TIF285 Learning from data, Project 1: Parameter estimation for a toy model of an effective field theory

Tomas Lundberg and Jonas H. Fritz*
Chalmers University of Technology
(Dated: December 8, 2021)

Based on data from [1], we estimate low energy constants of an effective field theory using Bayesian statistics. In particular, we compare the usage of different priors for the uncertainty of the parameter estimation and fitting to the experimental data. This is done by reproducing results obtained in [1], namely Figures 1,3 and 4, as well as Table III. In addition, we generate own data sets of different size and with different noise to test model performance.

Contents

I. Introduction

Bayesian parameter estimation is a powerful tool which, a priori, incorporates theoretical knowledge about your parameters. Furthermore, it offers a clear probabilistic interpretation in how to compare models with different parameter estimations. We will illustrate these main features of the Bayesian approach to parameter estimation in this report by considering the following function, which represents an effective field theory

$$g(x) = \left(\frac{1}{2} + \tan\left(\frac{\pi}{2}x\right)\right)^2. \tag{1}$$

This function can be Taylor expanded up to order k as

$$g_{\text{th}}(x) \equiv \sum_{i=0}^{k} a_i x^i. \tag{2}$$

While the underlying physics is interesting, we will only concern ourselves with the task of estimating the a_i's up to some truncation order k. If k is sufficiently small, then these a_i's are the so called low-energy constants. This is done using Bayesian parameter estimation with Markov Chain Monte Carlo sampling (MCMC). This gives us a joint probability density function (pdf) for all the a_i's. By marginalization to each possible tuple of parameters we can visualize the joint pdf in two dimensions. This is a reproduction of Figure 1 in [1]. In addition, we are interested in the impact the use of different priors will have. For that we consider both a flat and a Gaussian naturalness prior.

Once the parameters are estimated, we can use these to plot the estimated function with its 1σ intervals along with the true function Eq. (1) and the underlying data, which reproduces Figure 3 and 4 from [1]. [1] use randomly generated data for their modelling. Each data point d_j is given by

$$d_j = g(x_j)(1 + c\eta_j) \quad \Longrightarrow \quad \sigma_j = c d_j, \tag{3}$$

where η_j is drawn from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ and c is an arbitrarily chosen relative error. To make our results comparable to [1], we make use of the same data set used in that paper. This was data set $\text{D1}_{5\%}$ with 10 data points and $c = 0.05$. Further, we calculate the evidence using Laplace's approximation for a χ^2 distribution. Finally, we will use Eq. (3) to generate new data sets and evaluate how the uncertainty in the parameter estimation is affected by different data set sizes N and different values of the relative error c for the different priors.

* lutomas@student.chalmers.se, jonasher@student.chalmers.se

II. Method

In order to estimate the coefficients a_i in Eq. (2) we will use a Bayesian parameter estimation approach. In this section we go through the basics of the Bayesian framework, such as stating Bayes' Theorem, defining the likelihood function and different priors. We will also explain how to calculate the evidence and what it can tell us about model selection. Further, we briefly discuss the method of marginalization and how we will use it. We obtain the posterior pdf for the coefficients by using Markov Chain Monte Carlo (MCMC) sampling of which we will also explain the fundamentals.

A. Bayes' theorem

The *posterior* pdf for the parameters $\mathbf{a}$ given the data D and prior knowledge I is obtained through Bayes' theorem

$$\mathrm{p}(\mathbf{a} \mid D,I) = \frac{\mathrm{p}(D \mid \mathbf{a},I)\,\mathrm{p}(\mathbf{a} \mid I)}{\mathrm{p}(D \mid I)}, \tag{4}$$

where $\mathrm{p}(D \mid \mathbf{a},I)$ is the *likelihood* of the data D given the coefficients $\mathbf{a}$ and our prior knowledge I, $\mathrm{p}(\mathbf{a} \mid I)$ the *prior* of $\mathbf{a}$ given the prior knowledge I and $\mathrm{p}(D \mid I)$ the *evidence* of the data D given the prior knowledge I. We note that the evidence is simply a normalization factor for our posterior, and hence

$$\mathrm{p}(\mathbf{a} \mid D,I) \propto \mathrm{p}(D \mid \mathbf{a},I)\,\mathrm{p}(\mathbf{a} \mid I). \tag{5}$$

Furthermore, to achieve higher numerical stability, we will calculate the log of the posterior through

$$\log \mathrm{p}(\mathbf{a} \mid D,I) \propto \log \mathrm{p}(D \mid \mathbf{a},I) + \log \mathrm{p}(\mathbf{a} \mid I). \tag{6}$$

[1]

1. Priors

We follow [1] and use two different priors for our coefficients $\mathbf{a}$; a flat prior and a Gaussian naturalness prior. The flat prior is defined as

$$\mathrm{p}_{\mathrm{flat}}(\mathbf{a} \mid I) = \begin{cases} 1 & \text{if } |a_i| > \tilde{a} \ \ \forall i \in \{0,1,\ldots,k\} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\tilde{a}$ defines the limit of the flat prior. We use $\tilde{a} = 10^6$ throughout our analysis. This might seem unnecessarily large but in order to reproduce the results in Table III of [1] this $\tilde{a}$ was required for the higher orders of the flat prior. Furthermore, it did not affect the estimations of the lower order parameters. The Gaussian naturalness prior is defined as

$$\mathrm{p}_{\mathrm{Gaussian}}(\mathbf{a} \mid I) = \left(\frac{1}{\sqrt{2\pi}\bar{a}}\right)^{k+1} \exp\left(-\frac{\mathbf{a}^2}{2\bar{a}^2}\right), \tag{8}$$

where we use a fixed $\bar{a} = 5$ throughout our analysis.

2. Likelihood

The likelihood function for our data $D = \{d_j\}$ with corresponding error σ_j, as defined in Eq. (3), where $|D| = N_d$ is given by

$$\mathrm{p}(D \mid \mathbf{a},I) = \prod_{j=1}^{N_d} \left(\frac{1}{\sqrt{2\pi}\sigma_j}\right) e^{-\chi^2/2}, \tag{9}$$

where the χ^2 distribution is defined as

$$\chi^2 = \sum_{i=1}^{N_d} \left(\frac{d_i - g_{\mathrm{th}}(x_i)}{\sigma_i}\right)^2. \tag{10}$$

Here, $g_{\mathrm{th}}(x_i)$ denotes the model value at x_i from Eq. (2) given the coefficients $\mathbf{a}$. [1]

3. Calculating the evidence - Laplace's method

The denominator $\mathrm{p}(D \mid I)$ in Bayes' theorem, Eq. (4), is called the evidence. As mentioned, it does not affect our posterior pdf since it just shows up as a normalization factor. However, calculating the evidence for our models with different truncation order k is useful for comparing the models. If, for instance, the ratio of the evidence with truncation order k and $k+1$

$$\frac{\mathrm{p}(D \mid k,I)}{\mathrm{p}(D \mid k+1,I)}, \tag{11}$$

is $<< 1$ then the larger truncation order $k+1$ has a high probability of being the better model and vice versa.

Calculating the evidence can be a precarious endeavour, but when dealing with χ^2 distributions one can use Laplace's approximation [2]. Let's denote the evidence in Bayes' theorem as the normalizing constant Z_p of our unnormalized posterior pdf $\mathrm{p}^*(\mathbf{a})$, with $|\mathbf{a}| = k$, then

$$Z_p = \int \mathrm{p}^*(\mathbf{a}) d\mathbf{a}. \tag{12}$$

If we in particular consider a χ^2 probability distribution function, as is the case in this report; $\mathrm{p}^*(\mathbf{a}) = \exp\left(-\frac{1}{2}\chi^2(\mathbf{a})\right)$, then we obtain for our best estimates $\mathbf{a}_0$ the evidence as

$$Z_p \approx \exp\left(-\frac{1}{2}\chi^2(\mathbf{a}_0)\right) \sqrt{\frac{(4\pi)^k}{\det \Sigma^{-1}}}, \tag{13}$$

where Σ^{-1} is the inverse of the covariance matrix for our sampled posterior pdf over $\mathbf{a}$. This is Laplace's approximation for calculating the evidence.

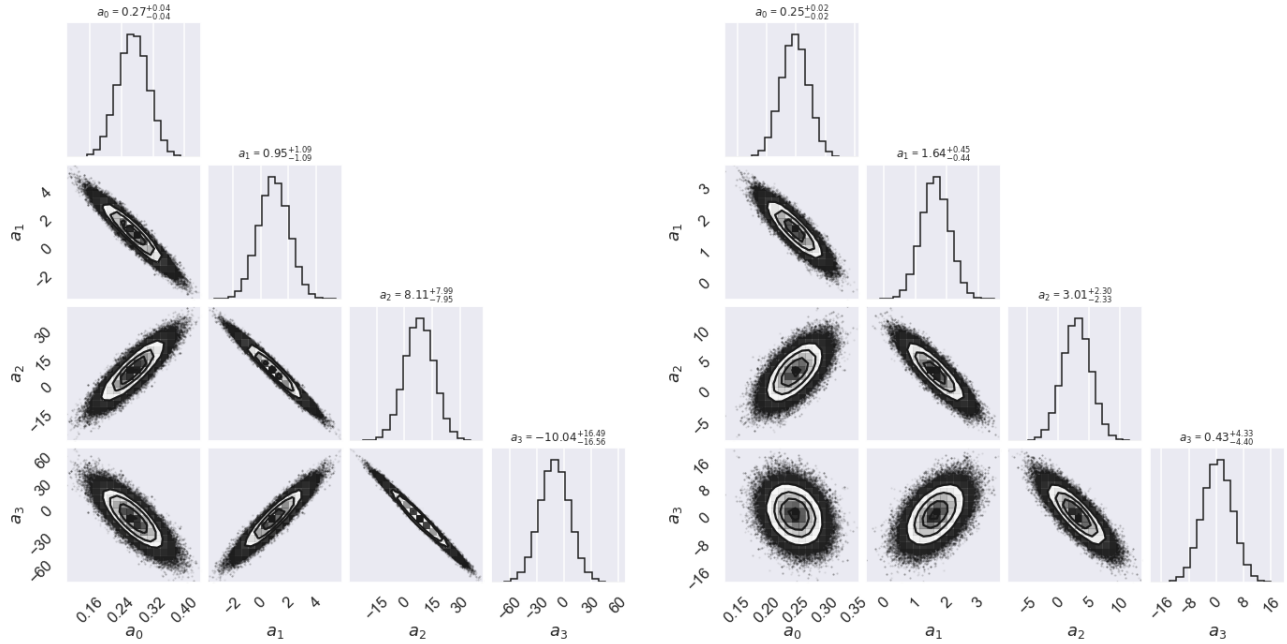

FIG. 1: Corner plot of the marginalized joint probabilities for the parameters a_i up to third order. The left corner plot shows the use of a flat prior and the right one shows the use of a Gaussian prior. For the flat prior, the joint pdfs are thinner, corresponding to a stronger (anti-)correlation between the parameters. This figure shows excellent agreement with Figure 1 of [1].

B. Marginalization

Marginalization is a method one can use to obtain the posterior pdf over a subset of the parameter space given a posterior pdf over the full parameter space. If we, for instance, are only interested in the posterior pdf for a_j given $\mathrm{p}(\mathbf{a} \mid D, I)$ with parameters $\mathbf{a} = \{a_i\}$ with $i \in \{0, 1, \ldots, k\}$ we can retrieve the relevant posterior through

$$\mathrm{p}(a_j \mid D, I) = \int \prod_{i \neq j}^{k} d_i \, \mathrm{p}(\mathbf{a} \mid D, I). \tag{14}$$

This is done on the diagonal in Fig. 1. Similarly, we can obtain the joint posterior for two parameters a_j and a_m to evaluate their correlation as is done off-diagonal in Fig. 1.

Moreover, we can use marginalization to account for higher order terms to avoid underfitting while only extracting the lower order terms. That is, we marginalize parameters from $k+1$ to $k_{\max}$ and obtain the posterior for parameters $\mathbf{a} = \{a_i\}$ with $i \in \{0, 1, \ldots, k\}$. This is done in Table I.[1]

C. Markov Chain Monte Carlo sampling

Markov Chain Monte Carlo (MCMC) sampling is a technique that can be used to solve integrals over a large multidimensional space. In our case, we use it to draw samples distributed according to our posterior pdf. Instead of integrating over all k parameters from $-\tilde{a}$ to $+\tilde{a}$ in the case of the flat prior or from $-\infty$ to $+\infty$ as in the case of the Gaussian prior we use MCMC which performs a random walk over our posterior pdf that produces samples that are proportional to our posterior pdf. [3]

In order to obtain a reliable estimation of our posterior probability distribution we take n_{samples} samples from n_{walkers} different random walks. Furthermore, we discard the first n_{warmup} samples such that each random walk has had time to stabilize.

III. Results and discussion

In this section we discuss how we reproduced the results in [1], that is, Figure 1, 3 and 4 as well as Table *III*. We also offer an interpretation of the result where we in particular compare the outcomes of the different priors. Finally, we discuss how data sets of different sizes and with different magnitudes of error effect the model performance.

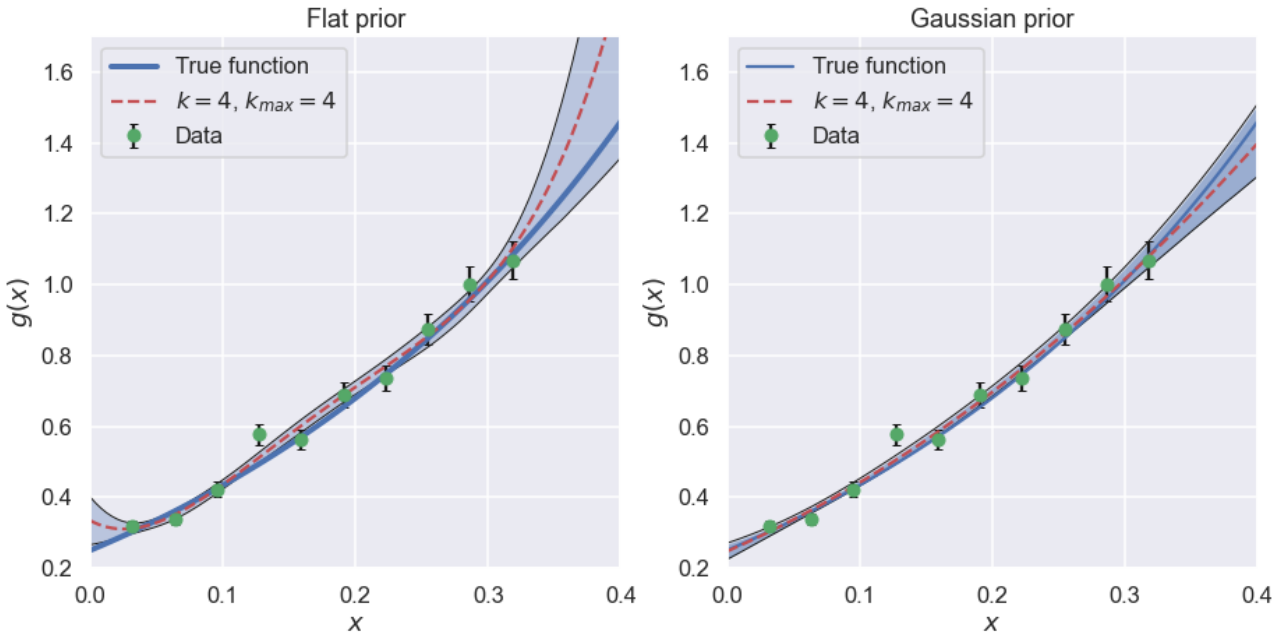

FIG. 2: Eq. (1) along with the generated data and the estimated function using a flat prior (left) and a Gaussian prior (right) with corresponding 1σ DoB intervals. These results are in good agreement with Figure 3 and Figure 4 of [1]

A. Projected posterior plots

We use MCMC with $n_{\text{walkers}} = 50$, $n_{\text{warmup}} = 2000$ and $n_{\text{samples}} = 10000$ to sample the joint posterior pdf of the a_i's up to third order, $k = 3$. By marginalizing down to two parameters we can plot all the pairwise joint pdf's for the parameters in a corner plot. Doing this both for a flat and for a Gaussian prior results in Fig. 1, which shows excellent agreement with Figure 1 from [1].

We see that all single parameter marginalized distributions are Gaussian, which follows from the fact that multiplying a Gaussian likelihood Eq. (9) with a flat prior results in a truncated Gaussian and multiplying it with a Gaussian prior also results in a Gaussian. For the flat prior, the joint pdf's are thinner, corresponding to a stronger (anti-)correlation between the parameters. If say, one of the higher order parameters takes a large value with high uncertainty, lower order parameters will have to compensate for that, explaining the correlation. Note for instance in Fig. 1 for the flat prior how the rather large value of a_2 is compensated for the large negative value of a_3. We will see further on that this is a clear example of overfitting. With a Gaussian prior the parameter values are restricted, decreasing the need for this kind of compensation and therefore also decreasing correlation. The Gaussian prior also decreases the uncertainty for each parameter.

B. Comparing the true model to the data

Taking the same approach as in producing Fig. 1, we again estimate coefficients with MCMC sampling using the same settings as before, but now for order $k = 4$. In Fig. 2, we use these coefficients to plot Eq. (2), together with the corresponding 1σ degree-of-belief (DoB) intervals, the data and Eq. (2) with the true expansion coefficients to the same order. Again, this shows excellent agreement with Figure 3 and 4 from [1].

For both priors we can see that the DoB intervals are tight where there is data. However, where there is now data, the DoB intervals widens, which is much more pronounced for the flat prior. This follows from the fact that the Gaussian prior restricts the parameter range, thus also restricting the range of possible functions and the DoB intervals. For the flat prior the opposite is true: since the parameters can take on higher values, the function can vary a lot more. This is only restricted by the presence of data points, which clearly define where the function has to lie. We can also note, if we look close, that the best estimate for the flat prior is drawn to the data points to a higher degree than for the Gaussian prior. This is a good example of overfitting which becomes even clearer in the next section when we compare the estimated values to the true values. We conclude that a Gaussian prior is much better suited to make predictions from a limited amount of data.

TABLE I: Estimates of coefficients up to second order, $k = 2$ for expansion up to sixth order, $k_{\max} = 6$, using marginalization, along with 68% DoB. The χ^2/dof and evidence for the flat and Gaussian prior respectively are also included. The true values are shown on the last row. Again we find good agreement with the results of Table III in [1].

		Flat prior				Gaussian prior			
k	$k_{\max}$	χ^2/dof	a_0	a_1	a_2	Evidence	a_0	a_1	a_2
0	0	67	0.48 ± 0.01	-	-	~ 0	0.48 ± 0.01	-	-
1	1	2.2	0.2 ± 0.01	2.6 ± 0.1	-	6e+02	0.2 ± 0.01	2.6 ± 0.1	-
2	2	1.6	0.25 ± 0.02	1.6 ± 0.4	3.3 ± 1.3	3.2e+03	0.25 ± 0.02	1.7 ± 0.4	3 ± 1
2	3	1.9	0.27 ± 0.04	0.95 ± 1	8.3 ± 8.1	2.7e+03	0.25 ± 0.02	1.7 ± 0.5	2.9 ± 2
2	4	2	0.33 ± 0.07	-2 ± 3	46 ± 34	2.6e+03	0.25 ± 0.02	1.6 ± 0.4	3 ± 2
2	5	1.6	0.57 ± 0.13	-15 ± 7	280 ± 120	2.7e+03	0.25 ± 0.02	1.6 ± 0.5	3.1 ± 2
2	6	2.2	0.6 ± 0.29	-17 ± 20	330 ± 400	2.7e+03	0.25 ± 0.02	1.7 ± 0.4	2.9 ± 2
True values			0.25	1.57	2.47		0.25	1.57	2.47

C. Coefficient estimates at higher orders

Further, we are interested in how the parameter estimation improves when expanding to higher orders. We do the expansion up to order $k_{\max} = 6$ and extract the parameters up to order $k = 2$ using marginalization. The sampled posterior pdf's from where we take the median (or equivalently the mean since we have a Gaussian posterior pdf) as the best estimate and the 1σ DoB for the coefficients a_0, a_1 and a_2 are obtained through MCMC sampling using the same settings as before. For the flat prior, we also calculate for each order the cost function per degree of freedom, χ^2/dof, where the degrees of freedom are the number of data points subtracted by the order k. For the Gaussian prior we calculate the evidence using Laplace's approximation.

The results are shown in Table I Which is in strong agreement with Table III from [1]. The flat and Gaussian prior display roughly equivalent results up to order $k_{\max} = 2$. At higher orders, the parameter uncertainty for the flat prior rapidly increases. This is due to overfitting, where the parameters will take on extreme values to minimize the cost function. In contrast, the Gaussian prior limits the sudden growth of the parameters and thus the uncertainty for higher orders. This is in agreement with what we find in Fig. 2, where the confidence interval is smaller for the Gaussian prior. χ^2/dof greatly decreases when going from $k_{\max} = 0$ to $k_{\max} = 1$ and only fluctuates thereafter. While χ^2 decreases with higher orders, so do the degrees of freedom, such that their ratio remains roughly constant. When also taking the parameter uncertainty into consideration, we can conclude that $k_{\max} = 2$ is the best degree for fitting the data with the flat prior. The same is true for the Gaussian prior, where the evidence is greatest at $k_{\max} = 2$. This means, as mentioned in Section II A 3, that the model with $k_{\max} = 2$ has the highest probability of being the best one given the data. Contrary to the flat prior, the parameter uncertainty does not increase for the Gaussian prior when adding higher order terms and the best estimates remains more or less the same. To reiterate from previous sections, this is due to the fact that the possibility of overfitting is reduced when using the Gaussian prior since the parameter space is restricted around zero.

D. Evaluating the dependence of data set size and noise on accuracy

Finally, we want to consider the influence of the data set size N and the measurement uncertainty parameter c, as defined in Eq. (3), on the parameter uncertainty σ. This is particularly interesting from an experimentalist's perspective, as this can serve as guide whether improving measurement precision or performing additional measurements is more beneficial. This was done for order $k = 2$ with 15×15 generated data sets with Eq. (3) using logarithmically spaced values of c between 0.01 and 1 and N between 5 and roughly 300. Due to the computationally expensive task of performing 15×15 MCMC samplings for the two different priors we reduced our settings for the MCMC to $n_{\text{walkers}} = 50$, $n_{\text{warmup}} = 1000$ and $n_{\text{samples}} = 2000$.

The results are shown in Fig. 3. We can clearly see that a big data set can make up for bigger noise, and that conversely, precise measurements can compensate a low number of measurements. Unsurprisingly, having both big data sets and precise measurements is best for minimizing parameter uncertainty. We can also note similar patterns as from previous sections where the Gaussian prior results in smaller errors for all parameters since the contour lines are not as tight as for the flat prior. This is especially true for a_2. Other than that, the patterns for the dependence of N and c for the different priors are very similar.

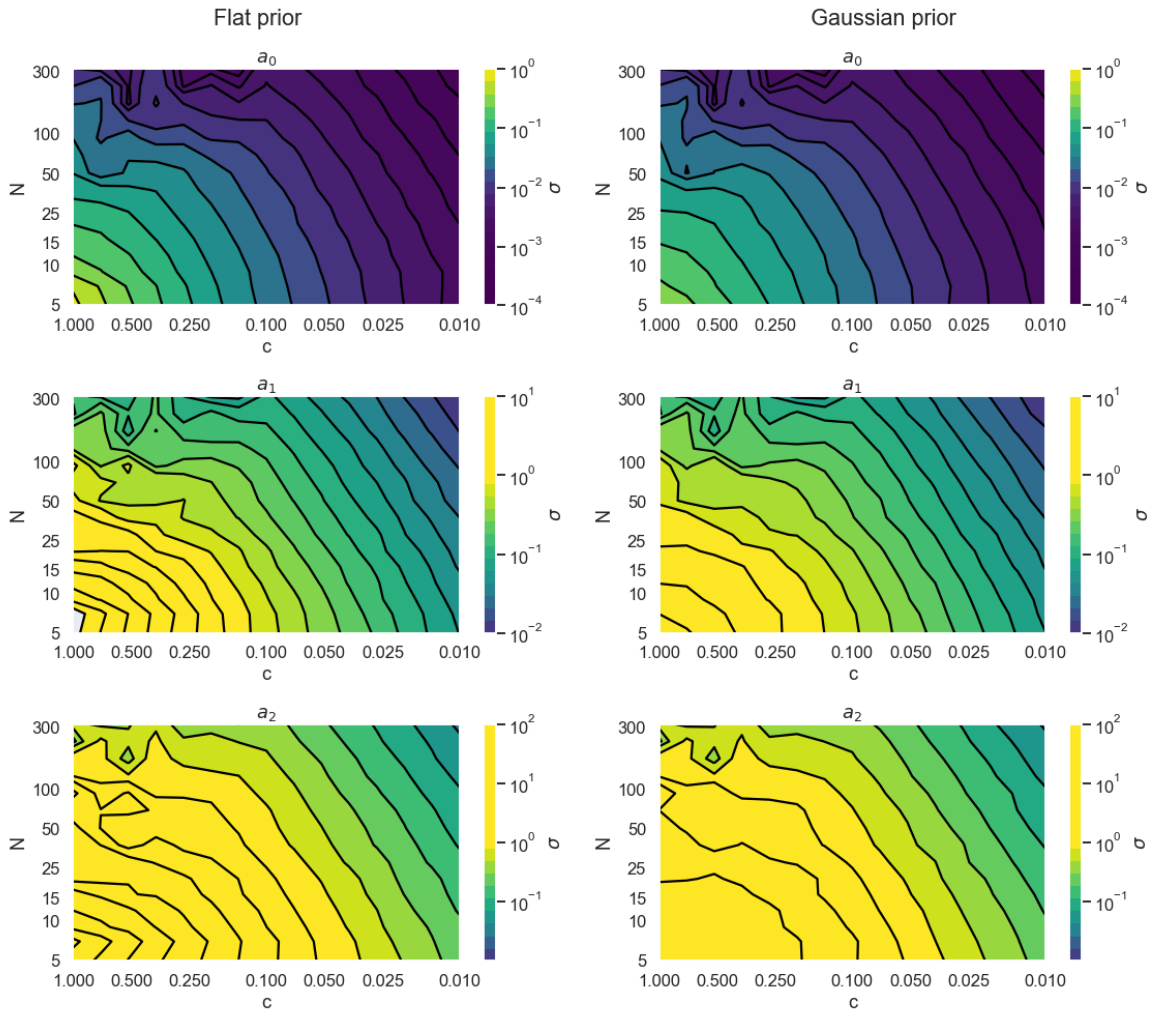

FIG. 3: Colormap of the uncertainty σ associated with the parameters a_0, a_1 and a_2 in dependece of data set size N and noise amplitude c. Note that the axis for the c values is inverted since having low values result in smaller values for σ. The opposite is true for N. Also note that both axis are log scaled. A large data set can make up for large noise and conversely, a small dataset can be compensated by low measurement uncertainty.

IV. Conclusion

We have used Bayesian parameter estimation to find Taylor expansion coefficients of an effective field theory based on data from [1]. In particular, we have studied the effect the use of two different priors, a flat and a Gaussian, has on this estimation. To sample the pdf of the parameters, called the posterior pdf in the Bayesian framework, we have used Markov Chain Monte Carlo (MCMC) sampling. Our results are reproductions of the results obtained in [1], with which we find excellent agreement throughout.

We have also applied our model to newly generated datasets of varying size and noise. We find that larger data sets can compensate for a increase in noise and vice versa.

Throughout our analysis we find that the Gaussian prior performs better, in the sense that it decreases the uncertainty in parameter estimation and increases the models ability to predict the functions behavior where there is no underlying data. The Gaussian prior incorporates the prior knowledge that the lower order expansion coefficients should be of order 1 [4] into the model. This highlights the strength of the Bayesian approach in general: that our previous knowledge of certain properties of the parameters can be taken into account to improve the model.

[1] S. Wesolowski, N. Klco, R. J. Furnstahl, D. R. Phillips, and A. Thapaliya, Journal of Physics G: Nuclear and Particle Physics **43**, 074001 (2016).
[2] C. Forssén, "Learning from data: Hypothesis testing and bayesian model selection," Chalmers course literature TIF285 (2019).
[3] C. Forssén, "Learning from data: Markov chain monte carlo sampling," Chalmers course literature TIF285 (2019).
[4] I. Svensson, "Tif285 learning from data," Private communication, Chalmers (2020).