Project 1 - Learning from Data

Parameter estimation for a toy model of an effective field theory

Mathias Arvidsson CID: matharv, Carl Strandby CID: carstran

Program: Teknisk Fysik.

Course: Learning from Data, [Tif285], Chalmers, Fall 2022.

Contents

1 Introduction **1**

2 Description of methods used **1**

3 Results **3**

3.1 Basic 3

3.2 Optional 6

4 Discussion **8**

1 Introduction

Having performed an experiment in physics one is often faced with the challenge of finding a function that explains the observed behaviour. This can include both identifying which observed variables are correlated or anti-correlated with each other and determining their parameter values, and identifying which variables can be seen as noise and neglected. One framework for parameter estimation is Bayesian statistics, which not only attempts to infer the values of the parameters, but also attempts to determine the degree of belief (DoB) that those are the correct parameter values.

This project explores a Bayesian approach to estimating parameters by determining coefficients of a Taylor expansion. Synthetic data representing experimentally measured data points was generated from the function

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

and includes independent data, x_j, the dependent function values, $d(x_j)$, and standard deviations, σ_j, used to offset the function values from the true function $g(x)$. The taylor expansion of the function $g(x)$, whose coefficients this project tries to determine are

$$g(x) = 0.25 + 1.57x + 2.47x^2 + 1.29x^3 + ... \tag{2}$$

The specific goals of this project is to adopt the methodology described in, and reproduce results from Sarah Wesolowski, et al [4]. Specifically we have chosen to reproduce figures 1, 3 and 4 with $k_{max} = 3$, which corresponds to estimating four coefficients of the Taylor expansion. These figures describe correlations between posterior distributions of coefficients and their degree of beliefs, resulting from an MCMC-sampling using a Uniform and a Gaussian prior. Figures 3 and 4 also compare the fit produced with the estimated parameters with the measured datapoints. Furthermore this project attempts to reproduce results displayed in Table 3 in Wesolowski et al, and generating extra synthetic data to examine the impact of varying the number of datapoints.

2 Description of methods used

Bayesian statistics attempts to calculate a posterior probability distribution for the parameter values. The "best" parameter estimation is then the parameter values that maximizes this posterior. The posterior itself is calculated with Bayes rule and is proportional to a likelihood function times a prior. The likelihood function describes the probability of measuring the given data using the parameter values being tested. The function used here is a cost function defined by Wesolowski as

$$\mathrm{pr}(D \mid \mathbf{a}, I) = \prod_{j=1}^{N_d} \left(\frac{1}{\sqrt{2\pi}\sigma_{j,\mathrm{exp}}} \right) e^{-\chi^2/2}, \tag{3}$$

where the χ^2-function is defined to be

$$\chi^2 = \sum_{i=1}^{N_d} \left(\frac{d_i - g_{\mathrm{th}}\left(x_i\right)}{\sigma_i} \right)^2. \tag{4}$$

The prior encodes previous knowledge about the problem and choosing a prior that is as uninformed as possible remains a challenge in Bayesian statistics. Here two priors are compared, one uniform and one Gaussian, whose differences are discussed more in section 4. The uniform prior is simply one for parameter values in the specified range and zero for values outside. The Gaussian prior used here is defined by Wesolowski as

$$pr(\boldsymbol{a}|\bar{a}, I) = (\frac{1}{\sqrt{2\pi}\bar{a}})^{k+1}\exp(-\frac{\boldsymbol{a}^2}{2\bar{a}^2}). \tag{5}$$

The space of possible combinations of parameter values grows quickly as the number of parameters increases. This has two notable consequences. The first is that the likelihood and posterior values can be very small for a given combination of parameters, which is why logarithmic versions of the prior, likelihood and posterior are used. The second is that it becomes computationally intense to calculate the posterior distribution for every combination of parameter values. One way of circumventing this problem is to use an MCMC-sampling-algorithm to identify and evaluate the posterior distribution in interesting regions. Meaning regions where the posterior evaluation is more likely to impact the overall posterior distribution. A successful implementation of an MCMC-sampling algorithm would thus yield an approximate posterior distribution. In the implementation of an MCMC-sampling-algorithm this project makes use of the EnsambleSampler function in the python package emcee [1]. To define and run the sampler requires four inputs; the number of dimensions of the parameter space, the function to be explored, the number of walkers and finally an initial position. Below follows a short description of the four inputs.

The number of dimensions for the parameter space is simply the number of terms in the Taylor-expansion the sampler is instructed to look for. Note that this project adopts the convention of Wesoloski and describes figures and results in terms of truncation order, k, meaning that $k = k$ describes a $k + 1$-dimensional parameter space [4]. The function to explore is the logarithmic posterior, that is the log likelihood plus the log prior described above. The evidence here is taken to be a normalization constant but later calculated and compared for values found using the gaussian prior. The evidence is calculated with Laplace's method

$$Z_P \approx P^*(\theta_0)\sqrt{\frac{(2\pi)^K}{\det \Sigma^{-1}}}, \tag{6}$$

where Z_P is the evidence, θ_0 is the vector of optimal parameter values, K is the number of dimensions ($K = k + 1$), and Σ^{-1} is the Hessian [3].

A walker specifies a sequence of proposed positions which are accepted according to rules specified by a policy used by the sampling-algorithm. The rough idea behind these policies is to guide the walkers to more often evaluate the posterior distribution for parameter values that are estimated to be more probable and to sometimes evaluate the less likely parameter value in order to explore the parameter space. How this is done varies. Commonly, a walker will not choose actions based on it's own history of evaluated positions, but on the evaluations of other walkers in the same timestep. If a specific combination of parameter values evaluated by a walker is deemed to have a high probability, other walkers may gravitate towards trying similar parameter values. The specific policy used in this project was emcee's standard policy, "stretch move". The number of walkers was chosen to be 32. The initial position for the walkers was calculated by estimating the maximum of

the posterior distribution using the function ”optimize.minimize” in the python package scipy. This estimation was then offset for each walker randomly according to a normal distribution around the estimated maximum. This follows the procedure used in the documentation for emcee [2].

3 Results

This section presents figures and tables representing the parameter estimations for both the basic tasks and the extra tasks. The figures are interpreted and explained briefly. For further discussion see section 4.

3.1 Basic

Figures 1 and 2 show corner-plots of the posterior probability distributions for $k_{max} = 3$, that is four taylor coefficients using a uniform and gaussian prior respectively. Each pane shows a marginalized two-dimensional projected posterior for two of the parameters listed on the axes. The bottom left pane, for example, shows the relationship between the parameters a_3 on the y-axis, and a_0 on the x-axis. The true values of the given taylor expansion are indicated in blue. The uppermost pane of each column show the marginalized one-dimensional probability distribution for said parameter. The three dashed lines indicates the mean $\pm\sigma$ corresponding to a 68 degree of belief interval and the numerical values indicates the mean and the two sigmas respectively. The best fit parameters for the uniform prior was determined to be $a_0 = 0.27 \pm 0.04, a_1 = 0.96 \pm 1.1, a_2 = 8.04 \pm 8.05,$ and $a_3 = -9.93 \pm 16.5$ while the corresponding values for the gaussian prior were $a_0 = 0.25 \pm 0.02, a_1 = 1.65 \pm 0.45, a_2 = 2.96 \pm 2.35,$ and $a_3 = 0.30 \pm 4.3$.

How two parameters correlate with each other is also implicated by the contour plots. In general a circular contour plot indicates a negligible correlation, while elliptical contours do indicate a correlation. The thinner the ellipse, the more strongly the two parameters are correlated. The direction of the ellipse implies whether the parameters are correlated, indicated by a ”positive slope”, or anti-correlated, indicated by a ”negative slope”. The anti-correlation indicating that if one parameter is increased it is likely that the other is decreased. Note that the correlations are not all scaled one to one as the axes are scaled differently.

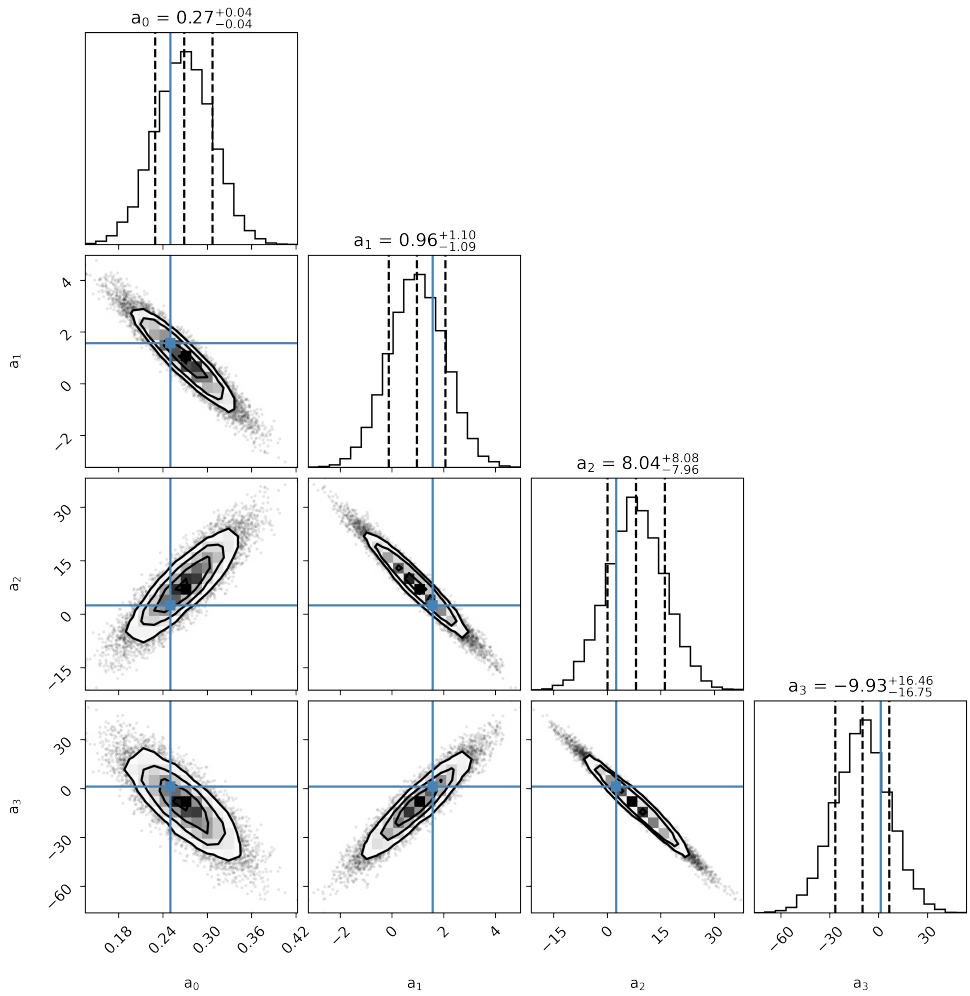


Figure 1: Projected posterior distributions for the different parameters $[a_0, a_1, a_2, a_3]$. Calculated using a Uniform prior. Presented in a corner plot where each pane shows a projected distribution of either one or two parameters. The blue lines indicate the true parameter values, while the dashed lines show the mean and 68% degree of belief intervals for the estimations.

It is worthwhile noting that both posteriors in figure 1 and 2 imply that coefficients for even powers of x are anti-correlated with the coefficients for odd powers of x. It is also notable that ellipses of the "neighbouring" coefficients, a_n and a_{n+1}, are thinner which indicates a stronger correlation while coefficients further away from each other are more spread out.

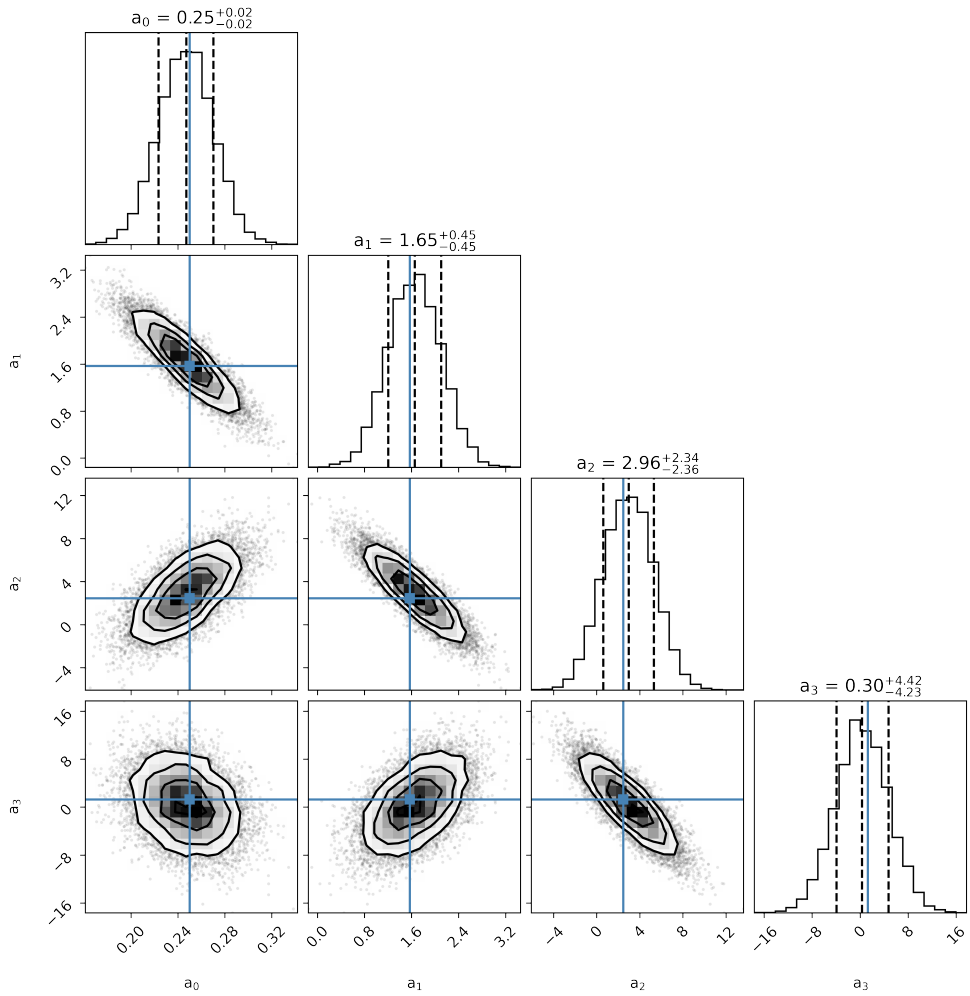
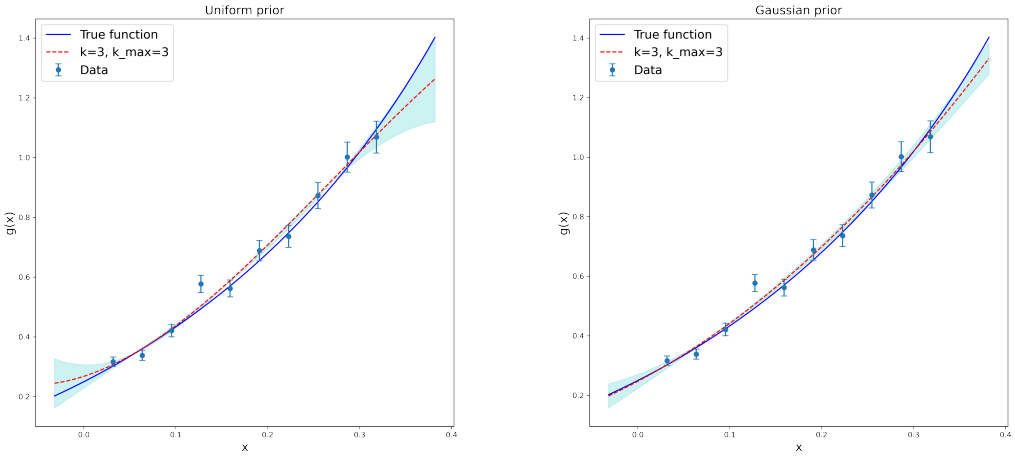


Figure 2: Projected posterior distributions for the different parameters $[a_0, a_1, a_2, a_3]$. Calculated using a Gaussian prior. Presented in a corner plot where each pane shows a projected distribution of either one or two parameters. The blue lines indicate the true parameter values, while the dashed lines show the means and 68% degree of belief intervals for the estimations.

Figures 3a and 3b show the true function, equation 1 and the synthetic datapoints generated from the true function in comparison to the approximated Taylor function with coefficients a_0 to a_3 described above. In both figures one can see that the approximated function is relatively close to the true function within the range of x-values covered by the synthetic datapoints. Outside of this range the approximated function diverges from the true function. The cyan background enfolding the approximated function describes the 68% degree of belief intervals. For the uniform prior this degree of belief interval quickly diverges and spreads out, while the corresponding interval for the gaussian prior can be seen to be slightly more stable around the mean.



(a) Model from uniform prior, k_{max}=3. (b) Model from gaussian prior, k_{max}=3.

Figure 3: Comparison plots between the true function and the estimated model for two different priors. The dots show the data points, while the light blue transparent area corresponds to the 68 % DoB margins for the model.

3.2 Optional

In order to reproduce table 3 from Wesolowski et al [4], the truncation order k_{max} was set to range from zero to six and the resulting values can be seen in table 1. Comparing these two tables one can see that the resulting error measurement of the uniform prior, χ^2 over degree of freedom $(k_{max}+1)$, is slightly higher in table 1 for most values of k_{max} but about ten times higher for $k_{max}=0$. It is also notable that while the values for a_2 remains relatively stable as k_{max} increases in table 1, they quickly degrade in Wesolowski's findings. This observation of degrading coefficients is however compatible with the divergence outside of the range of x-values seen in figures 3a and 3b. Thus the difference could stem from the error measurement being made on different ranges of x-values. The resulting parameter values and evidence measurements calculated with equation 6 for the gaussian prior is seen to be more similar, however it is worth noting a difference in coefficient values compared to values described in section 3.1. A slight difference might stem from a small modification in the code however, since in order to accommodate for the longer autocorrelation time of MCMC-sampling for high k_{max} a greater number of steps were discarded.

Table 1: Parameter estimations of the first three coefficients in the taylor expansion from equation (2) for different orders of k_{max}, and different priors. The final 7 columns uses the gaussian prior, while the columns 2-8 use the uniform prior. The evidence is calclulated using equation (6), and the χ^2/dof is calculated using equation 4 and dividing with $k_{max} + 1$.

k_{max}	χ^2/dof	a_0	$\pm\sigma$	a_1	$\pm\sigma$	a_2	$\pm\sigma$	Evidence	a_0	$\pm\sigma$	a_1	$\pm\sigma$	a_2	$\pm\sigma$
0	599.0	0.48	0.01						0.48	0.01				
1	8.9	0.20	0.01	2.6	0.1			599	0.20	0.01	2.5	0.1		
2	3.8	0.25	0.02	1.6	0.4	3.3	1.3	3282	0.25	0.02	1.6	0.4	3.2	1.3
3	2.8	0.27	0.04	1.0	1.1	8.1	8.1	2856	0.25	0.02	1.6	0.4	3.0	2.3
4	2.2	0.27	0.04	0.8	1.2	10.5	11.3	2852	0.25	0.02	1.6	0.5	3.0	2.4
5	1.8	0.27	0.04	0.8	1.3	11.1	11.6	2782	0.25	0.02	1.6	0.5	3.0	2.4
6	1.6	0.27	0.04	0.8	1.3	10.6	11.8	2741	0.25	0.02	1.6	0.5	3.0	2.4

Continuing the optional tasks, new datapoints were generated. The procedure of the Basic task was repeated to compare how well the parameter estimation would work when the number of datapoints were changed from ten, to five and fifteen respectively. The three sets were generated using equation 1 and errors taken from a standard distribution that scaled with the value of x. Figure 4 presents the comparison between the true function and estimated parameters for 5 generated data points, while figure 5 presents the results for 15 data points.

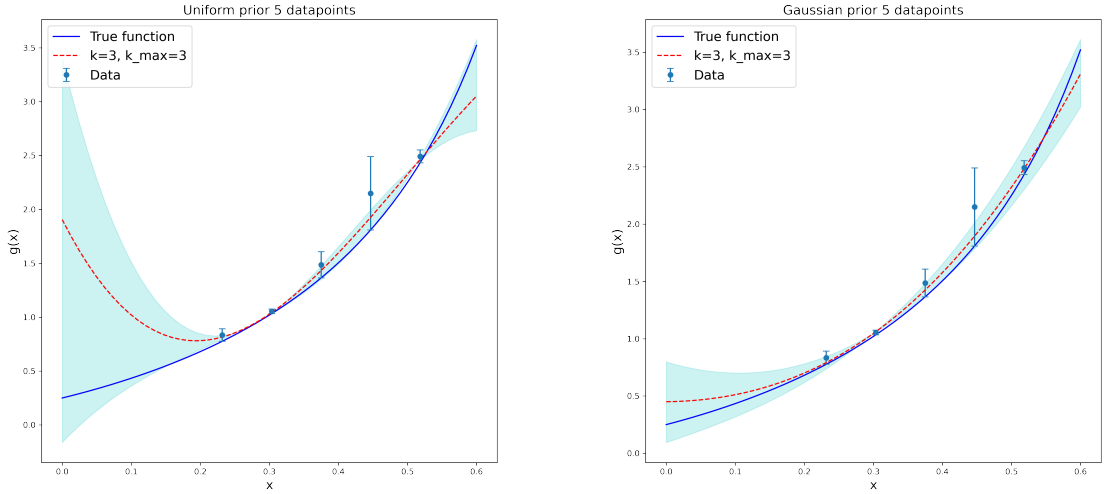


Figure 4: Comparison plots between the true function and the estimated model using 5 data points. The figure on the left was generated using a uniform prior, and the figure on the right using a gaussian prior. In both cases $k_{max} = 3$. The dots show the data points, while the light blue transparent area corresponds to the 68 % DoB margins of the model.

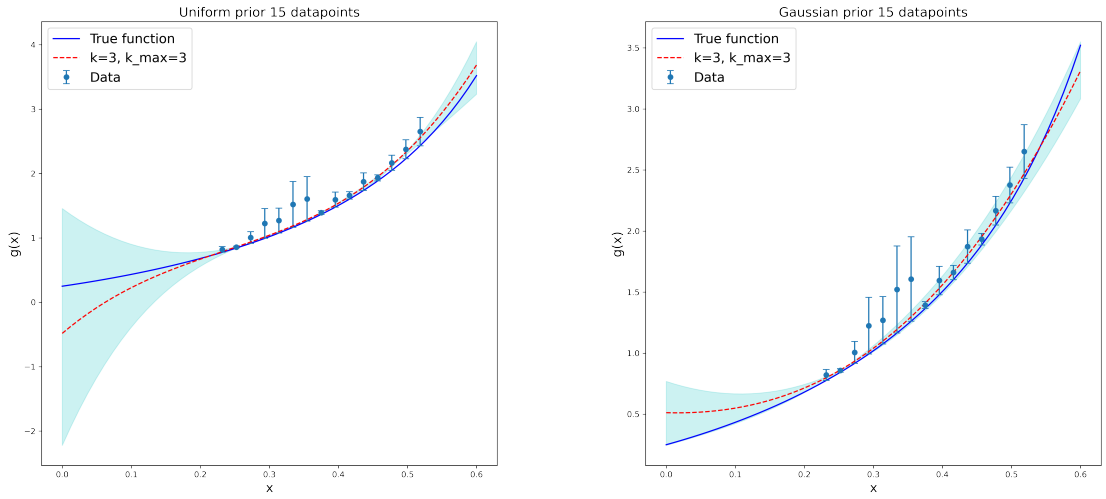


Figure 5: Comparison plots between the true function and the estimated model using 15 data points. The figure on the left was generated using a uniform prior, and the figure on the right using a gaussian prior. In both cases $k_{max} = 3$. The dots show the data points, while the light blue transparent area corresponds to the 68 % DoB margins of the model.

For only 5 datapoints the estimated functions quickly diverges outside of the data-range, especially for the uniform prior. The gaussian prior on the other hand seems to perform better outside of the measured area. For 15 data points similar patterns are visible. A small decrease in error margins might be seen when comparing the degree of belief intervals for the gaussian prior smaller and larger set of datapoints.

4 Discussion

It can be noted both in the corner-plots, figures 1 and 2, that the degree of belief intervals for $\pm\sigma$ grows with the order of the parameters. This could be at least partly explained by the nature of taylor expansions. Since the expansion is the most precise around the expansion point, in this case $x = 0$, the further one strays from the expansion point the more terms are needed to be added to keep the expansion relevant. Around the expansion point the higher order terms barely contribute meaning that they remain more free to vary and are thus less strictly determined, and consequently have larger error margins.

Having performed the emcee-samplings of the posteriors with two different priors, one uniform and one gaussian, a natural question to ask is which was better and why? It's important to note that this depends heavily on what problem one is trying to solve, but given the resulting posteriors in this experiment, one is led to the conclusion that the gaussian prior was the better. One argument for this is that the errors, the coefficient degree of belief intervals, scaled better with higher order terms, see table 1. This indicates that the gaussian prior is more stable when dealing with higher-dimensional posterior distributions. Another argument comes from comparing figures 3a and 3b, where it can be noted that the degree of belief intervals remain more stable outside of the interval of data points used to estimate the parameters. If one would validate the parameter estimations

on x-values outside of the forementioned range, it is likely that the gaussian prior would perform better. Thirdly it is interesting to note that when varying the number of datapoints in figures 4 and 5, the posterior for the gaussian priors only changes marginally, indicating that it is a good choice of prior even for small datasets. A small note is that determining the most appropriate prior can be theoretically boiled down to choosing the prior which maximizes entropy [3]. It can be shown that if the mean and variance is known, such as in this problem, the assignment of a normal distributed prior is the most appropriate.

Another interesting question comes from comparing the calculated evidences of different truncation orders k_{max} in table 1 and from this trying to infer which of the seven models were best. While the value of the evidence itself can't be used to compare two different models, the ratios between them can. Comparing the values it is noted that $k_{max} = 2$, three coefficients, has a slightly larger evidence that the following truncation orders. The bayesian framework thus implies that if we need to choose one of the seven models, we should prefer the model with three terms in the taylor expansion [3].

References

[1] Daniel Foreman-Mackey et al. "emcee: The MCMC Hammer". In: 125.925 (Mar. 2013), p. 306. DOI: 10.1086/670067. arXiv: 1202.3665 [astro-ph.IM].

[2] Foreman-Mackey et al. *Fitting a model to data*. URL: https://emcee.readthedocs.io/en/stable/tutorials/line/.

[3] Christian Forssén. *Learning from data*. 2022. URL: https://cforssen.gitlab.io/tif285-book/content/Intro/welcome.html.

[4] S Wesolowski et al. "Bayesian parameter estimation for effective field theories". In: *Journal of Physics G: Nuclear and Particle Physics* 43.7 (May 2016), p. 074001. DOI: 10.1088/0954-3899/43/7/074001. URL: https://doi.org/10.1088%2F0954-3899%2F43%2F7%2F074001.