathrm{d}\theta p(\theta|y, M) \cdot \theta \end{aligned} \tag{4}$$

Throughout this paper we will use $\langle\theta\rangle$ as our best fit estimate. [7]

C. Model comparison

To compare two (or more) competing models M_i that describe a dataset D let us write Bayes theorem once

* http://www.github.com/krausejm; krause@hiskp.uni-bonn.de
† http://www.github.com/dschuechter; dschuechter@uni-bonn.de

again

$$p(M_i|y) = \frac{p(M_i) \cdot p(y|M_i)}{p(y)}. \tag{5}$$

We recognize $p(y|M_i)$ as the marginal likelihood of the model M_i. Let now the two competing models be those of Alice and Bob introduced in section (I).Thus,

$$p(y|M_i) = \int_{-\infty}^{\infty} \mathrm{d}\lambda p(y|\lambda, M) \cdot p(\lambda|M)$$

follows [7, Chap. 3].

1. *Bayes Factor*

One way of comparing two models is to compare their posterior odds, that is the ratio quote lecture A. Wirzba?

$$O_{ij} := \underbrace{\frac{p(M_i|y)}{p(M_j|y)}}_{\text{posterior odds}} = \underbrace{\frac{p(y|M_i)}{p(y|M_j)}}_{\text{Bayes Factor}} \cdot \underbrace{\frac{p(M_i)}{p(M_j)}}_{\text{prior odds}} = B_{ij} \cdot \frac{p(M_i)}{p(M_j)}. \tag{6}$$

Usually we can assume the prior odds to be ~ 1 because we do not favor one model over the other just on prior beliefs. With this equation (6) simply becomes

$$O_{ij} = B_{ij}.$$

Calculating B_{ij} will be referred to as calculating the Bayes factor *in favor of* model M_i. We can then evaluate the result of B_{ij} as follows

$$B_{ij} = B_{ji}^{-1} = \begin{cases} \geq 1 & \text{model } M_i \text{ favored} \\ < 1 & \text{model } M_j \text{ favored} \end{cases}$$

Nevertheless a Bayes Factor ≥ 1 does not mean that the favored model is far superior. There is an empirical scale which points to the strength of evidence for a given B_{ij}. This is displayed in table (I).

$\lvert \ln B_{ij} \rvert$	Odds	Probability	Strength of evidence
< 1.0	$\lesssim 3:1$	< 0.750	Inconclusive
1.0	$\sim 3:1$	0.750	Weak evidence
2.5	$\sim 12:1$	0.923	Moderate evidence
5.0	$\sim 150:1$	0.993	Strong evidence

TABLE I: Empirical scale for evaluating the strength of evidence when comparing two models M_i vs. M_j, adapted from [10]

Bayes factor vor allem attraktiv für genestete modelle

2. *Bayes Complexity*

When comparing models with a different amount of parameters, we obviously denote more parameters as more complex. A fundamental question of model-selection is then if the given data supports more or less parameters. A naïve measure of complexity would be the number of free parameters a model has, we denote this as $\mathcal{C}_0$. A more sophisticated measure of complexity is given by the so called Bayesian complexity, first introduced by Spiegelhalter et al [8], which can be written as [5]

$$\mathcal{C}_b = -2 \int \mathrm{d}\boldsymbol{\theta} p(\boldsymbol{\theta}|y, M) \log(\mathcal{L}(\boldsymbol{\theta})) + 2 \log(\mathcal{L}(\tilde{\boldsymbol{\theta}})), \tag{7}$$

with the likelihood $\mathcal{L}(\boldsymbol{\theta}) = p(D|\boldsymbol{\theta}, M)$ and $\tilde{\boldsymbol{\theta}} = \langle \boldsymbol{\theta} \rangle$. $\mathcal{C}_b$ describes how many model parameters the data is able to constrain [5] and is thus a useful tool for examining models with an increasing number of parameters. If we define a χ^2 as $\mathcal{L}(\boldsymbol{\theta}) \propto \exp(-\chi^2/2)$, we can write

$$\mathcal{C}_b = \overline{\chi^2(\boldsymbol{\theta})} - \chi^2(\tilde{\boldsymbol{\theta}}), \tag{8}$$

where $\overline{\chi}$ denotes the mean taken over the posterior PDF. The definition of $\mathcal{C}_b$ is chosen such that $\mathcal{C}_b \to \mathcal{C}_0$ for highly informative data [5].

III. METHODS

We will now describe the functions of the used algorithms for the model selection. Since the implementation of the so called nested sampling is sufficiently dealt with by Gruppe - Bayesian parameter fitting our implementation uses the `PyMC3`-Python Library [2]. `PyMC3` offers simple solutions to create models and a wide variety of sampling algorithms. Furthermore the `ArviZ`-Library provides methods for posterior analysis and visualization [1].

A. Monte-Carlo Sampling

First, the goal of Monte-Carlo sampling is discussed. Often equations (3), (7) and (8) have no closed analytical solutions, so we are left with a numerical ansatz (or analytical approximations) [11]. One possibility is given by Markov-Chain-Monte-Carlo Sampling (MCMC) [?]. The main idea is to generate a chain of parameter samples $\boldsymbol{\theta}^{(t)}$ that follow the posterior PDF (2). For such a chain the mean with respect to the posterior is given by

$$\langle \boldsymbol{\theta} \rangle \approx \int p(\boldsymbol{\theta}|y) \boldsymbol{\theta} d\boldsymbol{\theta} = \frac{1}{N} \sum_{t=0}^{N-1} \boldsymbol{\theta}^{(t)}, \tag{9}$$

since the samples follow the PDF $p(\boldsymbol{\theta}|y)$. Analogously the expectation value of any function with respect to the posterior is

$$\langle f(\boldsymbol{\theta}) \rangle \approx \frac{1}{N} \sum_{t=0}^{N-1} f(\boldsymbol{\theta}^{(t)}). \tag{10}$$

Having established a Markov chain, one can compute the marginal posterior (3) by binning θ_i in the given parameter range and ignoring all other parameters [10].

B. Metropolis-Hastings Algorithm

Because the Metropolis-Hastings-algorithm is one of the fundamentals of understanding SMC, see section (III C), we briefly state it here. The main characterization of a Markov chain – which consists of random parameters $\boldsymbol{\theta}^{(t)}$ – is that each element is only determined by the previous element. In our case we want to sample in the parameter space according to the distribution $p(\boldsymbol{\theta}|y, M) := p(\boldsymbol{\theta})$ (2). We can achieve this by randomly proposing a new vector in parameter space $\boldsymbol{\theta'}$ according to a arbitrary proposal distribution $p_p(\boldsymbol{a}|\boldsymbol{b})$ and accept it with a probability

$$A(\boldsymbol{\theta'}, \boldsymbol{\theta}^{(t)}) = \min\left(1, \frac{p(\boldsymbol{\theta'})p_p(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta'})}{p(\boldsymbol{\theta}^{(t)})p_p(\boldsymbol{\theta'}|\boldsymbol{\theta}^{(t)})}\right).$$

This ensures the sampled values follow the desired PDF [11]. This is only a very brief overview of MCMC. More information on the underlying theory can be found e.g. in [6]

C. Sequential Monte Carlo (SMC)

For our purposes we chose the *Sequential Monte Carlo* sampling algorithm provided by `PyMC3` [3], because it grants easy access to the marginal likelihood. This in turn is needed for model comparison via the Bayes-factor.

Since the SMC algorithm joins several statistical concepts, including *importance sampling*, *tempering* and and MCMC kernel (Metropolis-Hastings) [3], it is a highly non-trivial algorithm. Thus, a detailed description is beyond the scope of this paper. We will now sketch the main idea of the algorithm. First let us introduce an auxiliary *parameter* β and write equation (2) as

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

with $Z_\beta = \int \mathrm{d}\boldsymbol{\theta} p(y|\boldsymbol{\theta})^\beta \cdot p(\boldsymbol{\theta})$. For $\beta = 1$ we get the same equation as before, that is $p(\boldsymbol{\theta}|y)_\beta = p(\boldsymbol{\theta}|y)$. The idea is to gradually sample from $\beta = 0$ to $\beta = 1$ using β to control the transition from an easy to sample distribution to a harder one. The final result is a collection of samples from the true posterior [3]. We can then estimate the marginal likelihood as [4]

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

The hat denotes, that this is a numerical estimation of $p(y)$ [3]. Using SMC we can in one run do parameter inference and also compare different models using the Bayes-factor (6).

D. Savage Dickey Density Ration (SDDR)

Now we introduce an alternative way of computing the Bayes-factor (6) for *nested* models; 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 furhter assume seperable priors (which is usually the case [9])

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

We can then write the Bayes-factor as [9]

$$B_{ij} = B_{ji}^{-1} = \left.\frac{p(\omega|y, M_1)}{p(\omega|M_1)}\right|_{\omega=\omega_\star} \text{(SDDR)}. \tag{11}$$

IV. EXAMPLES

A. Betabinomial example (coin flip)

Let us now consider as a starting example, the flipping of a two-sided coin, i.e. an experiment where we can measure either heads (H) or tails (T) with 50% probability, respectively. This, while simple, allows us an intuitive approach to Bayesian inference and model selection as well as to the MCMC techniques discussed before. Furthermore is this example easily altered to many real-life problems, such as birth rates, ..., or anything with the option of either success or failure.

1. Analytical approach?

Assume we throw a coin 20 times. We observe 6 H and 14 T. "Is this a fair coin?" might be a question to ask yourself since the bias in outcome is quite large. Naively expecting a fair coin we could assign a *prior* to the probability of heads θ as centred around 0.5, so for example a gaussian with $\mu = 0.5, \sigma = 0.1$.

2. Numerical approach

B. Fitting a polynomial of unknown degree

1. Analytical approach?

2. Numerical approach

V. DISCUSSION

VI. SUMMARY

[1] Arviz: Exploratory analysis of bayesian models — arviz dev documentation. URL https://arviz-devs.github.io/arviz/. last visit: March 2, 2021.

[2] Pymc3 documentation — pymc3 3.10.0 documentation, . URL https://docs.pymc.io/. last visit: March 2, 2021.

[3] Pymc3 documentation — sequential monte carlo, . URL https://docs.pymc.io/notebooks/SMC2_gaussians.html. last visit: March 2, 2021.

[4] D. Gunawan, K.-D. Dang, M. Quiroz, R. Kohn, and M.-N. Tran. Subsampling sequential monte carlo for static bayesian models. 2020. URL https://arxiv.org/pdf/1805.03317.pdf. last visit: March 2, 2021.

[5] M. Kunz, R. Trotta, and D. R. Parkinson. Measuring the effective complexity of cosmological models. *Phys. Rev. D*, 74:023503, Jul 2006. doi: 10.1103/PhysRevD.74.023503. URL https://link.aps.org/doi/10.1103/PhysRevD.74.023503.

[6] R. M. Neal. Probabilistic inference using markov chain monte carlo methods. *Technical Report CRG-TR-93-1*.

[7] D. Sivia and J. Skilling. *Data Analysis - A Bayesian tutorial*, volume 2. Oxford University Press, 2006.

[8] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4):583–639, 2002. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/3088806.

[9] R. Trotta. Applications of Bayesian model selection to cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, 378(1):72–82, 05 2007. ISSN 0035-8711. doi:10.1111/j.1365-2966.2007.11738.x. URL https://doi.org/10.1111/j.1365-2966.2007.11738.x.

[10] R. Trotta. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49(2):71–104, Mar 2008. ISSN 1366-5812. doi: 10.1080/00107510802066753. URL http://dx.doi.org/10.1080/00107510802066753.

[11] U. von Toussaint. Bayesian inference in physics. *Rev. Mod. Phys.*, 83:943–999, Sep 2011. doi: 10.1103/RevModPhys.83.943. URL https://link.aps.org/doi/10.1103/RevModPhys.83.943.