Master's Thesis

Sparse Identification of Coupled Stochastic Differential Equations from Data

Tobias Wand

Advisor and Examiner:
Dr. Oliver Kamps

Main Examiner:
PD Dr. Svetlana Gurevich

July 2020

Contents

"Data alone has no value — it's just masses of numbers or words."

— Steven J. Bowen

1 Introduction and Motivation

Complex systems in physics, chemistry, biology, economics and other quantitative sciences can be subjected to similar methods of analysis [27]. Especially since the advent of big data technologies, time series analysis became an important part of the toolbox of nonlinear sciences [40, 41, 42]. Estimating the underlying differential equations from observed data can give important insight into the system dynamics. Great advances have already been made with regards to deterministic differential equations [26].

However, many dynamical systems show a separation of time scales [27] and "fast dynamics can often be treated as fluctuations" [2]. Hence, being able to deal with such fluctuations via stochastic differential equations is a promising strategy to analyze dynamical system with separate time scales. Similarly, the limited accuracy of data measurement also requires a robust interpretation algorithm which does no fail in the face of noisy data. Because stochastic differential equations occur in various fields of science, their analysis can be very beneficial for a variety of disciplines [2, 3, 5]. Yet, they include additional challenges for data analysis and require a more refined analysis than the SINDy method presented in [26]. Often, analyzing stochastic ODEs is done with a known shape of equations in mind [5], but this limits the approach to well-understood systems. An ideal tool for the analysis of data would be able to construct a stochastic ODE only based on the data and without any further knowledge about the system dynamics.

Moreover, it is usually desired to receive a sparse ODE without any superfluous terms because of several reasons: First, a sparse ODE can avoid the problem of an overfitted solution that is unable to be used for predicting future data. Second, if the ODE only includes few terms, then it might be possible to interpret every term in order to find out its real world counterpart in the dynamical system's physics. And finally, "often the behavior of

complex systems that are far from equilibrium can be traced back to rather simple laws" ([2]): complex dynamics being caused by simple and sparse ODEs is part of the beauty and raison d'être of nonlinear sciences. Methods such as the Lasso regularization can be used to promote sparsity [22], but the analysis of deterministic ODEs in [26] suggests a hard threshold as the best practise for the analysis of ODEs.

This thesis describes how to estimate stochastic ODEs with dynamical noise from data without any prior knowledge about the shape of the ODEs. Hence, a very general approach to data analysis with a wide range of applicability is outlined in this thesis. Recently, Bayesian methods have been used to quantify the uncertainty of noisy data and to subsequently use hyper-parameters to estimate the underlying differential equations with a similarly general framework for the equations [4]. Instead, I apply the Bayesian methods to the optimization of the hyper-parameter's impact on the reconstruction of the ODEs. Furthermore, I use Markov Chain Monte Carlo samples to gain insight into the posterior distribution of the unknown model parameters and therefore regard a fully Bayesian interpretation of the data. These methods prove to be crucial tools in successfully estimating ODEs from noisy data.

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

— Sir Arthur Conan Doyle

2 Theoretical Foundations

Probability theory uses mathematics, the language of logical truth, to describe what cannot be known for certain. This section is an introduction to the mathematical theory of probability with a focus on Bayesian methods. First, this section gives a brief overview over the most important aspects of probability theory and its applications in statistics in section 2.1. Then, the Bayesian approach to statistics is discussed in section 2.2 including the two crucial aspects of constructing a posterior density based on prior knowledge and assessing models via the Bayesian information criterion. Finally, section 2.3 discusses several instructive examples to highlight the importance of Bayesian methods. More detailed information and proofs can be found in e.g. [38], [8] and [28] or in the more unorthodox approach to probability theory in [37].

2.1 Introduction to Probability Theory and Statistics

Here, a short introduction to the basics of probability theory is outlined. The most important definitions and limit theorems are briefly explained and the Parzen estimator is discussed as a simple, yet effective method to interpret data with statistical methods. Finally, this subsection introduces the concept of sampling algorithms.

2.1.1 Distributions and Their Characteristic Values

The distribution of random variables X representing experiments or events with an unpredictable outcome is a central aspect of probability theory. The distribution is given by a density function $f(x)$ that describes the probability $\mathbb{P}([a,b])$ of the outcome of X being in an interval $[a,b]$ and is

defined as

$$\mathbb{P}\left([a,b]\right) = \int_a^b f(x)\,\mathrm{d}x \qquad \text{with} \qquad \int f(x)\,\mathrm{d}x \stackrel{!}{=} 1 \tag{1}$$

as the normalization. For discrete random variables X that only show a countable number of outcomes $x_{1,2,\ldots}$, the normalization simplifies to $\sum_i \mathbb{P}(X = x_i) = 1$ with the identity $f(x_i) = \mathbb{P}(X = x_i)$.

Instead of always dealing with the full density function, it is often useful to summarize the information of the density function in two values, the expectation value $\mathbb{E}[X]$ and the variance $\mathbb{V}(X)$. This is especially advisable in the case of unimodal distribution functions, where most values can be expected to lie in the range of $\pm\sqrt{\mathbb{V}(X)}$ around $\mathbb{E}[X]$. These values are defined as

$$\mathbb{E}[X] = \int x f(x)\,\mathrm{d}x \quad \text{and} \quad \mathbb{V}(X) = \int (x - \mathbb{E}[X])^2 f(x)\,\mathrm{d}x \tag{2}$$

or in the case of discrete X

$$\mathbb{E}[X] = \sum_i x_i \mathbb{P}(X = x_i) \quad \text{and} \quad \mathbb{V}(X) = \sum_i (x_i - \mathbb{E}[X])^2 \mathbb{P}(X = x_i). \tag{3}$$

Since the probability measure $\mathbb{P}$ is only defined for sets, the notation $\mathbb{P}(X = x_i)$ is an abbreviation of $\mathbb{P}(\{X = x_i\})$.

2.1.2 Limit Theorems

Two limit theorems, the central limit theorem and the law of large numbers, are so important for probability theory that they are sometimes taken as axiomatic properties of mathematical statistics. Yet, they can be derived from even more fundamental ideas and theorems.

Law of Large Numbers

The law of large numbers, in its most basic form, is a corollary to the Chebychev inequality, which itself is a special case of the Markov inequality. For a

monotonously increasing function $h(x) \geq 0$, it holds that

$$\begin{aligned} h(a)\mathbb{P}(X \geq a) &= \int_a^\infty h(a) f(x)\,\mathrm{d}x \leq \int_a^\infty h(x) f(x)\,\mathrm{d}x \\ &\leq \int_0^\infty h(x) f(x)\,\mathrm{d}x = \mathbb{E}[h(X)]. \end{aligned} \tag{4}$$

Thus, the Markov inequality $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[h(X)]}{h(a)}$ holds. Specifying $h(x) = x^2$ and replacing X with the centralized random variable $Y = |X - \mathbb{E}[X]|$ directly yields the Chebychev inequality

$$\begin{aligned} \mathbb{P}(Y \geq a) &\leq \frac{\mathbb{E}[Y^2]}{a^2} \text{ and hence} \\ \mathbb{P}(|X - \mathbb{E}[X]| \geq a) &\leq \frac{\mathbb{V}(X)}{a^2}. \end{aligned} \tag{5}$$

The Chebychev inequality shows that there is an upper boundary to the probability that X deviates from its expectation value. This is especially useful for the mean value of a sample of independent and identically distributed random variables Z_i with the sample mean $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$. If the variance of each Z_i is finite $\mathbb{V}(Z_i) < \infty$, then simple algebra shows that $\mathbb{V}(\overline{Z}_n) = \frac{1}{n^2}\mathbb{V}(Z_1)$ and $\mathbb{E}[\overline{Z}_n] = \mathbb{E}[Z_1]$. Hence, the Chebychev inequality becomes

$$\mathbb{P}(|\overline{Z}_n - \mathbb{E}[Z_1]| \geq a) \leq \frac{\mathbb{V}(Z_1)}{a^2 n^2} \xrightarrow{n\to\infty} 0. \tag{6}$$

This means, that any deviation a between sample mean $\overline{Z}_n$ and the expectation value $\mathbb{E}[Z_1]$ occurs with negligible probability, if n is large enough: the sample mean converges to the expectation value. This property is so fundamental to "everyday statistics" that mean and expectation value are frequently used synonymously, but in fact, the **weak law of large numbers (WLLN)** in equation (6) describes the true connection between those two entities. The law of large numbers can be refined to show an even stronger convergence level as can be seen in [8].

Central Limit Theorem

The central limit theorem is another property that is frequently taken for granted when dealing with mathematical statistics. It states that the sum of independent and identically distributed random variables X_i is a Gaussian random variable.

Consider such random variables $X_{1,...,n}$ with $\mathbb{E}[X_j] = \mu$ and $\mathbb{V}(X_j) = \sigma^2$. The transformation $Y_j = \frac{X_j - \mu}{\sigma}$ will yield random variables with a zero expectation value and unit variance. The transformed sum S_n of the X_j is given as

$$S_n = \frac{\sum_{j=1}^{n}(X_j - \mu)}{\sqrt{n\sigma^2}} = \sum_{j=1}^{n} \frac{Y_j}{\sqrt{n}}. \tag{7}$$

The central limit theorem can be easily derived via the use of Fourier transforms, which in probability theory are usually referred to as the characteristic function $\phi_Z(t) = \mathbb{E}[\exp(itZ)]$ of a random variable Z. Because of the convolution theorem and the identical distribution of the Y_j, it follows

$$\phi_{S_n}(t) = \phi_{\sum_j (Y_j/\sqrt{n})}(t) = \left[\phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n. \tag{8}$$

With a Taylor series expansion for large n and therefore small $\frac{t}{\sqrt{n}}$

$$\phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + \mathcal{O}\left(\frac{t^2}{n}\right) \tag{9}$$

and hence

$$\phi_{S_n}(t) \approx \left[1 - \frac{t^2}{2n}\right]^n \xrightarrow{n\to\infty} e^{-\frac{1}{2}t^2}. \tag{10}$$

This proves the convergence of S_n to a Gaussian distribution because of the invariance of Gaussian functions under Fourier transforms. Variable transformation hence yields that $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to a Gaussian $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ distribution. A more formal proof of the **central limit theorem (CLT)** can be found in [8].

2.1.3 Parzen Estimator

The Parzen estimator or kernel density estimation (KDE) is a method to reconstruct the underlying density $f(x)$ for a given sample $x_1, ..., x_n$ with an estimator $f_n(x)$. It uses a superposition of kernel functions $K(x)$ which are probability densities without the normalization constant. $f_n(x) = (1/nh)\sum_i K((x - x_i)/h)$ defines a smoothened version of the histogram of the sample, where the exact choice of $K(x)$ (Gaussian, Cauchy etc.) may not even be relevant. The smoothening parameter or bandwidth $h > 0$ should be low enough to still acknowledge much of the underlying distribution, but

high enough to prevent the formation of a highly multimodal, oversmoothed $f_n(x)$. For further discussion about an optimal h and the asymptotic properties of $f_n(x)$, cf. e.g. [20].

2.1.4 Sampling Methods

Given a probability density $f(x)$, it is often useful to draw a representative sample from this distribution. Given a representative sample, most properties of the distribution (expectation value, variance, correlations etc.) can be inferred and integrals $\int \gamma(x) f(x)\,\mathrm{d}x$ (such as the advanced model comparison in (55)) can easily be calculated. However, getting a representative sample is not as easy as it may seem and it is especially difficult, if $f(x)$ does not have an closed algebraic shape. One has to identify the areas with high probability in order to get a representative sample, which itself can be challenging for a multimodal, multidimensional distribution $f(x)$, since the 'curse of dimensionality' also applies to the simple sampling methods. Here, the basics of sampling methods will be outlined with an emphasis on the application of such methods. A more detailed discussion of sampling methods and their theoretical background can be found in chapters 8 and 9 of [28].

Sampling and Integrals

An weighted integral over a function $\gamma(x)$ and the density $f(x)$ can be expressed in terms of a sample of x with length $x_{1,\dots,N}$ as

$$\frac{\int \gamma(x) f(x)\,\mathrm{d}x}{\int f(x)\,\mathrm{d}x} \;\widehat{=}\; \mathbb{E}[\gamma(x)] = \frac{1}{N}\sum_{i=1}^{N} \gamma(x_i) \tag{11}$$

as long as the law of large numbers holds. Then, the uncertainty of $\mathbb{E}[\gamma(x)]$ decreases as $N^{-1/2}$ independently of the dimension of x. Hence, the 'curse of dimensionality' is lifted by such a sampling method.

Rejection Sampling

Rejection sampling is one of the most basic methods to draw a sample from a known probability density $f(x)$ for $x \in X$. Note that f can be a discrete distribution. One needs another probability density $g(x)$ with $f(x) > 0 \implies g(x) > 0$, a scalar M given by $\forall x \in X : f(x) \leq M g(x)$ and uniform random variables U_i on $[0, 1]$. To acquire a sample of f:

1. draw x_i according to $g(x)$ and u_i from U_i
2. if $u_i < f(x_i)/(Mg(x_i))$, add x_i to the sample; otherwise reject it
3. repeat until the sample has a sufficient size

An average of M iterations is needed for one sample element. This method is used, if the distribution f is neither analytically known nor given by a programming package. For further details, cf. [21]. If the scalar M is chosen correctly, $f(x)$ does not even have to be a normalized density.

2.2 Bayesian Probability Theory

Several aspects of probability theory and statistics fall under the brand of "Bayesian methods". An exhaustive discussion of these aspects and the underlying theory can be found in [6] and in a more general context in [8]. Their common aspect is the use of conditional probabilities, whose mathematical formulation and analysis is usually attributed to reverend Thomas Bayes. The core idea of Bayesian methods is to regard the probability of the occurrence of A given the occurrence of B. The significance of this idea is easily understood, if one regards A as "a given theory is correct" and B as the data resulting from experiments related to this theory. Depending on the known data, what can be inferred about the validity of the theory?

This section deals with the theoretical framework that is necessary to express this line of thought in the language of mathematics. Constructing a posterior distribution based on prior knowledge is one of the major aspects of this section and the Bayesian model selection criterion in 2.2.4 is a powerful criterion to evaluate data in section 4.

2.2.1 Conditional Probabilities

The fundamental expression of Bayesian theory is the conditional probability $\mathbb{P}(A|B)$ expressing the probability of an event A given that B is already known. For $\mathbb{P}(B) > 0$, this property is defined as the relative share of A and B occurring measured against the occurrence of B:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{12}$$

Note that this definition is a probability measure for all sets $A_B \subseteq B$ with normalization $\mathbb{P}(B|B) = 1$. Every set A can be cut down to an $A_B \subseteq B$ as $A_B = A \cap B$, since this does not affect $\mathbb{P}(A|B) = \mathbb{P}(A_B|B)$. Subadditivity and additivity of disjoint sets are still valid for this measure.
If there are mutually exclusive $B_{1,\dots,n}$ forming a partition of the full sample space $\Omega = \bigcup_{i=1}^{n} B_i$ with $\mathbb{P}(B_i) > 0$, then the conditional probability leads to the law of total probability: Just like the sample space Ω can be partitioned by the B_i, every $A \subseteq \Omega$ can also be partitioned. The total probability is then given as

$$\mathbb{P}(A) = \mathbb{P}(A|\Omega) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i) \tag{13}$$

which follows directly from $\mathbb{P}(A|B_i)\mathbb{P}(B_i) = \mathbb{P}(A \cap B_i)$ and the disjoint partition B_i of Ω.

Bayes's Theorem

In the context of evaluating the validity of a model M with respect to known data D, there is still a problem: Usually, only the distribution of the data based on a specific model $\mathbb{P}(D|M)$ is known analytically, but the desired quantity is $\mathbb{P}(M|D)$. Bayes's theorem links these two probabilities via

$$\mathbb{P}(M|D) = \frac{\mathbb{P}(D|M)\mathbb{P}(M)}{\mathbb{P}(D)} \tag{14}$$

and can be easily derived via $\mathbb{P}(D|M)\mathbb{P}(M) = \mathbb{P}(D \cap M) = \mathbb{P}(M|D)\mathbb{P}(D)$ from equation (12). However, this does not yet solve the problems of evaluating the model M as the two probabilities $\mathbb{P}(D)$ and $\mathbb{P}(M)$ still need to be computed. These terms are most easily understood by expanding the approach to the experiment: Instead of only regarding one model M, other models M^C should be included as alternative propositions. This makes sense, because if there only was reason to believe that one model M could have caused the data, this model would have already implicitly been regarded as true. With mutually exclusive models M^C, the probability $\mathbb{P}(D)$ is given by the law of total probability as

$$\mathbb{P}(D) = \mathbb{P}(D|M)\mathbb{P}(M) + \mathbb{P}(D|M^C)\mathbb{P}(M^C)\ . \tag{15}$$

Of course, M^C could not include one, but many alternative models, meaning

that for a countable set of models, $\mathbb{P}(D|M^C)\mathbb{P}(M^C)$ would again correspond to another law of total probability. The most subtle aspect of equation (14), however, is $\mathbb{P}(M)$. This corresponds to the a priori belief that M is the correct model. Here, "a priori" is regarded as "before conducting the experiment". If older experiments or other theoretical derivations give a strong hint that M might or might not be correct, this can and should be included in $\mathbb{P}(M)$. If there is no prior knowledge given, one might asses all possible models $M_{1,\dots,m}$ with the discrete uniform probability $\frac{1}{m}$. Thus, one can infer the probability of a model M given observations D.

This is the basic framework of Bayesian theory. A nomenclature had been established in the field of Bayesian statistics which will also be used in this thesis. In a general version of equation (14) with events A and B and $\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$, the stochastic dependence of A on B is supposed to be known: $\mathbb{P}(A|B)$ is called the *likelihood* of A under B. The probability $\mathbb{P}(B)$ is known as the *a priori distribution* or simply the *prior* and consequently, the updated probability and usually desired quantity $\mathbb{P}(B|A)$ called the *a posteriori distribution* or the *posterior* of B given A. The total probability of A given as $\mathbb{P}(A)$ is called the *evidence* and often treated with little care since it does not explicitly depend on B. Hence, a simplified Bayes's theorem is sometimes given as

$$\mathbb{P}(B|A) \propto \mathbb{P}(A|B)\mathbb{P}(B) \tag{16}$$

if the normalization factor of the evidence can be neglected.

Sometimes, the law of total probability is referred to as "marginalization rule". It may be useful to mention an alternative, more general formulation of equation (13): Consider sets A_i forming a disjoint partition of Ω with $\mathbb{P}(A_i) > 0$ and sets B and C with $\mathbb{P}(C) > 0$. Then, equation (13) can be reformulated as

$$\mathbb{P}(B|C) = \sum_{i=1}^{\infty} \mathbb{P}(B|A_i, C)\mathbb{P}(A_i|C) \ . \tag{17}$$

The inclusion of the A_i is reminiscent of including a unity matrix to a multiplication. The condition C serves as a background condition for all of the probabilities in equation (17).

Example: Monty Hall Problem

One of the most famous examples of the basic Bayesian approach leading to an counterintuitive conclusion is the Monty Hall problem. Monty Hall's game show included a game, in which the participant chooses one of three closed doors: Behind one of the doors is a prize P, behind the other two doors is a goat P^C. The participant fist chooses one of the three closed doors and Monty Hall then opens one of the two other doors to reveal one of the two goats. Then, the participant is given the chance to switch his choice to the other closed door or to remain at the initially selected door. The question of course is, whether switching to the other door increases the probability of success.

First, one has to model the problem. P_i represents the prize being behind the door i, X_i the player initially choosing door i and H_i Monty Hall opening door i. A priori, $\mathbb{P}(P_i) = \frac{1}{3}$ for all i. Consider that the player chooses door 1 (X_1) and Monty Hall opens door 3 (H_3). Because of the symmetry of the problem with respect to permutations, this case is representative of all possible outcomes. After the player chooses X_1, only H_2 and H_3 can happen. If P_2 is true, then H_2 cannot happen and H_3 must happen and vice versa for P_3. If P_1 is true, then H_2 and H_3 happen with equal probability. Hence, one concludes the probabilities

$$\mathbb{P}(H_3|X_1 \cap P_1) = \frac{1}{2}, \quad \mathbb{P}(H_3|X_1 \cap P_2) = 1, \quad \mathbb{P}(H_3|X_1 \cap P_3) = 0\ . \quad (18)$$

The desired quantity is $\mathbb{P}(P_2|X_1 \cap H_3)$ and via the definition of conditional probabilities and Bayes's theorem, this can be expressed as

$$\mathbb{P}(P_2|X_1 \cap H_3) = \frac{\mathbb{P}(P_2 \cap X_1 \cap H_3)}{\mathbb{P}(X_1 \cap H_3)} \overset{Bayes}{=} \frac{\mathbb{P}(H_3|X_1 \cap P_2)\mathbb{P}(X_1 \cap P_2)}{\mathbb{P}(X_1 \cap H_3)}\ . \quad (19)$$

Since X_i and P_i are independent, $\mathbb{P}(X_1 \cap P_2) = \frac{1}{3^2}$. Equation (18) gives $\mathbb{P}(H_3|X_1 \cap P_2) = 1$ and combined with equation (17), it yields with $\mathbb{P}(P_i) = \frac{1}{3}$ that $\mathbb{P}(H_3|X_1) = \frac{1}{2}$. Hence:

$$\mathbb{P}(P_2|X_1 \cap H_3) = \frac{1 \cdot \frac{1}{9}}{\mathbb{P}(H_3|X_1)\mathbb{P}(X_1)} = \frac{6}{9} = \frac{2}{3} \quad (20)$$

and therefore, switching to door 2 increases the chance of winning the prize.

This seemingly counterintuitive result makes sense, if it is regarded in the context of information theory: Initially, all three doors have the same chance of success: $\frac{1}{3}$. Door 1 has the chance $\frac{1}{3}$ and the sum of the chances of 2 and 3 is $\frac{2}{3}$. This is still valid after door 3 has been opened, meaning that still, the combination of 2 and 3 has a success chance of $\frac{2}{3}$. But the additional information about door 3 means that within the combination of 2 and 3, the chance is no longer distributed uniformly, but concentrated on door 2: door 2 "inherits" the combined success probability of the combination 2 and 3.

2.2.2 Continuous Probability Densities

The concept of a conditional probability and the resulting Bayesian equations make sense intuitively, but equation (12) has a significant flaw, as it is only valid for $\mathbb{P}(B) > 0$. This does not matter in experiments with discrete distributions, but requires a careful generalization in order to apply the concepts of conditional probabilities to continuous distributions. In mathematical probability theory, this concept is usually motivated in the context of conditional expectations. Hence, the first part of this subsection will briefly introduce conditional expectations, whereas the second part will focus on the conditional probability density. Readers without the desire for mathematical rigour, however, may skip this section.

Conditional Expectations

Consider two random variables X and Y with respect to the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The notation $\mathbb{E}[X|Y]$ represents the conditional expectation of X given Y and it is defined via a minimizing problem: the conditional expectation of X given Y is the best approximation of X in the Hilbert space $\mathfrak{L}^2(\sigma(Y))$. This is the function space

$$\mathfrak{L}^2(\sigma(Y)) = \{Z : Z \text{ is } \sigma(Y)\text{-measurable and } \mathbb{E}[Z^2] < \infty\}$$

and the minimization is conducted with respect to the L^2 norm:

$$\mathbb{E}[X|Y] = \underset{Z \in \mathfrak{L}^2(\sigma(Y))}{\arg\min} ||X - Z||_2 \tag{21}$$

and is $\mathbb{P}$-a.s. ($\mathbb{P}$ almost surely) unique. One can think of the conditional

expectation in a geometrical way as the projection of the random variable X on the Hilbert space $\mathfrak{L}^2(\sigma(Y))$ with the $\mathfrak{L}^2(\sigma(Y))$ being the function space that encodes the knowledge given by Y.

The conditional expectation then fulfills the following two conditions:

1. $\mathbb{E}[X|Y]$ is $\sigma(Y)$-measurable
2. for every $A \in \sigma(Y)$ is $\int_A \mathbb{E}[X|Y]\,\mathrm{d}\mathbb{P} = \int_A X\,\mathrm{d}\mathbb{P}$.

This is a very technical expressions corresponding to the following line of thought: X should be replaced with an easier version of itself that only includes the knowledge given by Y. Hence, the conditional expectation is $\sigma(Y)$-measurable, but no longer necessarily $\mathcal{A}$-measurable. The remaining information in $\mathcal{A} \setminus \sigma(Y)$ is no longer needed. However, for every piece of information $A \in \sigma(Y)$, the conditional expectation $\mathbb{E}[X|Y]$ must accurately represent X. Therefore, the equation $\int_A \mathbb{E}[X|Y]\,\mathrm{d}\mathbb{P} = \int_A X\,\mathrm{d}\mathbb{P}$ has to hold for every $A \in \sigma(Y)$.

The conditional expectation allows a mathematically clear definition of conditional probabilities that holds even for continuous distributions. The probability of an event $B \in \mathcal{A}$ can always be expressed as the expectation value of its indicator function $\mathbb{E}[\mathbb{1}_B] = \mathbb{P}[B]$. This can be adapted to the conditional expectation and defines

$$\mathbb{P}(B|Y) = \mathbb{E}[\mathbb{1}_B|Y] \tag{22}$$

for every set $B \in \mathcal{A}$ and every random variable Y. Since the minimization criterion (21) does not depend on Y being distributed discretely, equation (22) generalizes the naive approach of equation (12).

This short digression about conditional expectations shows the procedure to arrive at a mathematically well-defined conditional probability. The next section will deal with the conditional probability density as a more practical tool to calculate conditional probabilities. Further properties of the conditional expectation can be found in chapter 8 of [8] where they are discussed in great theoretical detail.

Conditional Probability Densities

Probability density functions define a given distribution and every piece of information about a distribution can be derived from its density. Hence, the next step in the analysis of conditional probabilities is to derive a conditional probability density function.

Again, consider two random variables X and Y, but now, the vector (X, Y) is the random variable with respect to the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. If (X, Y) has a joint probability density $f(x, y)$ with respect to $\mathrm{d}\lambda^2$, then one can define

$$f_Y(y) = \int f(x, y)\,\mathrm{d}x \ . \tag{23}$$

This is similar to the Bayesian concept of marginalization. Then, one can prove that

$$\mathbb{E}[X|Y = y] = \frac{1}{f_Y(y)} \int x f(x, y)\,\mathrm{d}x \tag{24}$$

and therefore for general values of Y

$$\mathbb{E}[X|Y] = \frac{1}{f_Y(Y)} \int x f(x, Y)\,\mathrm{d}x \quad \mathbb{P}\text{-a.s.} \tag{25}$$

Hence, $\frac{f(x,Y)}{f_Y(Y)}$ acts as a conditional density for X given Y. Although the concept of a conditional density is frequently treated with little care and it is often derived from pure mathematical intuition, it is still noteworthy to know that there exists a mathematically well-defined theory behind this idea. Moreover, it is important to stress that equation (25) provides a uniqueness condition for the conditional density with $\mathbb{P}$-almost-sureness.

2.2.3 Bayesian Parameter Estimation

Non-Bayesian statistics introduced methods like the maximum likelihood estimation or the method of moments to estimate the value of an unknown parameter θ based on observations x_i of random variables $X_i \sim \mathbb{P}_\theta$ with $X_i \in \chi$ for a known distribution shape $\mathbb{P}_\theta$ with density $f_\theta(x)$. Bayesian methods help to incorporate prior knowledge into this estimation problem via probability densities. This section illustrates several aspects and subtleties of Bayesian methods (e.g. the selection of a suitable prior function) and it might be instructive for readers unfamiliar with Bayesian methods. Other readers might directly skip to section 2.2.4.

The idea behind Bayesian parameter estimation is to regard the true parameter θ as being distributed according to an a priori distribution $\alpha(\theta')$ for all $\theta' \in \Theta$: the fixed θ is replaced by a random variable τ with the a priori density α encoding the prior information. The observations x then depend on a two-stage experiment: x is distributed according to $f_{\tau=\theta}$ and τ according to α. Of course, the distribution α might depend on even more parameters with further a priori distributions, which would lead to a hierarchical Bayesian model, but for now, a two-stage experiment will be considered. This Bayesian estimation will be discussed in this section following the theory presented in [7].

The aim of Bayesian estimations is to derive an estimation $T(x)$ based on the observations x of random variables X for a parameter θ. In the Bayesian context, the error of the estimator is analyzed with respect to a risk function that characterizes the gravity of such an error. A standard approach leading to simple and elegant mathematics is $L(\theta, T(x)) = (\theta - T(x))^2$ as a quadratic loss function. This, however, still includes the unknown parameter θ and therefore, it cannot be computed exactly. Instead, the risk function $\mathcal{R}$ and the Bayesian risk function r are used:

$$\mathcal{R}(\theta, T) = \mathbb{E}_\theta[L(\theta, T(X)] = \int_\chi L(\theta, T(x)) \, \mathrm{d}\mathbb{P}_\theta(x) \tag{26}$$

is the expected loss function for a given θ. The θ is still unknown, but the random fluctuations of the observations x of X are now taken into account. Here, the $\mathbb{E}_\theta$ indicates that the expectation value is computed with respect to the true θ as an underlying parameter.
Since the true θ is still unknown, one can then use the a priori distribution α and calculate the expectation value of $\mathcal{R}(\theta, T)$ with respect to the distribution α of θ. Then,

$$r(\alpha, T) = \int_\Theta \mathcal{R}(\theta, T)\alpha(\mathrm{d}\theta) = \int_\Theta \int_\chi L(\theta, T(x))\mathbb{P}_\theta(\mathrm{d}x)\alpha(\mathrm{d}\theta) \tag{27}$$

defines the **Bayes risk** and only depends on the function of the observation $T(x)$ and on the prior knowledge $\alpha(\theta)$. The optimal estimator $T(x)$ is canonically defined as the T which minimizes the Bayes risk: $r(\alpha, T) \leq r(\alpha, T')$.

The advantage of using a quadratic loss function is that it is directly related to the definition of the conditional expectation via the L^2 norm in equation (21). The Bayesian theorem can be adapted to probability densities as

$$f(Y|X=x) = \frac{f_{X|Y=y}(x) f_Y(y)}{\int f_{X|Y=y'}(x) f_Y(y')\,\mathrm{d}y'} \tag{28}$$

with the familiar shape Posterior $= \frac{\text{Likelihood}\cdot\text{Prior}}{\text{Evidence}}$ and the *Evidence* being given by an integral version of the law of total probability. With the use of Fubini's theorem and the Bayesian theorem for conditional densities, then

$$r(\alpha, T) = \int_\Theta \int_\chi L(\theta, T(x)) \mathbb{P}_\theta(\mathrm{d}x)\alpha(\mathrm{d}\theta) = \int_\Theta \int_\chi L(\theta, T(x)) \mathbb{P}^{X|\tau=\theta}(\mathrm{d}x)\mathbb{P}^\tau(\mathrm{d}\theta) \tag{29}$$

$$\overset{Bayes}{=} \int_\chi \int_\Theta L(\theta, T(x)) \mathbb{P}^{\tau|X=x}(\mathrm{d}\theta)\mathbb{P}^X(\mathrm{d}x) = \int_\chi \mathbb{E}[L(\tau, T(x))|X=x]\mathbb{P}^X(\mathrm{d}x).$$

Therefore, the minimization of the Bayes risk r is equivalent to minimizing the conditional expectation of the quadratic loss function. But this is already the definition of the conditional expectation for the random variable τ. The best Bayesian estimator $\hat{\theta}_B(x)$ of θ is therefore

$$\hat{\theta}_B(x) = \mathbb{E}[\tau|X=x] \tag{30}$$

and the best Bayesian estimator of a function $\gamma(\theta)$ is

$$\hat{\gamma}_B(x) = \mathbb{E}[\gamma(\tau)|X=x]. \tag{31}$$

Note that equation (31) holds for any function $\gamma(\theta)$ of the unknown parameter θ and is not just restricted to e.g. linear transformations.

Conjugate Priors and an Example

Common prior densities for such estimation problems are the β and γ distributions. They are used for two reasons: First, their dependence on further hyper-parameters makes them very flexible and allows them to potentially encode many kinds of prior knowledge without changing the morphology of the equation. Second, they are so-called conjugate priors to many other distributions allowing them to lead to algebraically elegant calculations which

will be demonstrated in a brief example:

Consider independent random variables $X_1, \ldots, X_n$ with a Bernoulli-p distribution, meaning $\mathbb{P}(X_i = 0) = p = 1 - \mathbb{P}(X_i = 1)$ and a $\beta(a, b)$ distribution as a prior. The prior density is

$$g_{ab}(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \tag{32}$$

for $p \in (0, 1)$ and has $\mathbb{E}[\beta(a, b)] = \frac{a}{a+b}$. As a matter of formality, we introduce the random variable τ that is distributed according to the prior distribution $\beta(a, b)$. The joint probability density of the $X_{1,\ldots,n}$ for a fixed p is

$$f_p(x) = p^s(1-p)^{n-s} \tag{33}$$

with $s = \sum x_i$. For a quadratic loss function, the posterior distribution therefore becomes

$$h(p) = C(a, b, s)p^{a+s-1}(1-p)^{b+n-s-1}\mathbb{1}_{(0,1)}(p). \tag{34}$$

This has again the structure of a β-distribution, but with the parameters $\beta(a + s, b + n - s)$, meaning that the normalization constant is given as $C(a, b, s) = \frac{\Gamma(a+b+n)}{\Gamma(a+s)\Gamma(b+n-s)}$. According to equation (31), the Bayesian estimator for p is given by the expectation value of τ under the posterior density. Therefore, it is the expectation value of the new $\beta(a+s, b+n-s)$-distribution:

$$\hat{p}_B(x) = \frac{a+s}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{\sum x_i}{n}. \tag{35}$$

The Bayes estimator is therefore a weighted mean between the a priori estimator $\mathbb{E}[\beta(a, b)] = \frac{a}{a+b}$ and the maximum likelihood estimator $\overline{x} = \frac{1}{n}\sum x_i$ and for large n, the contribution of the prior decreases to 0. This makes sense, since for a large data sample, the prior assumptions should no longer have a strong impact on the evaluation, whereas they should matter significantly for a small sample size.

This example shows that the β-prior leads to a β-posterior. Similar calculations can be observed for γ-priors for other distributions and they simplify the analytical calculation. This characteristic is meant when addressing the

conjugate prior: the β-distribution is a conjugate prior for the Bernoulli distribution. However, the β- and γ-distributions have another characteristic that is very beneficial for prior distributions: They can be tailored to suit almost any shape by changing their parameters. For example, the prior in equation (32) turns into a uniform distribution for $a = b = 1$ and into a distribution concentrated at the extremes 0 and 1 for $a, b \gg 1$ which is still symmetrical for $a = b \gg 1$.

Frequent Priors: Uniform and Gaussian

One should also note that the uniform and the Gaussian distribution are also frequently used as prior distributions for parameters. Usually, they are chosen to simplify the algebraic expression of the posterior distribution. However, a more profound reasoning can be gained via the methods of information theory and maximum entropy as discussed in the appendix in A.1.

2.2.4 Information Criteria

Model selection based on data is an important aspect of empirical sciences. One approach to model selection is the calculation of an information criterion. The most well-known is the Aikake information criterion (AIC) and it includes a punishment for overly complicated models to prevent overfitting and promote sparsity. For k parameters in the model and the maximum likelihood value L of the model, the AIC is given by

$$\mathrm{AIC} = 2k - 2\log L \tag{36}$$

and a lower AIC is preferred according to the principle of Ockham's razor. Introducing an a-priori-distribution for the unknown parameters, the AIC can reevaluated in Bayesian terms. Chapter 9 of [35] shows that under certain conditions, the prior can be integrated out. Then, the Bayesian Information Criterion BIC

$$\mathrm{BIC} = k\log n - 2\log L \tag{37}$$

can be derived with n being the sample size of the data. Again, Ockham's razor is represented by minimizing the BIC and results in sparse model selection. A comparison of the two information criteria can be found in chapter

11.7 of [28].

2.2.5 Bayesian Philosophy

The inversion of the conditional probabilities in Bayes's equation (14) had stirred up a rather philosophical debate amongst statisticians. The inversion of probability, meaning the transition from $\mathbb{P}(D|M)$ to $\mathbb{P}(M|D)$ caused a feeling of uneasiness in statistics: the probability of the underlying model M based on the observed data D felt like trying to mess with causality and for a period of time, people refused to accept that the probability of "before" could be measured conditional on "after". Chapter 13 of [9] - correspondingly titled "The Bayesian Heresy" - gives a vivid description of how much criticism Bayesian ideas originally faced. There, Salsburg mentions that one of the pioneers of modern statistics, Ronald Fisher, once had to defend himself against the accusation of having used "inverse probability", as it was simply considered to be too unorthodox of a philosophy.

Another philosophically debatable aspect was ignored in section 2.2.3 but is worth to discuss. There, the prior distribution for the true parameter θ was incorporated into the statistical theory by replacing the fixed θ with a random variable τ whose distribution is given by the a priori density. But in fact, there is no random variable τ: θ is a fixed value and has no inherent element of uncertainty. If θ for example represents the probability of a radioactive decay of a specific atom, then θ is always the same, whereas the Bayesian approach pretends that θ is merely the result of a dice toss τ according to the a priori distribution. An albeit philosophically unsatisfactory way of dealing with these doubts is to regard the prior distribution as a mathematical artefact and auxiliary technique. However, one might also switch the perspective and claim that even although the true θ_{true} is indeed constant, one can only "subjectively" talk about a guess θ for this true parameter. Even the most precise measurement would still have some kind of measurement error and such uncertainties can be encaptured by the prior distribution.

A useful advantage of Bayesian methods is their application for learning processes. The prior encodes the initial knowledge and is updated based on the data to form a new prior for the next sample of data and so on.

Yet, this is only really successful if the true model is not ruled out by the prior distribution: if the possibility of a perfectly unfair coin is ruled out completely by the prior with probability 0, no streak of "heads" could force the posterior to favour the possibility that the coin always shows heads. The importance of not just "inductively" relying on updating the prior is discussed in [31]. There, the authors argue that the true strength of Bayesian methods will only be harnessed if the "falsificationist" view of classical statistics is also incorporated. They compare such an improved Bayesian analysis to the scientific theory of Kuhn and the falsification of a Bayesian model and its corresponding prior to Kuhn's transition to another scientific paradigm. They claim including model falsifications in Bayesian analyses was especially important in social sciences, since essentially all models in social sciences are known to be false, meaning that falsifying certain aspects of the model would be the essential scientific process in order to show the boundaries of these models.

2.3 Practical Examples of Bayesian Statistics

It is instructive to examine the concepts of Bayesian statistics by regarding some examples to test if the Bayesian approach actually makes a difference. The two examples of this section deal with frequent problems of statistical sciences, the linear regression and testing two hypotheses against each other. Especially the latter example shows a more nuanced and less obvious advantage of Bayesian methods.

2.3.1 Linear Regression

Linear regression means to interpret a data sample by assuming an underlying linear relationship between the different variables and to calculate a reasonable approximation for this linear function. Linear regression is one of the most important methods of data interpretation, as many functional relationships in nature and other sciences can be approximately described by a linear function. In this section, it will be demonstrated how a Bayesian model can help to create a more robust estimation for the linear function. For simplicity, we will restrict ourselves to the one dimensional function

$$y = \alpha + \beta x = f(x) \tag{38}$$

with the unknown parameters α and β. The measured data pairs (x_i/y_i) will be given by

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{39}$$

with the independent Gaussian random variables $\epsilon_i \sim \mathcal{N}(0, \sigma_0^2)$.

Least Squares: a Non-Bayesian Approach

A simple approach to estimate α and β is the least squares estimator (LSE) $(\hat{\alpha}/\hat{\beta})$ for (α/β), which is derived via minimizing the residual $R = \sum(y_i - f(x_i))^2$. This leads to the estimator

$$\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2} \qquad \text{and} \qquad \hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n \tag{40}$$

with the sample means and sample (co-)variance

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{41}$$

$$s_{xx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \tag{42}$$

$$s_{yy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 \tag{43}$$

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n). \tag{44}$$

This simple approach already leads to a very useful result, since it can be proven that the LSE is an unbiased estimator [7]. However, an outlier in the sample can distort the results dramatically. Simply choosing to neglect seemingly unreasonable data is an alluring solution to this problem, but does not correspond to the ethics of science and could also lead to negative effects such as ignoring the discovery of new phenomena. An example of this is given in the 1986 CODATA, where the authors commenting on the constant K_V concede that the "large change in K_V and hence in many other quantities between 1973 and 1986 would have been avoided if two determinations of F which seemed to be discrepant with the remaining data had not been deleted in the 1973 adjustment" [10].

Bayesian Regression

A different approach to deal with seemingly wrong data is to assume that the statistical error σ_0 is only a lower boundary to the 'true' error σ. This is a fairly realistic assumption, since the 'bad' data has been caused by *something*, which is apparently unknown to us. Attributing the apparent misfit of the data to an unknown error source and therefore to a $\sigma > \sigma_0$ is a sensible conclusion. If most data seem to correspond to the initial error σ_0, it makes sense to assume a prior distribution $p(\sigma)$ which attributes a high probability to σ close to σ_0. A possible choice is

$$p(\sigma) = \frac{\sigma_0}{\sigma^2} \mathbb{1}_{[\sigma_0,\infty)} \tag{45}$$

with the normalization $\int p(\sigma)\,\mathrm{d}\sigma = 1$ already included. For a single datum D, it is now possible to calculate the marginalized likelihood. The model prediction will be denoted as f and marginalization will make the unknown σ disappear. This leads to the likelihood

$$p(D|f,\sigma_0) = \int_0^\infty p(D|f,\sigma)p(\sigma|\sigma_0)\,\mathrm{d}\sigma \tag{46}$$

which can be solved by assuming a Gaussian density in $p(D|f,\sigma)$ consistent with the initial least squares assumption. This leads to

$$p(D|f,\sigma) = \frac{\sigma_0}{\sqrt{2\pi}} \int_0^{1/\sigma_0} te^{-t^2(f-D)^2/2}\,\mathrm{d}t \tag{47}$$

with the substitution $t = 1/\sigma$. This integral can be solved analytically and it is simplified by introducing the residual $R = (f-D)/\sigma_0$. This results in

$$p(D|f,\sigma_0) = \frac{1}{\sigma_0\sqrt{2\pi}} \frac{1-\exp(-R^2/2)}{R^2}. \tag{48}$$

If a uniform prior distribution is assigned to the parameters of the model f, the posterior distribution can be derived. Analogously to the standard least squares approach, the posterior probability $p((\alpha,\beta)|\vec{D}))$ is maximized to derive (α/β) with the vector $\vec{D}$ representing the measured data (x_i, y_i). This is equivalent to the easier maximization of the logarithm of the posterior. We find

$$\ln p(\alpha,\beta|\vec{D}) = \text{constant } + \sum_{k=1}^{N} \ln \frac{1-e^{-R_k^2/2}}{R_k^2} \tag{49}$$

with (α, β) included implicitly in $R_k = (f(x_k) - y_k)/\sigma_0$. Maximizing (49) leads to new, but more complicated estimators $(\hat{\alpha}/\hat{\beta})_B$. Figure 1 compares the least squares and the Bayesian estimations by using a sample set of data with one outlier. Whereas the LSE does not capture the real data trend, the Bayesian estimator proves to be quite resistant against the outlier, although the outlier is still included in the analysis and not just merely ignored. For a more detailed analysis of this example, cf. chapter 8.3.1 in [1], whereas an even more general approach to Bayesian regression problems is found in chapter 9 of [28].

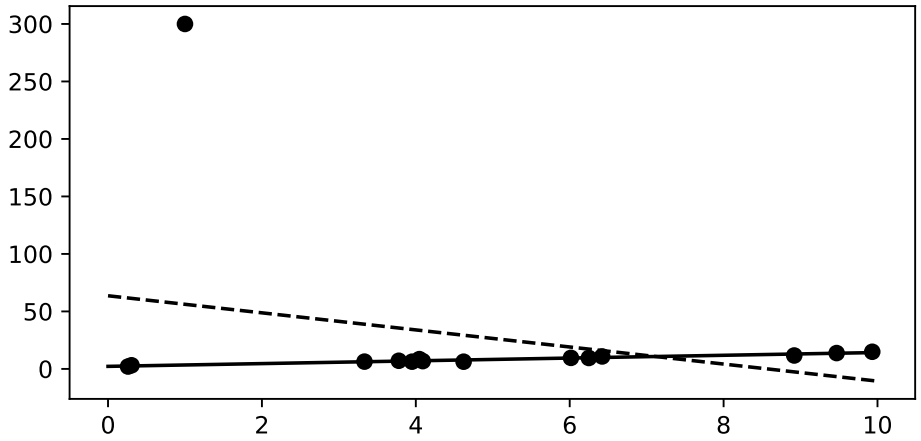


Figure 1: 10 data points (x/y) generated via equation (39) with $\alpha = 2.1$ and $\beta = 1.2$ including one outlier. LS estimates $(\hat{\alpha}/\hat{\beta}) = (63.6/ - 7.4)$ depicted as a dashed line, Bayesian methods estimate $(\hat{\alpha}/\hat{\beta})_B = (2.2/1.2)$ depicted as a solid line.

2.3.2 Bayesian Hypothesis Testing

The Bayesian framework allows a different approach to the problem of hypothesis testing which is most easily understood by regarding the case of comparing two different hypotheses or models M_1 and M_2 and observed data D. Note that the models do not have to be exhaustive, meaning that $\mathbb{P}(M_1 \cup M_2) < 1$ is permitted: the Bayesian approach does not aim to find the "one and only" correct model, but only compares two models against each other. The Bayesian approach regards the posterior odds ratio R given by

$$R = \frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)} \tag{50}$$

whose structure is reminiscent of equation (14). R factorizes into the prior odds ratio $\frac{P(M_1)}{P(M_2)}$ and the Bayes factor $B_{12} = \frac{P(D|M_1)}{P(D|M_2)}$ fulfilling $B_{12} = B_{21}^{-1}$. The prior odds ratio reflects, if there are a priori reasons to favour one model over the other and is frequently set to 1 for simplicity, making the Bayes factor the more significant component of equation (50). Following the description of R and B_{12} in chapter 11 of [28] and under a prior odds ratio of 1, a Bayes factor $B_{12} \in (0.1, 10)$ is usually regarded as being too close to 1 in order to make any claim that one of the models should be favoured over the other, but this insignificance interval can be tailored according to the preferences of the statistician.

The Bayesian approach has several advantages over the frequentist model selection. An overview of several examples can be found in [29] and an especially striking example is briefly discussed in [30] referring to a frequentist analysis of sociological data which has basically no other option than to reject a very promising hypothesis accounting for 99.7% of the data because of the large sample size of $n \approx 10^5$. The Bayesian analysis, however, comes to a different conclusion that supports the rejected hypothesis. Here, the Bayesian model comparison goes beyond the Bayes factor and relates to the Bayesian information criterion (BIC) which will be discussed later in this thesis. According to chapter 11.1 of [28], the major advantage of Bayesian hypothesis comparison over the frequentist approach is that it can successfully deal with systems, in which the data has a very low probability even under the true model. Another, more subtle advantage can be seen by regarding the following example from chapter 11.8 of [28]:

Example: Failure of the Frequentist Approach
Astronomers observed $n = 102$ stars including $r = 5$ white dwarves. The hypothesis M states that in such surveys, 10% of the observed stars are white dwarves. The true fraction of white dwarves is denoted as p and the corresponding statistical test is a two-sided test of $p = 0.1$ against $p \neq 0.1$.

Statistician A assumes that out of the n measurements, r white dwarves were observed and comes to the conclusion that these observations are given

by a binomial model with the distribution

$$\mathbb{P}(r|p,n) = \binom{n}{r} p^r (1-p)^{n-r} \tag{51}$$

and for the two-sided test, the doubled probability of observing no more than five white dwarves is given as

$$\mathbb{P}_A = 2\sum_{r=0}^{5} \mathbb{P}(r|p,n) \approx 0.102 > 0.05 \tag{52}$$

and therefore, the hypothesis $p = 0.1$ cannot be rejected at the 95% level.

However, statistician B assumes that the astronomers observed stars until they found $r = 5$ white dwarves. Then, the probability model is no longer binomial, but a negative binomial with n as the random variable and

$$\mathbb{P}_{\text{neg}}(n|p,r) = \binom{n-1}{r-1} p^r (1-p)^{n-r} \tag{53}$$

as the distribution. Therefore, the analysis of the two-sided test results in

$$\mathbb{P}_B = 2\sum_{n=102}^{\infty} \mathbb{P}(n|p,r) \approx 0.044 < 0.05 \tag{54}$$

and thus, statistician B would reject $p = 0.1$ at the 95% confidence level.

Both analyses are correct and which of them is applicable to the real situation depends on the measurement process. But this also means that in order to perform a frequentist analysis, one has to decide in advance which approach to follow. As pointed out in [28], this can lead to problems in the data analysis: e.g. if one decides to measure until 5 white dwarves are observed, but simply fails to observe the fifth white dwarf, then the data becomes problematic.

Example: Bayesian Success

The Bayesian approach to evaluating the data does not care about the details of the experimental setup, meaning that it is irrelevant whether statistician A or B performs the correct analysis of the data. Comparing the hypothesis

$p = 0.1$ with any other $p' \neq 0.1$ by computing the posterior odds ratio as in equation (50), the binomial coefficients cancel each other out. Hence, both approaches A and B arrive at the same posterior odds ratio R.

However, the task is not to design a pointwise test of $p = 0.1$ versus another p', but to test $p = 0.1$ versus every other $p' \neq 0.1$. This is done via marginalization and choosing a suitable prior density $f_{pr}(p)$ for p. The corresponding posterior odds ratio becomes

$$R = \frac{\mathbb{P}(\text{data}|p) f_{pr}(p)}{\int \mathbb{P}(\text{data}|p') f_{pr}(p') \, \mathrm{d}p'} \tag{55}$$

with $\mathbb{P}(\text{data}|p)$ representing either the binomial or the negative binomial distribution. However, the binomial coefficients still cancel each other because they do not depend on the integration variable p'. In both cases, the Bayesian analysis arrives at

$$R = \frac{p^r (1-p)^{n-r} f_{pr}(p)}{\int p'^r (1-p')^{n-r} f_{pr}(p') \, \mathrm{d}p'} \tag{56}$$

and the evaluation of the statistical significance is independent of the way the data was collected allowing for a greater flexibility than the frequentist approach.

"I pass with relief from the tossing sea of Cause and Theory to the firm ground of Result and Fact."

— Sir Winston Churchill

3 Methods

How can data $(x_i)_i$ be interpreted in order to construct a sparse ODE $f(x) = \dot{x}$? This section follows a Bayesian approach to this problem derived from Bayes's theorem "Posterior $\propto$ Likelihood $\cdot$ Prior".
With a general function $f_{\vec{C}}$ in mind, whose shape depends on coefficients C_i, we construct a likelihood density based on the propagator method of [36]: With given parameter values C_i, the transition probability from x_i to x_{i+1} is modelled as a Gaussian random process with transition density $\rho(x_{i+1}|x_i, \vec{C})$ in section 3.1.
A maximum likelihood estimation of the compound density

$$\max_{\vec{C}} \prod_i \rho(x_{i+1}|x_i, \vec{C}) \tag{57}$$

might be sufficient to accurately reproduce the data, but it can easily fall into the trap of overfitting, because it is highly unlikely that noisy or stochastic x_i will lead to sparse coefficients $\vec{C}$. But since many dynamical systems are only governed by simple and sparse ODEs, a prior distribution can be introduced to the estimation procedure to include the prior knowledge "it is likely to only have few coefficients $C_i \neq 0$". This prior density is encoded via hyper-parameters in section 3.2. Hence, the estimation problem becomes a Bayesian maximum posterior estimation of the hyper-parameterized likelihood function.
Finally, it can be insightful to regard the full posterior distribution instead of reducing it to e.g. its maximum value. This analysis can show correlations between different parameters and possibly reveal secondary maxima in the posterior density corresponding to unknown system dynamics. Such a fully Bayesian approach can be accomplished by expanding the sampling algorithm of 2.1.4 to a Markov Chain Monte Carlo sampler and is described in section 3.3.

3.1 Likelihood Approach to Diffential Equations

Estimating ordinary differential equations (ODEs) from data is a challenging task, especially if the ODE has a diffusion term. In this thesis, the ODEs were modelled with a general polynomial ansatz and the measured data is incorporated with a likelihood approach by computing the transition probabilities between the different measurements. This approach is discussed in this section.

3.1.1 Differential Equations with Dynamic Noise

Consider a differential equation $\dot{x} = f(x)$. If these dynamics are subject to dynamic noise, then the noise is not simply a measurement error, but part of the dynamics. For example, if $f(x)$ describes the movement of a particle through a gas of other particles, then random collisions with other particles can stochastically affect the trajectory. For simplicity, the noise is usually treated as a random variable $\epsilon_\sigma \sim \mathcal{N}(0, \sigma^2)$. One can heuristically derive a method to incorporate the dynamic noise into the numerical treatment of the differential equation:
Numerically, it makes sense to introduce the transformation $\sigma\epsilon = \epsilon_\sigma$ in order to regard random variables $\epsilon = \mathcal{N}(0, 1)$ as the sources of dynamic noise. In the context of collisions with other particles, the overall stochastic displacement is the sum of the individual collisions $\sigma\epsilon_{tot} = \sigma \sum_{i=1}^{n} \epsilon^{(i)}$. For independent random variables $\epsilon^{(i)}$, this results in the scaling law $\epsilon_{tot} \sim \mathcal{N}(0, n)$. Because of the properties of the normal distribution it follows that $\epsilon_{tot} \sim \sqrt{n}\mathcal{N}(0, 1)$, meaning the total noise can again be regarded as a standard normal distribution. For a given time interval $\mathrm{d}t$, the number of collisions should scale proportionally $n = \alpha\,\mathrm{d}t$. Introducing $q = \alpha\sigma$, the overall displacement D_{tot} in this time interval is distributed according to

$$D_{tot} = \sigma\epsilon_{tot} \sim \sigma\alpha\sqrt{\mathrm{d}t}\mathcal{N}(0, 1) = q\sqrt{\mathrm{d}t}\mathcal{N}(0, 1). \tag{58}$$

Hence, integrating the differential equation $\dot{x} = f(x)$ and taking into account the noise can be done via the Euler method. The time is discretized according to Δt and with equation (58), it follows that

$$x_{t+\Delta t} = x_t + f(x_t)\Delta t + q\sqrt{\Delta t}\epsilon \text{ with } \epsilon \sim \mathcal{N}(0, 1). \tag{59}$$

3.1.2 Stochastic Differential Equations with Propagator

Although not usually regarded as a Bayesian method, the use of propagators to analyze stochastic differential equations makes use of the conditional probabilities discussed in this thesis. Generalizing the ideas from section 3.1.1, the stochastic differential equation can be written as

$$\dot{z} = D_1(z,t) + D_2(z,t)\Delta t\sqrt{2}\epsilon \text{ with } \epsilon \sim \mathcal{N}(0,1) \tag{60}$$

with a rescaling factor of $\sqrt{2}$ for the stochastic component. The propagator defines the probability density for $z_{t+\Delta t}$ conditional on the value of z_t. In the Gaussian formulation of equation (60), this results in the propagator

$$\rho(z_{t+\Delta t}|z_t) = \frac{1}{\sqrt{2\pi V}} \exp\left(-\frac{1}{2V}(z_{t+\Delta t} - M)^2\right) \tag{61}$$

with mean $M = z_t + D_1(z_t,t)\Delta t$ and variance $V = 2\Delta t D_2(z_t,t)^2$. This distribution corresponds to a random walk with diffusion term $D_2(x,t)$.

Generalized to d dimensions, regard a Gaussian diffusion process $z \in \mathbb{R}^d$, D_1 as a vector and D_2 as a quadratic diffusion matrix. Then the propagator has the probability density

$$\rho(z(t+\tau), t+\tau | z(t), t) = \left((2\pi\tau)^d \det D_2\right)^{-1/2} \exp\left(-\frac{\tau}{2}\tilde{d}^T \cdot D_2(z,t)^{-1} \cdot \tilde{d}\right) \tag{62}$$

with $\tilde{d} = \frac{z(t+\tau)-z(t)}{\tau} - D_1(z,t)$. This approach to stochastic ODEs is discussed in detail in [36].

3.1.3 Maximum Posterior Estimation of the ODE

ρ in equation (62) is the likelihood for transitions from $z(t)$ to $z(t+\tau)$ based on the stochastic ODE in equation (60). This ODE consists of the deterministic $D_1(z,t)$ and the stochastic term $D_2(z,t)$. If our goal is to find out the shape of the ODE, one can make an ansatz

$$D_i(z) = \sum_{j=0}^{n} C_j^{(i)} \alpha_j^{(i)}(z) \tag{63}$$

with e.g. $\alpha_j^{(i)}(z) = z^j$ as model functions. A similar approach was done in [5], but with a more informed ansatz for D_i that already corresponded to a certain model of ODEs. The coefficients $C_j^{(i)}$ have to be estimated to decide which model functions are relevant or negligible. If the shape of equation (63) is inserted into the likelihood function (60) with known measurements $(z_t)_{t\in T}$, then one can maximize the likelihood ρ with respect to the coefficients $C_j^{(i)}$. Hence, the coefficients of the ODE and therefore the ODE itself can be estimated via this maximum likelihood approach.

The following equations show the general ansatz to this problem for N data points $z_{1\leq k\leq N} \in \mathbb{R}^d$ with $z_k = z(t_0 + kh)$. $D_1(z_k) \in \mathbb{R}^d$ is a vector, $D_2(z_k) \in \mathbb{R}^{d\times d}$ is a diffusion matrix evaluated at a point z_k of the trajectory. The log-likelihood function $\mathcal{L}$ is then given by

$$\mathcal{L} = \sum_{k=1}^{N} \frac{-r(z_k)}{2h} - \log\left(\sqrt{4\pi h}^d \sqrt{\det D_2(z_k)}\right) \tag{64}$$

with the temporal discretization h and the residual

$$\begin{aligned} r(z_k) &= \tilde{D}_1^T(z_k) \cdot (D_2^{-1}(z_k) \cdot \tilde{D}_1(z_k)) \text{ with} \\ \tilde{D}_1(z_k) &= (z_{k+1} - z_k) - hD_1(z_k). \end{aligned} \tag{65}$$

This is derived from a multi-dimensional Gaussian treatment of the propagator outlined in [36]. The likelihood function L is basically a combination of all transition probabilities $\rho(z_{t+\Delta t}|z_t)$ and has to be maximized to estimate the most suitable $C_j^{(i)}$.

In the Bayesian approach to statistics, this maximum likelihood approach can be modified to a maximum posterior approach by introducing prior distributions for the coefficients $C_j^{(i)}$. One simple example of this is to attribute a flat prior to all deterministic $C_j^{(1)}$ and to regard the diffusion term as a constant $D_2(z,t) = q$. Then, an exponential prior distribution for q can ensure that the noise is always strictly positive in order to ensure a mathematically sensible estimation. However, the prior density can also be included via hyper-parameters which are subject of the next section.

3.2 Hyper-Parameters

In the research on machine learning, hyper-parameters are all parameters of the algorithm that are not updated by the learning process, but instead provide the framework of how the algorithm has to work. In an analogy, one can regard a Runge-Kutta method with adaptive step-size h to approximate the solution of a given ordinary differential equation numerically. The step-size h is changed by the algorithm to suite the needs of the given problem, but the Butcher tableau is given beforehand and remains unchanged throughout the entire calculation. In this sense, one could interpret the Butcher tableau as the hyper-parameter of the Runge-Kutta methods.

3.2.1 Hyper-Parameters as the Prior Density

Usually, hyper-parameters are used to enforce sparsity in an estimation problem. This means that out of the coefficients C_1 to C_n of the estimation problem, only a few of them should have a nonzero value. There are two reasons why desiring sparsity can be justified: First, it prevents an overfitting of the estimation algorithm. Overfitting means that the estimation algorithm constructs an extremely complicated function that perfectly fits the data but is unable to produce a reasonable prediction for unknown data. Second, a sparse solution helps to identify the actually relevant contributions to e.g. the system dynamics and their real world counterparts.

Fundamentally, introducing sparsity to a system is equivalent to encoding the a priori information "only few of the coefficients are larger than zero". Hence, it is unsurprising that there exists a peculiar relationship between prior densities and hyper-parametrized regularization methods. For example, the *Lasso* method described in [22] can change the standard least squares estimation of a function with coefficients $\vec{C}$

$$\min_{\vec{C}} \left(\sum_i (f_{\vec{C}}(x_i) - y_i)^2 \right) \tag{66}$$

to become

$$\min_{\vec{C}} \left(\sum_i (f_{\vec{C}}(x_i) - y_i)^2 + \lambda \sum_j |C_j| \right). \tag{67}$$

How does this promote sparsity? Equation (67) does not only have to mini-

mize the distance between data y_i and the estimator $f_{\vec{C}}$, but simultaneously has to minimize the sum of the absolute values of coefficients $\vec{C}$ that contribute to f. Hence, an overly complicated f with many nonzero C_i will receive a large penalty because of the second term in (67) and will be discouraged by the minimization algorithm.

If equation (66) with its quadratic residua is interpreted as the logarithm of a Gaussian likelihood function, then how to interpret the second term in (67)? Inserting it into an exponential function yields a multidimensional $Exp(\lambda)$ distribution for the absolute values $|C_i|$ and this can be interpreted as a compound prior distribution for the coefficients $|C_i|$. Similarly, if the second term in (67) does not summate the absolute values, but the quadratic values of C_i^2 (this is called the *Ridge* method), then it can be interpreted as a Gaussian prior density with variance $\sigma^2 \propto \lambda^{-1}$.

3.2.2 Hyper-Parameter Optimization

Obviously, the choice of hyper-parameters can change the outcome of the algorithm drastically. Depending on what the algorithm is supposed to achieve, there exist "better" and "worse" hyper-parameters. Ideally, one hopes to select the optimal hyper-parameters for the given problem, or at least parameters close to the optimum. Thinking back to the Runge-Kutta example, the Runge-Kutta 4 method might qualify as the most optimal method for many applications, since it provides a good accuracy with relatively little computational effort. Without divine inspiration, it may however be hard to find a suitable set of hyper-parameters; especially when the algorithm itself is so time-consuming that one cannot simply perform a large-scale test of all possible parameters. This problem will be discussed in the following sections of this thesis which are based on the articles [17], [18] and [19]. There are several possibilities to search for an optimal set of hyper-parameters, including intuitive methods like random searches or an iteration along a discrete grid. One of them, the sequential model-based optimization, will be discussed in this thesis.

3.2.3 Sequential Model-Based Optimization

Consider an objective function $f : \chi \to \mathbb{R}$ that maps hyper-parameters $x \in \chi$ to a score $y \in \mathbb{R}$. E.g. one can imagine an algorithm fitting a function F to a set of given data X according to an interpolation method specified by x and f returns the mean squared difference between the original data and the interpolated function. In this case, an optimal set of hyper-parameters suffices $f(x_o) \stackrel{!}{=} \arg\min_x f(x)$, but other problems might require a maximization of f. The transformation $f \mapsto -f$ naturally combines both of these expressions. For simplicity, we will restrict ourselves to calculating the minimum of such functions.

Often, the evaluation of $f(x)$ is very time-consuming. Therefore, it is reasonable not to calculate $f(x)$ in every single iteration, but to define a suitable replacement for f: First, a (probabilistic) model $\mathcal{M}_0$ is defined using some samples of χ evaluated under f. These initial data points constitute the history H of the evaluations of f. Based on this knowledge, another set of hyper-parameters $x \in \chi$ is selected by evaluating the surrogate function S that is supposed to accurately approximate f at a much lower computational cost. The new x_n is defined via $x_n = \arg\min S(x|\mathcal{M}_0)$ and once it has been calculated, the true function $f(x_n)$ is evaluated. $(x_n, f(x_n))$ is added to the history of data, which allows us to update the model $\mathcal{M}_0$ to $\mathcal{M}_1$. As a pseudo code, this is given in algorithm 1.

$H = ((x_1, ..., x_n), f(x_1, ..., x_n))$
$\mathcal{M}_0 = \mathcal{M}(H)$
while $t<T$ **do**
  $x_{n+t} = \arg\min S(x|\mathcal{M}_t)$
  $H = H \cup \{(x_{n+t}, f(x_{n+t}))\}$
  $\mathcal{M}_t = \mathcal{M}(H)$
end
return H

Algorithm 1: Pseudo code of a hyper-parameter estimation.

But how can we transform the given history H of the data into a suitable model $\mathcal{M}$ that the surrogate function S can evaluate? A reasonable and frequent choice for S is the expected improvement $\mathbb{E}[I_{y^*}(x)]$ defined for a threshold value y^* of f that is derived from the quantile γ as $p(y < y^*) = \gamma$

of the observed $y = f(x)$ values. The expected improvement reads

$$\mathbb{E}[I_{y^*}(x)] = \int_{-\infty}^{y^*} (y^* - y) p(y|x) \, \mathrm{d}y \tag{68}$$

and can be interpreted as follows: As $y = f(x)$ should be minimized, only $y < y^*$ should be regarded, if e.g. y^* is selected as the previously best value. For such y, the improvement compared to the old y^* is (positively) measured as $(y^* - y)$. The improvement given by choosing a hyper-parameter x is therefore $((y^* - f(x))$ or zero, if $f(x)$ is not an improvement.
However, the improvement is not completely known, as evaluating $f(x)$ is supposed to be avoided. Therefore, the model $\mathcal{M}$ is used: a conditional probability distribution $p(y|x)$ is defined to estimate which values of $y = f(x)$ might occur for a given x. The model $\mathcal{M}$ therefore reflects the Bayesian characteristic of this approach.

3.2.4 Probabilistic Regression Model: Tree-Structured Parzen Estimator TPE

There are several suitable models $\mathcal{M}$, but the tree-structured Parzen estimator (TPE) described in [17] has a significant advantage: defining a probability distribution $p(y)$ of the objective function values is not necessary. The conditional probability $p(y|x)$ in (68) can be transformed via Bayes's rule $p(y|x) = p(x|y)p(y)/p(x)$. With the TPE approach, $p(x|y)$ is defined as

$$p(x|y) = \begin{cases} l(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^* \end{cases} \tag{69}$$

and is therefore considered as a combination of two distributions $l(x)$ and $g(x)$. The density $l(x)$ is formed by using those observations x_i from H which fulfil $f(x_i) < y^*$, whereas the remaining observations are used to form $g(x)$. The densities $g(x)$ and $l(x)$ are calculated via a Parzen estimator (cf. 2.1.3) based on their respective sample of observations. Why this particular separation is used will become clear at the end of this chapter.

With the Bayesian theorem, the expected improvement can be rewritten

$$\mathbb{E}[I_{y^*}(x)] = \int_{-\infty}^{y^*} (y^* - y) \frac{p(x|y)p(y)}{p(x)} \, \mathrm{d}y = \frac{1}{p(x)} \int_{-\infty}^{y^*} (y^* - y) p(x|y)p(y) \, \mathrm{d}y \tag{70}$$

and since $\gamma = p(y < y^*)$, it also holds that via marginalization

$$p(x) = \int p(x|y)p(y) \, \mathrm{d}y = \gamma l(x) + (1 - \gamma) g(x). \tag{71}$$

Additionally, the integral in equation (70) can be simplified to

$$\begin{aligned} \int_{-\infty}^{y^*} (y^* - y) p(x|y)p(y) \, \mathrm{d}y &= l(x) \int_{-\infty}^{y^*} (y^* - y) p(y) \, \mathrm{d}y \\ &= l(x) y^* \gamma - l(x) \int_{-\infty}^{y^*} y p(y) \, \mathrm{d}y. \end{aligned} \tag{72}$$

It is noteworthy, that the literature usually forgets the y in the final integral of (72), but this does not change the basic conclusion from this calculation. Combining these simplifications yields

$$\mathbb{E}[I_{y^*}(x)] = \frac{l(x) y^* \gamma - l(x) \int_{-\infty}^{y^*} y p(y) \, \mathrm{d}y}{\gamma l(x) + (1 - \gamma) g(x)} = \left(\frac{\gamma l(x) + (1 - \gamma) g(x)}{l(x) y^* \gamma - l(x) \int_{-\infty}^{y^*} y p(y) \, \mathrm{d}y} \right)^{-1} \tag{73}$$

$$= \left(\frac{\gamma + (1 - \gamma) \frac{g(x)}{l(x)}}{y^* \gamma - \int_{-\infty}^{y^*} y p(y) \, \mathrm{d}y} \right)^{-1} \propto \left(\gamma + (1 - \gamma) \frac{g(x)}{l(x)} \right)^{-1}$$

with the proportionality being valid despite this correction of the literature, since the denominator has no direct dependence on x. Note that to derive equation (73), no approximations were used. The maximum of the expected improvement is therefore achived by minimizing the ratio $g(x)/l(x)$. The new x_n selected via the surrogate model is therefore given by $\arg\min_x(g(x)/l(x))$. This observation helps to justify the separation of $p(x|y)$ into $l(x)$ and $g(x)$. Knowing that $l(x)$ reflects the probability density of x based on "good" measurements $y < y^*$ and $g(x)$ the density based on "bad" measurements, the criterion for selecting x_n reads as follows: x_n should simultaneously have a high probability under the distribution based on the good measurements $l(x)$ **and** a low probability under the distribution based on the bad measure-

ments $g(x)$. One can imagine that in particularly difficult problems, $g(x)$ and $l(x)$ are close and overlapping distributions. Simply maximizing $l(x)$ would not be a sufficient criterion on its own, since the same x could also have a high probability under the undesirable condition $y \geq y^*$. Hence, the ratio $(g(x)/l(x))$ ensures to be a useful criterion even for such a difficult problem.

3.2.5 Testing the TPE

Before considering the problem of estimating hyper-parameters of real models, one can test the strength of the TPE Estimator with a mock model. Here, a quadratic function is used as $f(x)$ with a minimum in $x = 1.5$ and a small amount of noise to act as a loss function for a real hyper-parameter x whose optimum should be estimated as $x_o \approx 1.5$ by the TPE estimator. Since the idea behind the TPE method is to minimize the calculation of the real function $f(x)$, only a $n = 5$ new value pairs $(x/f(x))$ are calculated in each iteration of the programme. Figure 2 (top) shows the final calculation of the densities $g(x)$ and $l(x)$ and of the estimated x_{est}. In 100 iterations, it the average estimator for $x_{min} = 1.5$ was 1.53 with an empirical variance of 0.03, thereby proving that the TPE method is a fairly reliable estimation for this problem.

However, one has to stress that the introduction of noise to this particular system proved to be significantly more challenging than the deterministic $f(x)$. Without a sufficiently strict threshold that ensures that only few values are used to calculate $l(x)$, the wide and noisy minimum of $f(x)$ allowed the same x to be sorted into both $l(x)$ or $g(x)$ if it was calculated repeatedly. Therefore, the system struggled to clearly differentiate between the "good" measurements for $l(x)$ and the "bad" ones for $g(x)$. Another problem of this method is that for large numbers n of iterations, the densities $g(x)$ and $l(x)$ become more and more similar. This is indicated by comparing the final density estimation at the top of figure 2 with the initial one at the bottom of figure 2. For large n and a not sufficiently strict threshold, the densities can converge to basically identical shapes which makes the evaluation $\arg\min g(x)/l(x)$ of the surrogate model rather a matter of noise than of actual statistical properties. However, the TPE method should only be used for small n anyways since its aim is to minimize the amount of time

spent on calculating $f(x)$. Hence, it may me justified to ignore the latter problem of the TPE method.

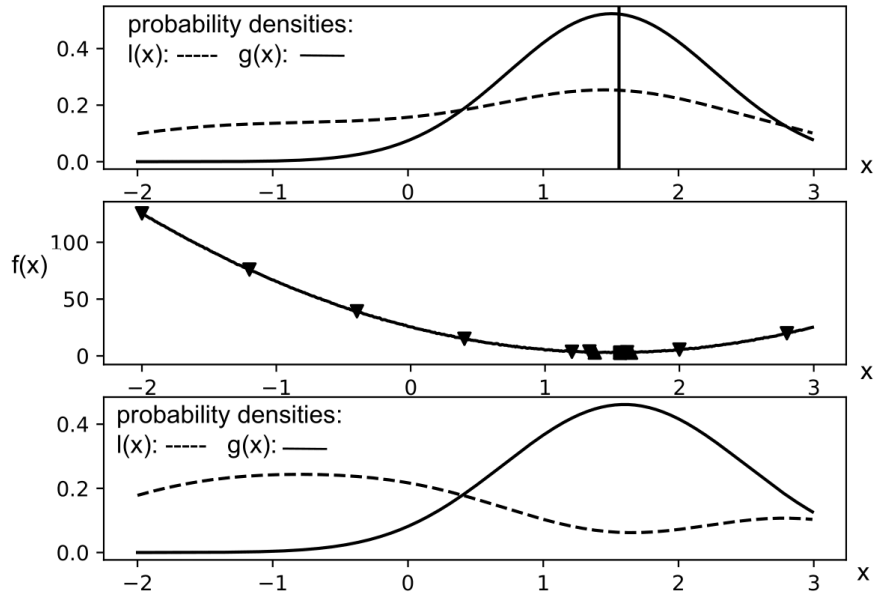


Figure 2: TPE results for the fifth estimation giving $x_{est} = 1.56$. Top: The densities $g(x)$ as a dashed and $l(x)$ as a solid line with the vertical showing the position of the estimated x. Centre: Data history H with upward triangles contributing to $l(x)$, downwards ones to $g(x)$ and the newly estimated x as a square. Bottom: $g(x)$ and $l(x)$ for the initially measured sample before the first estimation.

3.2.6 Comparison: Gaussian Process vs. TPE

Another example of hyper-parameter estimation is the use of a Gaussian process. Here, a Gaussian random process with mean μ and covariance matrix Σ explores the space of hyper-parameters. An exhaustive description of the construction and analysis of the Gaussian process can be found in [23] and in in [24]. Several comparative studies have been made to showcase the advantages and disadvantages of the Gaussian Process method and the TPE method. In [17] and [32] both methods were used on sample problems and compared with respect to convergence speed, quality of the mean solution and the variability of the solutions. Both of these empirical studies

concluded that the TPE method showed an overall superior performance for the examined problems. A less empirical comparison with respect to their applicability in neural networks can be found in [33] with specific regards to the technical details.

A conceptual disadvantage shared by both methods is their own reliance on hyper-parameters: For the Gaussian Process, the shape of the covariance matrix can be interpreted as a hyper-parameter, and for the TPE, the choice of kernel, the separation percentile indicated by y' in (69) and especially the bandwidth are hyper-parameters. Interestingly, there exists an optimality criterion for the "best" bandwidth as outlined in [34], but it is only applicable for high numbers of data. However, the possibly very restrictive choice of good data points to construct $l(x)$ often leads to a very low sample of points to calculate $l(x)$ via the kernel density method. Hence, the optimal bandwidth unfortunately does not apply to this situation.

3.2.7 More than one Hyper-Parameter: Elastic Net

Naturally, the question arises whether the hyper-parameter estimation can be extended to multidimensional problems. The multidimensional TPE and several regularization methods will be presented in this section.

Multidimensional TPE

Since the concept of kernel estimators is applicable to dimensions $n > 1$, there seems to be no formal obstacle to use the method described in section 3.2.4 for multidimensional situations. For simplicity, one can regard a problem with two hyper-parameters $\lambda_{1,2}$, a generalized regression problem with two penalty functions. These hyper-parameters are included in a minimization problem

$$\min_{\beta} \left(\sum_i (f_\beta(x_i) - y_i)^2 + \lambda_1 \sum_i |\beta_i| + \lambda_2 \sum_i \beta_i^2 \right) \tag{74}$$

known as the *Elastic Net* regularization. Because not only the residuum $(f-y)^2$, but also the absolute values and the quadratic values of the model parameters have to be minimized simultaneously in equation (74), the minimizing algorithm is encouraged to set superfluous parameters to 0. This

method is a generalization of two other regression methods: $\lambda_2 = 0$ corresponds to the Lasso or L^1 regression, $\lambda_1 = 0$ is known as the Ridge method or L^2 regularization.
In order to keep the problem simple, a strictly linear model

$$f_\beta(x) = \beta_0 + x\beta_1 \tag{75}$$

is regarded. The measured data is split into training data, based on which equation (74) suggests $\beta_{0,1}$ for fixed $\lambda_{1,2}$. With these parameters, the squared error $\sum_{x_{\text{test}}}(f_\beta(x) - x)^2$ is computed to test if the algorithm (and hence the choice of $\lambda_{1,2}$) is reasonably accurate. According to this error, the tested $\lambda_{1,2}$ are split into groups of low and high errors according to algorithm 1 and multidimensional densities $l(\lambda)$ and $g(\lambda)$ are calculated and evaluated.

For this given situation, 100 random initial $\lambda_{1,2}$ and their errors were computed and 100 further $\lambda_{1,2}$ were selected via the TPE method. The testing and training data were given uniformly distributed errors $y = f(x) + \epsilon$ with ϵ distributed according to $\text{Unif}_{(-\theta,\theta)}$ and $f(x) = 5x - 4$ as the true function. Figure 3 shows the TPE hyper-parameter estimation for this model for two different levels $\theta = 0.01$ and $\theta = 0.5$ of noise. Since $\theta = 0.01$ is an almost deterministic data generation, one would expect a fitting error close to zero which is achieved by the TPE method. The higher noise $\theta = 0.5$ prevents such an accurate regression, but the results in figure 3 show a convergence to an error of ≈ 0.7 which seems to be the best achievable error.

Besides this simple proof of concept, one can also regard a more practical example. The model of equation (75) can be modified to

$$f_\beta(x) = \beta_0 + x\beta_1 + x^2\beta_2 \tag{76}$$

in order to include the possibility of a parabolic shape. Since the real data follows a linear function, a good hyper-parameter estimation would result in a very low $|\beta_2|$. Figure 4 shows the error convergence for such a problem with the final iteration returning hyper-parameters $\lambda_{1,2}$ leading to the estimators $\beta_{0,1,2} \approx (4.96, -3.84, 0.001)$ and therefore, $|\beta_2| << |\beta_{0,1}|$ is achieved. The multidimensional TPE correctly predicts a linear instead of a quadratic model.

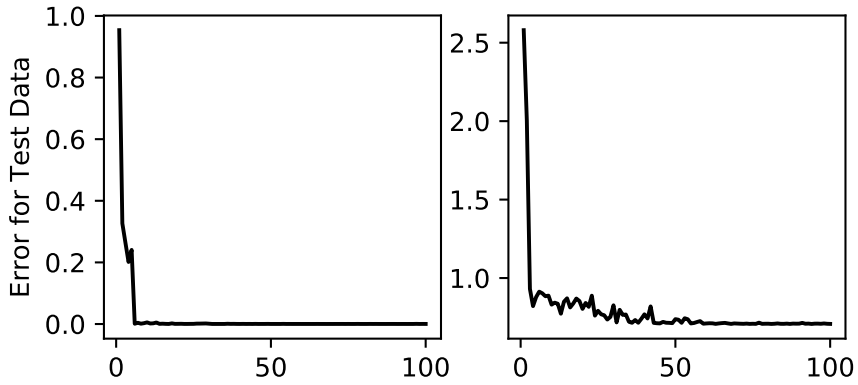


Figure 3: Estimating the optimal $\beta_{0,1}$ via TPE and the Elastic Net method as of equations (74) and $f(x) = 5x - 4 + \epsilon$ with uniform noise $\epsilon \in (-\theta, \theta)$. The number of iteration is on the ordinate.
Left: Low noise $\theta = 0.01$ shows quick convergence to ≈ 0.
Right: Medium noise $\theta = 0.5$ shows slower convergence and the apparently best achievable error is $0.7 > 0$.

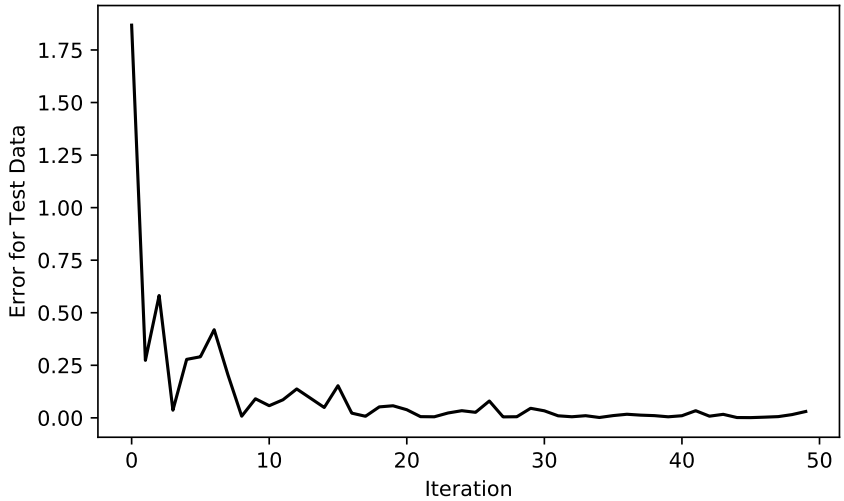


Figure 4: Estimating the optimal $\beta_{0,1,2}$ via TPE and the Elastic Net method as of equations (76) and $f(x) = 5x - 4 + \epsilon$ with uniform noise $\epsilon \in (-0.01, 0.01)$. The error converges to ≈ 0 within few iterations.

3.3 Markov Chain Monte Carlo

The simple rejection sampling algorithm discussed in section 2.1.4 has a major disadvantage, because much time is spent on regions with low probability that do not add any elements to the sample. The Markov Chain Monte Carlo (MCMC) algorithms alleviate this problem: they spend most time on regions of the sample space with high probability $f(x)$ and only waste little time on the regions with small probability $f(x)$, but manage to still draw a representative sample from the distribution.

MCMC is a random process $x_1, x_2, ...$ across the parameter space with the transition probability from x_n to x_{n+1} being given by $\tau(x_{n+1}|x_n)$ so that transitioning into regions with a high probability density $f(x)$ should be favoured. However, the Markov chain's stationary distribution has to be equal to the (normalized) density $f(x)$. The Metropolis-Hastings algorithm fulfills these requirements.

3.3.1 MCMC: Metropolis-Hastings

Similar to rejection sampling, the Metropolis-Hastings algorithm requires a distribution $Q(s|x_n)$ which can easily be sampled. Usually, Q is chosen as a multivariate Gaussian since most programming languages include a sampling method for Gaussian densities. Generally speaking, $Q(s|x_n)$ should favour points nearby x_n over those that are far away.

A candidate s is sampled from $Q(s|x_n)$. To decide, whether or not s is accepted, the Metropolis ratio

$$\rho = \frac{f(s)}{f(x_n)} \frac{Q(x_n|s)}{Q(s|x_n)} \tag{77}$$

is evaluated. If $\rho \geq 1$, s is accepted and $x_{n+1} = s$. If $\rho < 1$, s is only accepted with a chance ρ. If it is rejected, then $x_{n+1} = x_n$. Hence, the random process will always walk towards more probable regions $\rho \geq 1$, but still has a chance $0 < \rho < 1$ to explore the less probable areas. Although it is not obvious, the Metropolis-Hastings algorithm will converge to a representative set of samples for the distribution $f(x)$, although it may be advisable to discard the initial samples as a so-called "burn-in". A more in-depth discussion

of the Metropolis-Hastings algorithm can be found in chapter 8.5 of [28]. Note that usually, not one Markov Chain is used to sample data, but many "walkers" explore the parameter space.

3.3.2 MCMC: Advantages or the Posterior Distribution

The MCMC algorithm produces large samples of multidimensional data x_i. The single entries $x_j^{(i)}$ of x_i do not have to be free parameters of the observed model, but can be expanded to include nuisance parameters like hyper-parameters (which are discussed in section 3.2). This large sample can immediately be marginalized in the Bayesian sense to e.g. only show the distribution of one parameter $\{x_1^{(i)}\}_{1\leq i\leq n}$: since the other parameters $x_j^{(i)}$ for $j \neq i$ are already included in the sample in a ratio that is representative for the overall distribution, one can simply ignore them and only look at the first column of the sample table. Similarly, one can observe correlations between two parameters $x_j^{(i)}$ and $x_k^{(i)}$ by e.g. producing a scatter plot of the k^{th} and j^{th} column of the table, because (again) all other variables and their possible correlations are already implicitly included in the sampled data. Hence, MCMC can be used to gain much knowledge about a density $f(x)$ and the possible relationships between its parameters, whilst also being a computationally efficient and time-saving method.

"Premature optimization is the root of all evil."

— Donald Knuth

4 Results

The numerical data of two ODEs has been observed and analyzed in order to estimate sparse ODEs with the TPE-based MLE method and with the MCMC method. Since the MLE method is based on similar ideas as the SINDy method in [26], these two methods and their major differences will briefly be discussed in section 4.2.6, before moving on to the results of the MCMC method. Both approaches manage to yield sparse estimations of the ODEs.

4.1 Observerd Systems

In order to test the different methods for sparse ODE estimations, the Lorenz system and the van der Pol oscillator are analyzed. This section briefly describes their algebraic shape and their important properties

4.1.1 Lorenz System: Introduction

The Lorenz system is one of the most well-known dynamical systems. It is described extensively in most books on nonlinear systems, e.g. [27] devotes the entire chapter 9 to the study of the Lorenz system. Here, only the differential equation

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \vec{f}(x,y,z) = \begin{pmatrix} \sigma(y-x) \\ rx - y - xz \\ xy - bz \end{pmatrix} \tag{78}$$

is relevant, but more on the Lorenz system and its chaotic behaviour can be found in [27]. The data is generated with the standard parameters $\sigma = 10, b = 8/3, r = 28$ to observe trajectories in the chaotic regime and noise is added as dynamical noise $D_2(x,y,z) = D_2$ of (x,y,z).

4.1.2 Van der Pol Oscillator

The Van der Pol oscillator is an oscillator with a nonlinear damping and its dynamics follow the equation

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0. \tag{79}$$

Introducing the transformation $y = \dot{x}$, the differential equation becomes

$$\dot{x} = y \qquad \dot{y} = \mu(1 - x^2)y - x. \tag{80}$$

The trajectory approaches a limit cycle whose dependence on μ is illustrated in figure 5. The data used by the estimation algorithms is calculated numerically according to equation (80) with the Euler method and adding dynamical noise.

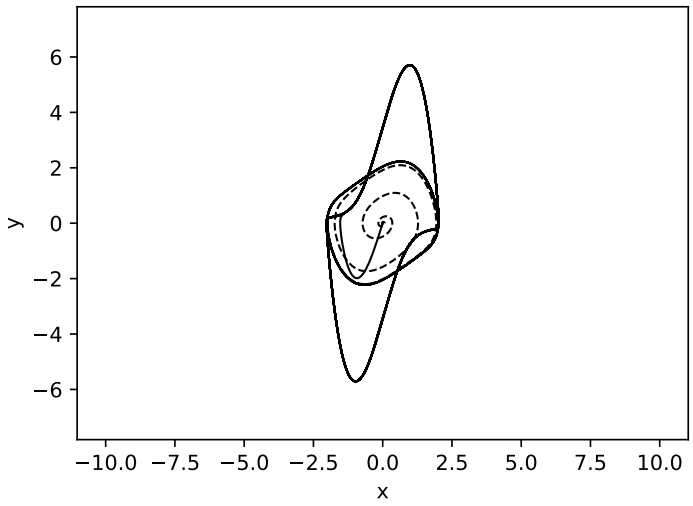


Figure 5: Van der Pol oscillators with $\mu = 3.5$ (solid) and $\mu = 0.5$ (dashed). The lower μ approaches a more circular limit cycle.

4.2 MLE Propagator with Hyper-Parameter

The propagator of equation (61) is analyzed with a polynomial ansatz for the time-independent deterministic part $D_1(z)$ of the differential equation.

This results in the approach

$$D_1(z) = C_0 + C_1 z_1 + C_2 z_2 + C_{11} z_1^2 + \dots \tag{81}$$

and can be treated with a maximum likelihood method with respect to the coefficients C. The dynamical noise $D_2(z)$ is also regarded as a variable by the maximum likelihood method. Constant noise will be denoted as $q = D_2 \neq D_2(z)$, but a case with a functional dependence of D_2 on the position in the phase space will also be regarded in section 4.2.4.

4.2.1 Suitable Score Function and Regularization Method

In order to separate good and bad hyper-parameters according to equation (69), a suitable score function is needed. Comparisons between test and training data have been examined for this purpose as well as the difference between forward integration and the remaining data. The most useful method turned out to be the Bayesian Information Criterion BIC (37).

Lasso and Ridge regularization methods have been tested for the MLE propagator method, but their dependence on the hyper-parameter turned out to be miniscule: unlike e.g. the estimation of a polynomial function from data, the estimation of the right-hand side of an ODE seems to be such a fragile concept that even a "bad" Lasso or Ridge hyper-parameter cannot thwart the basic trend of the parameter estimation. On the other hand, these methods also struggled to achieve real sparsity for systems with noticeable noise and still attributed small nonzero coefficients to several terms. One can introduce sparsity to these results by setting all coefficients with an absolute value below a certain threshold as zero. Then, however, it is more instructive to immediately use the thresholding method as regularization method and the threshold as the hyper-parameter. This method is also suggested as the best-practise in [26] for the SINDy method. Remembering that the Lasso and Ridge regularization can be interpreted as exponential or Gaussian prior distributions for the parameters, the thresholding can be understood as a prior density with a Heaviside shape.

Hence, the threshold λ was used as the hyper-parameter of a TPE-based hyper-parameter optimization with the BIC as the score function. After

thresholding once, the remaining parameters were used for a second maximum likelihood estimation, after which another thresholding and a final maximum likelihood estimation was performed. This procedure directly introduces a great level of sparsity to the system, because wrong nonzero contributions can no longer cancel each other out. In the following examples, a third order polynomial ansatz was used for the right-hand side of each ODE system.

4.2.2 Van der Pol Oscillator

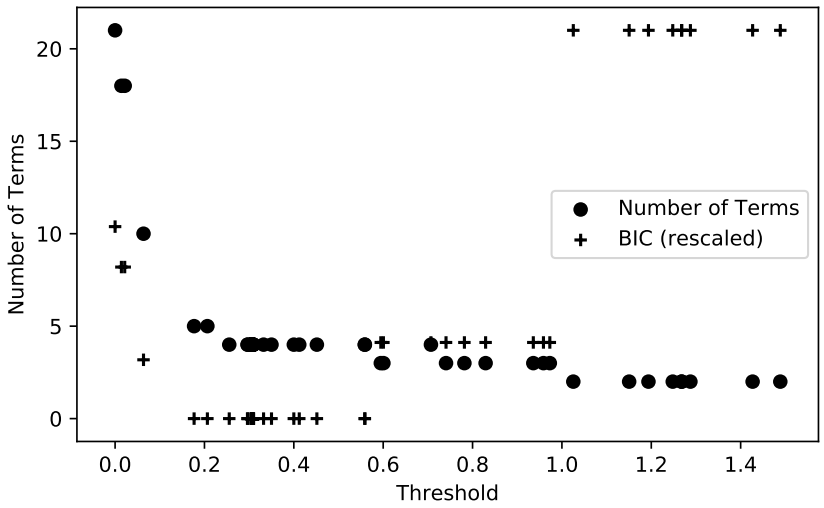


Figure 6: Van der Pol oscillator with $\mu = 5.5$ and dynamical noise $q = 0.25$. Dots show the number of nonzero terms estimated via MLE, crosses show the BIC (rescaled to fit into the graph).

Numerical data with $\mu = 5.5$ and $q = 0.25$ was used to estimate the right-hand side of the Van der Pol ODE and the level of dynamical noise q. As the results of the hyper-parameter optimization in figure 6 shows, the flat minimum of the BIC can be found for threshold values $\lambda \in (0.2, 0.6)$. As it is to be expected, the number of terms with nonzero coefficients decreases with increasing λ. Above the region of optimal λ, decreases of the number of terms sometimes coincide with increases of the BIC. These λ values signify that the threshold now starts to ignore terms that are essential to the system's dynamics. Below the region of optimal λ, the higher BIC values are caused

by the drastic increase in the number of terms adding a penalty to the BIC. The MLE procedure estimates the right-hand side and noise as

$$\begin{aligned} \dot{x} &= 0.90y \\ \dot{y} &= -0.99x + 5.33y - 5.28x^2y \\ q &= 0.25 \end{aligned} \tag{82}$$

and shows a high agreement with the real equations in (80). No superfluous terms are left after the threefold iterative thresholding: the estimated equations are perfectly sparse.

4.2.3 Lorenz System

The standard chaotic Lorenz system has been evaluated twice with the MLE method: Once with a high level of dynamical noise of $q = 10.$ and once in the deterministic limit of $q = 10^{-3}$. Figures 7 and 8 show the evaluations of the hyper-parameter optimizations in these cases. Again, a plateau of minimal BIC can be seen and the behaviour for λ above this flat minimum mimics the behaviour described in section 4.2.2. The increase of the BIC below the plateau can be seen, but is not as noticeable as it was in figure 6.

The parameter estimation for the noisy data with $q = 10$ results in

$$\begin{aligned} \dot{x} &= 9.88y - 9.95x \\ \dot{y} &= -0.39 + 27.88x - 0.99y - 1.00xz \\ \dot{z} &= 3.12 - 2.81z + 1.01xy \end{aligned} \tag{83}$$

$$q = 10.25 \tag{84}$$

and shows an accurate estimation of the Lorenz system parameters with only two superfluous terms remaining. Interestingly, these two terms are both constants in the right-hand side of the ODE. A longer time series of the data may have been sufficient to rule out the possibility of constant contributions to the ODEs, because it could have shown the absence of an exponential growth or decay in the phase space.

For the deterministic limit with $q = 10^{-3}$, the estimated coefficients

$$\dot{x} = 10.00y - 10.00x \tag{85}$$
$$\dot{y} = 28.00x - 1.00y - 1.00xz$$
$$\dot{z} = -2.67z + 1.00xy$$
$$q = 9.84 \cdot 10^{-4} \tag{86}$$

yield a basically perfectly accurate and sparse estimation of the right-hand side. Hence, the MLE method has successfully proven to be suitable for estimating stochastic differential equations from data.

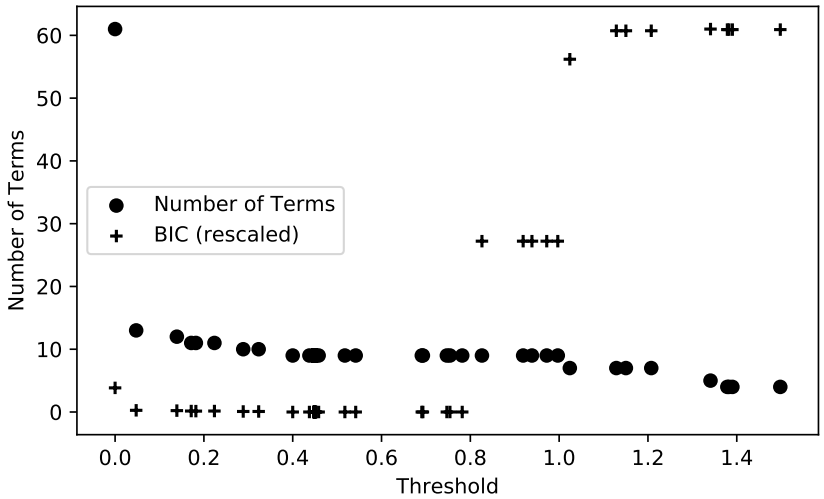


Figure 7: Standard chaotic Lorenz system with high dynamical noise $q = 10$. Dots show the number of nonzero terms estimated via MLE, crosses show the BIC (rescaled to fit into the graph).

4.2.4 Non-Constant Noise: Van der Pol

Instead of a constant noise $q = q_0$, it might make sense to assume a functional dependence of the noise on the position in the phase space. For example, if one regards the trajectory $(x, y, z)^{\mathrm{T}}$ of a particle through the atmosphere, the noise is caused by collisions with other particles. The number of collisions depends on the number of other particles or the particle density which in turn is z-dependent. Hence, $q = q(z)$.

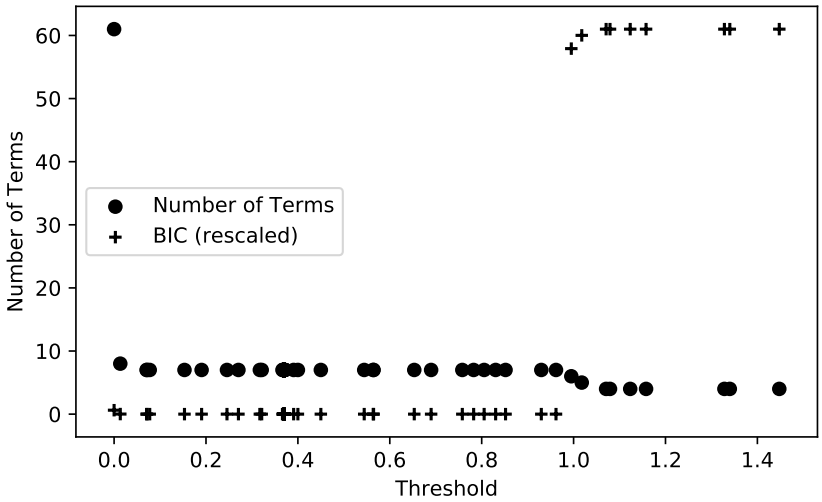


Figure 8: Standard chaotic Lorenz system in deterministic limit with low dynamical noise $q = 10^{-3}$. Dots show the number of nonzero terms estimated via MLE, crosses show the BIC (rescaled to fit into the graph).

Here, a Van der Pol oscillator has been analyzed with a parabolic noise

$$\dot{x} = y \qquad \dot{y} = \mu(1 - x^2)y - x \qquad q = q_0 + q_2 x^2 \tag{87}$$

with q acting as the dynamical noise for both variables x and y. $\mu = 2.0$, $q_0 = 0.05$ and $q_2 = 0.1$ were chosen as the parameters. Since no linear term is included in the noise in (88), the noise estimation also requires a certain sparsity, if the general noise function is given as

$$q = q_0 + q_1|x| + q_2 x^2 = D_2(x, y). \tag{88}$$

For technical reasons, the noise must never be allowed to reach 0 and thus become a deterministic system or, even worse, to reach negative values and ruin the mathematical formulation. This is why the absolute value of the linear contribution to the noise is used in (88). Also, this restriction prevents the use of the threshold for the noise parameters $q_{0,1,2}$ and makes the sparsity become a less objective thing to evaluate. One might choose to regard the estimation as successfully sparse, if the linear q_1 is in a lower order of magnitude than the other noise terms $q_{0,2}$.

Unfortunately, this model showed a strong dependence on the choice of starting parameters for the optimization of the likelihood via the *minimize* method of *scipy.optimize*. Choosing an optimization method without a starting value such as *differential_ evolution* gave rise to further technical problems. It is likely that the three degrees of freedom for the noise allow the occurrence of many deep local minima that make it difficult for the optimization algorithm to find the global minimum. Hence, the starting value for the parameters was chosen as the correct parameter value with some additional parameters by replacing some zeros with values of the order 10^{-1} or 1. With this setup, the TPE hyper-parameter estimation allowed the MLE method not only to sort out the superfluous parameters to promote sparsity, but also gave an accurate estimation of the noise function. The estimated coefficients up to order 10^{-3} are given as

$$\begin{aligned} \dot{x} &= 1.092y \\ \dot{y} &= -0.989x + 2.039y - 1.993x^2y \\ q &= 0.050 + 0.002|x| + 0.104x^2 \end{aligned} \tag{89}$$

and accurately reflect equation (88). The linear term of q_1 is one order of magnitude smaller than q_0 and two orders smaller than q_2. Hence, one can safely state that the MLE method also achieves sparsity for the noise function $q(x)$.

Based on the hypothesis that the large freedom in the noise value inhibits the optimization process, another hyper-parameter estimation was conducted. This time, the ODE coefficients were not given any "good" starting values, but the noise coefficients $q_{0,1,2}$ were given starting values in the correct order of magnitude. This lead to a surprisingly good result shown in figure 9. Rounded off after the order 10^{-3}, this resulted in the exact same coefficients as seen in equation (89). However, as seen in figure 9, a threshold of ≈ 0.97 was chosen as the optimal hyper-parameter and small derivations from this parameter already caused a noticeable difference. Hence, a threshold very close to the lowest relevant coefficient value of 1 in equation (88) was required to achieve a sparse and accurate solution. Unlike e.g. in figure 6, there is no plateau of optimal values, but a sharp minimum. This threshold value might have been overlooked by manual search and underlines the

usefulness of hyper-parameter optimization techniques.

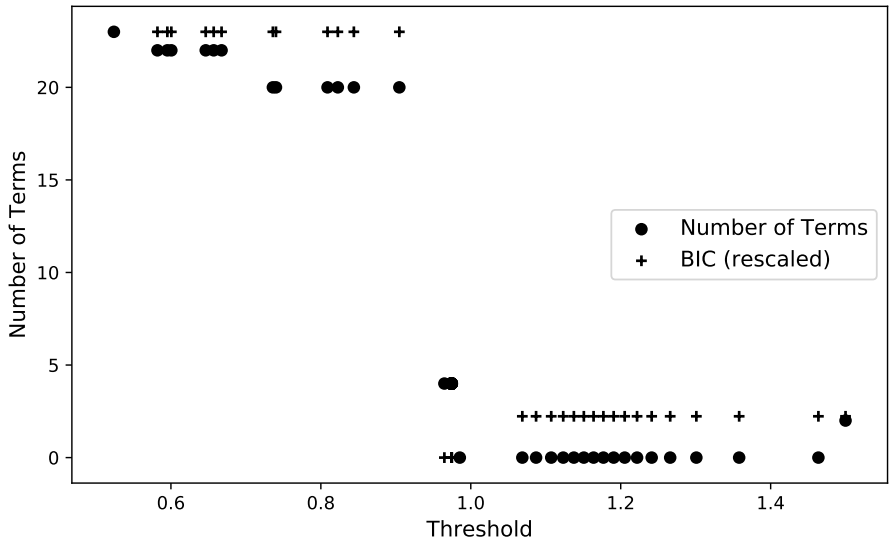
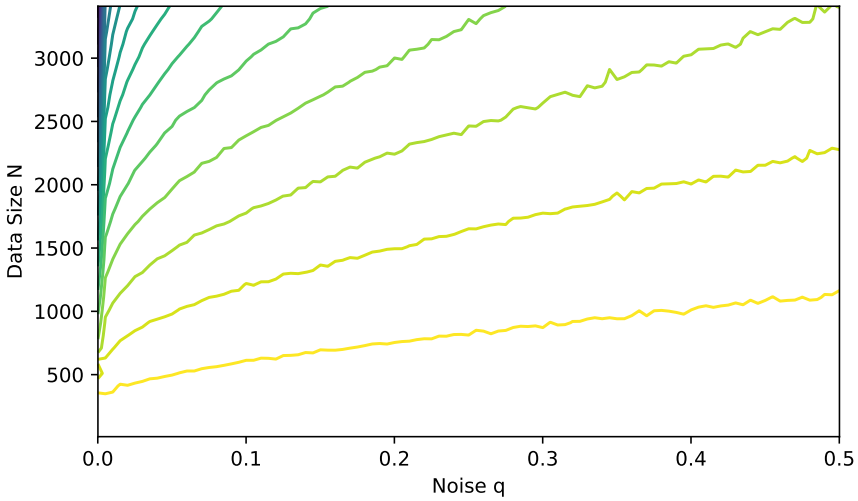


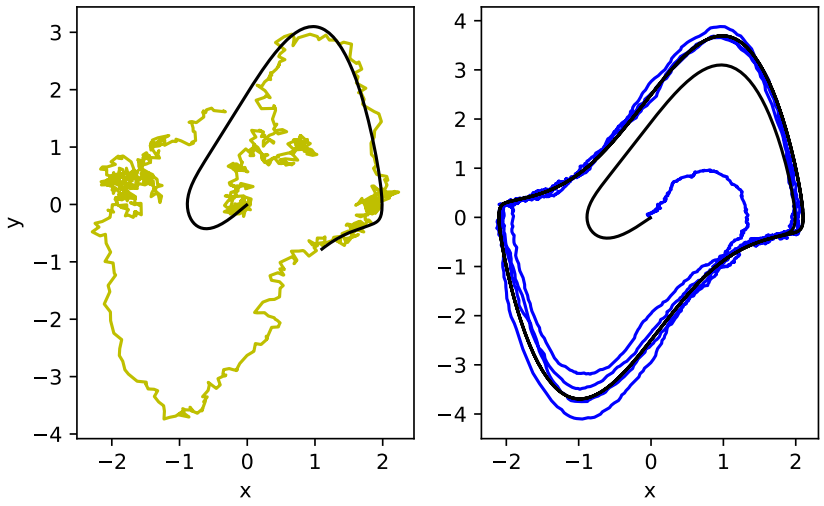
Figure 9: Van der Pol oscillator with $\mu = 2.0$ and nonconsant q as in equation (88). Dots show the number of nonzero terms estimated via MLE, crosses show the BIC (rescaled to fit into the graph).

4.2.5 Propagator Method: Dependence on Data Quality

Especially with regards to the evaluation of real data, the question arises, how much the quality of the propagator interpolation might depend on the quality of the data. To examine this problem, the van der Pol oscillator with constant $D_2 = q$ was evaluated for data samples with different sample sizes N and different noise levels q. Based on prior analysis of the system, the hyperparameter optimization was not conducted repeatedly for each set of data. Instead, the threshold was permanently set to $\lambda = 0.313$ as an optimal value for the ODE. It was investigated, how the BIC depends on N and q, meaning that the BIC was again used as a criterion for the quality of the approximation. The result is depicted in figure 10a indicating that the noise dependence becomes more relevant with a higher sample size, because the slope of the contour lines increases with higher N. To illustrate the different qualities of the approximations, figure 10b shows two trajectories from different areas of figure 10a to depict the clear differences in quality.



(a)



(b)

Figure 10: (a) Contour plot if the BIC for the van der Pol oscillator for different sample sizes N and noise levels q. Darker colours indicate a better approximation. Here, one full oscillation corresponds to approximately 800 data points. (b) Comparison between two van der Pol systems from different regions of figure 10a. On the left-hand side, $q = 0.4$ and $N = 900$, whereas on the right-hand side, $q = 0.01$ and $N = 3000$. The colours of the trajectories approximately correspond to the respective contour line in figure 10a, the data is shown in black.

4.2.6 Conclusion and Comparison with SINDy

MLE with Bayesian hyper-parameter optimization has proven to achieve sparse and accurate reconstruction of ODEs with dynamical noise from data. Knowledge about the structural shape of the ODEs was not necessary to achieve accurate results. The deterministic limit $D_2 \approx 0$ was not an obstacles for this method, but non-constant diffusion terms $D_2 = D_2(x)$ require a more careful treatment of the interpolation algorithm and might require e.g. starting guesses for the likelihood optimization that are in the correct order of magnitude.

Three major differences exist between the MLE method and the SINDy method from [26]: First, SINDy is a differential method, whereas the propagator for the MLE method is based on integrating the trajectory. Second, SINDy performs a multidimensional fit for a linear equation $\gamma = \sum_l \zeta_l$ via a least squares method. This implicitly assumes a symmetric uncertainty with respect to all variables ζ_l. But since ζ_l correspond to the various model functions, e.g. x^l, the same symmetric uncertainty cannot be guaranteed. This symmetry is not implicitly assumed in the general MLE approach in equation 64 in the presented method. Finally, there also exists a methodological difference between SINDy and the MLE method, because the SINDy method only considers deterministic ODEs. Since SINDy fundamentally relies on a linear regression, it should be able to deal with ODEs with a low amount of measurement noise, but dealing with dynamical noise in a diffusion term is quite challenging for SINDy.

4.3 Estimating ODEs via Markov Chain Monte Carlo

With the help of the python package *emcee* in [39], an MCMC algorithm was used to evaluate the likelihood function of the propagator (64) on the basis of numerical data of a van der Pol oscillator with noise level $q = 0.01$ and $\mu = 2$. Posterior samples of all coefficients c_i in a third-order-polynomial ansatz have been computed via MCMC. To see, whether the method is a successful approach to the estimation of ODEs, the trajectory was reconstructed to see if it matches the original data. For every coefficient c_i, the posterior sample was used to calculate the median $c^{(i)}_{\text{med}}$ and the mean $c^{(i)}_{\text{mean}}$. Hence, two trajectories were numerically reconstructed and can be compared

in figure 11. Both the mean-based trajectory in red and the median-based trajectory in blue seem to fit the data, meaning that MCMC is a promising approach to reconstructing ODEs from data.

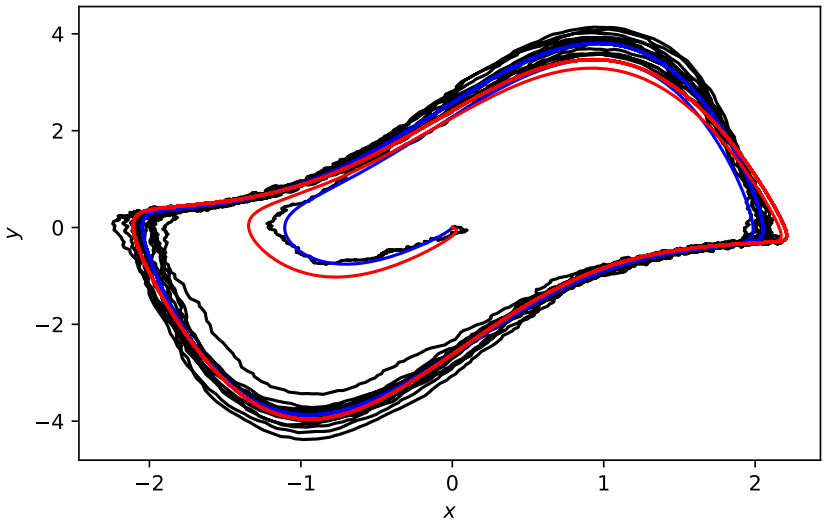


Figure 11: Van der Pol oscillator with $\mu = 2.0$ and $q = 0.01$ reconstructed via MCMC. The data is given in black, the median of the posterior samples in blue, the mean values in red.

4.3.1 Mean vs. Median: Sparsity and Accuracy

Since both the mean-based trajectory and the median-based trajectory manage to reconstruct the data correctly, the question arises whether to choose one over the other. And while the MCMC method manages to reproduce the trajectory, it does not incorporate a procedure to promote sparse solutions. Therefore, a threshold parameter λ was introduced to ignore all coefficients $|c^{(i)}_{\text{mean/med}}| < \lambda$. For the mean-based and the median-based reconstruction, the BIC was computed for various values of λ to evaluate the quality of the trajectory estimation for a λ-threshold. The BIC was chosen to evaluate the goodness of the approximation, because it also includes a sparsity-promoting term $n_{coeff} \log(n_{sample})$. Figure 12 illustrates this for the mean-based trajectory. The median-based approach does not show a qualitatively different shape.

This investigation leads to the following conclusions: First, the median-

based approach has a much lower optimal threshold $\lambda_{\text{med}}^{opt} \approx 0.007$ compared to $\lambda_{\text{mean}}^{opt} \in [0.02, 0.046]$ for the mean-based approach. Therefore, the optimal λ for the mean-based approach manages to ignore more terms than in the median-based approach and the resulting ODE has 9 nonzero coefficients for the mean-based approach compared to 14 for the median-based approach. However, the optimal BIC of the median-based approach is lower than $\text{BIC}(\lambda_{\text{mean}}^{opt})$ for the mean-based approach. Figure 13 compares the trajectories with the optimal sparsity and shows that both still fit the data.

Hence, in this particular example, the mean-based approach yields the more sparse estimation, but the median-based approach shows an overall higher quality of fit. The latter is further supported by regarding the true parameter values: the parameter $\mu = 2$ appears twice in the van der Pol equations (as μ and $-\mu$) and is therefore estimated twice by each method. The mean-based method yields the values 1.55 and -1.60, whereas the median-based method finds a closer numerical estimation with 1.83 and -1.93, respectively. However, for practical purposes and applications, the true value of μ is usually unknown. Therefore, it is very useful that the higher accuracy of the median-based method was also confirmed by the BIC without having to rely on any information concerning the true parameters.

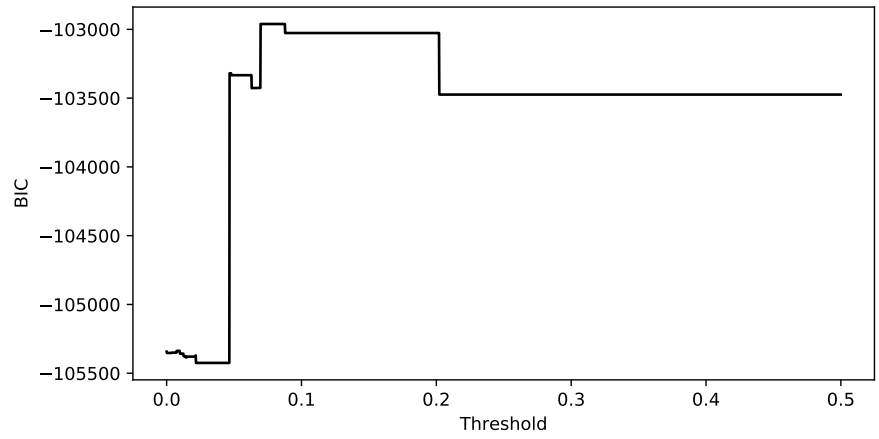


Figure 12: BIC as the quality of the trajectory reconstruction using only the coefficients with $|c_{\text{mean/med}}^{(i)}| < \lambda$ for a threshold parameter λ.

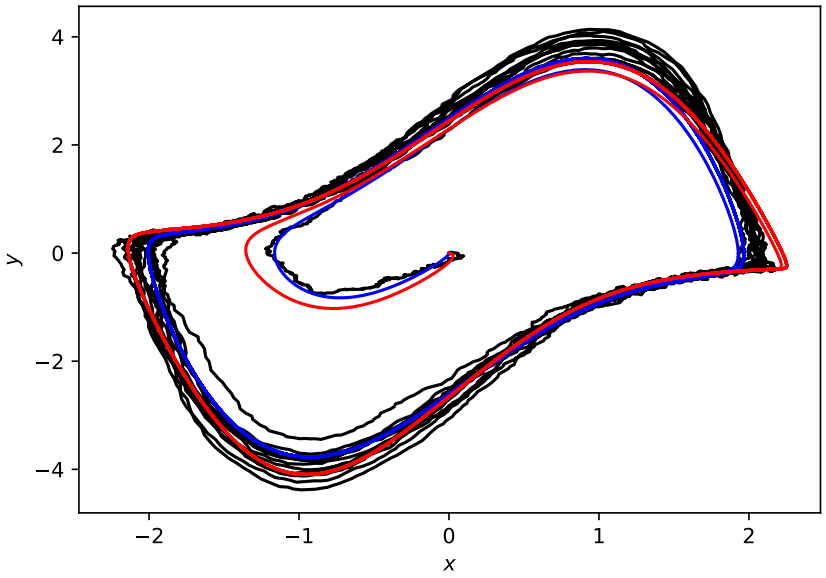


Figure 13: Van der Pol oscillator with $\mu = 2.0$ and $q = 0.01$ reconstructed via MCMC. The data is given in black. The blue/red trajectory shows the optimized sparse trajectory of the median/mean-based approach yielding the optimal BIC.

4.3.2 Posterior Distributions

Since we know the shape of the true van der Pol equations to be

$$\dot{x} = ay \qquad \dot{y} = bx + cy + dx^2 y$$
$$\text{noise} = q$$

with noise q, we can observe the posterior distributions of the relevant parameters a, b, c, d and q. Note that the true values for this particular set of data are $a = -b = 1$, $c = -d = 2$, $q = 10^{-2}$. The pairwise scatter plots for these parameters alongside the kernel density estimation for each parameter is depicted in figure 14.

Several interesting observations can be made: First, the kernel density estimations of all parameters with the exception of the noise q show two local maxima. This explains the differences between the mean-based and the median-based approaches from section 4.3.1. Second, the scatter plots show some well-defined tails that are especially noticeable for the parameters b, c

and d. The end of the tails correspond to the true parameter values, meaning that the tails represent the trails chosen by most of the MCMC-walkers. Finally, the scatter plots of c and d do not show a qualitative difference compared to the other ones. Because they both correspond to the same model parameter $c = -d = \mu$, one could have assumed to see a strong correlation in the scatter plot. Unfortunately, this does not seem to be the case.

The most striking observation is the presence of a secondary local maximum in the kernel densities. Ad hoc, it is not completely evident why these maxima exist. However, one could assume that they are caused by a too short burn-in period: The MCMC-samplers usually need a few iterations to initialize successfully, meaning that the first sampling results should be discarded.

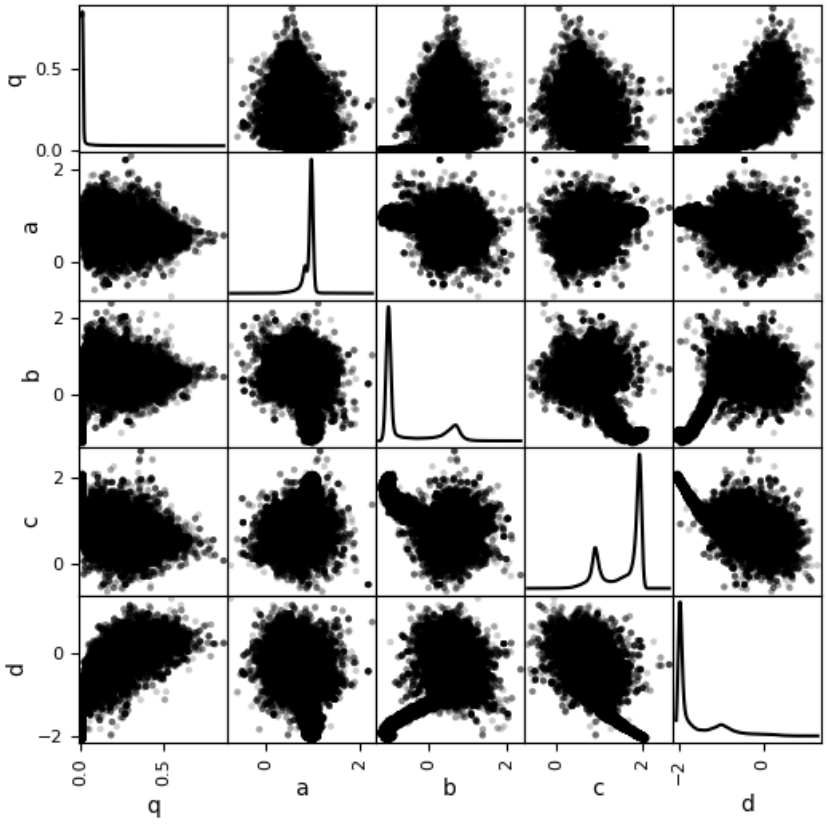


Figure 14: Scatter plot of the parameters given in equation (**??**) with the kernel density estimation on the diagonal axis.

4.3.3 Higher Burn-In

Based on the previous observations, the analysis was repeated with a higher burn-in number (300 instead of 10). This lead to noticeably different results: First, the coefficients yielded by regarding the mean and the median now only show minor differences. Second, the posterior distributions now have a different shape (cf. figure 15). Here, these two-dimensional distributions are concentrated in a much smaller region than in figure 14 and have a much more well-defined principal axis along which most of the distribution's variance is found. This is especially noticeable in the c/d scatter plot that basically depicts a straight line. This is a strong hint at a correlation between c and d, which is exactly what one would hope to receive as a result, because $c = -d = \mu$. Also, most of the secondary maxima in the kernel density estimations have vanished in figure 15. Therefore, it is safe to conclude that the higher burn-in improves the MCMC results.

The analysis of the estimated trajectory can be repeated as in section 4.3.1 to compare the quality and the sparsity of the mean-based and median-based estimation. Here, the minimum BIC is almost identical for both methods (approximately -113600 for the median and -113500 for the mean), but the methods show a much different result in terms of sparsity. The minimum BIC mean-based estimation still includes 15 nonzero coefficients (and the noise q), whereas the minimum BIC median-based estimation only includes exactly the four true coefficients (and the noise q). The best median-based estimation yields the ODE

$$\dot{x} = 0.96y \qquad \dot{y} = -1.02x + 1.96y - 1.98x^2y \qquad q = 0.01 \tag{90}$$

with coefficients rounded to two decimal places. This is both a perfectly sparse estimation of the ODE and a very accurate estimation of the true coefficients

$$\dot{x} = y \qquad \dot{y} = -x + 2y - 2x^2y \qquad q = 0.01 \tag{91}$$

and therefore proves the success of the MCMC approach to estimate sparse stochastic ODEs. The resulting trajectories shown in figure 16 depict the accurate representation of the data.

4.3.4 MCMC: Conclusion

Markov Chain Monte Carlo methods show promising results and successfully help to estimate sparse ODEs via the propagator-based likelihood. A sufficiently high burn-in number is necessary to receive informative posterior distributions which can help to infer correlations between the parameters of the ODE. Both the median-based and the mean-based approaches manage to accurately reproduce the trajectory of the data. However, especially when comparing the estimation to the real coefficient, the median-based approach yields the superior results, possibly because it is more robust to outliers of the MCMC samples.

Introducing a threshold parameter λ as in the SINDy method in [26] allows the MCMC method to perform sparse estimations. This is an important addition to the pure MCMC method which only yields a posterior distribution for the model parameters and does not directly yield an algebraic ODE as the "one and only" solution. Instead, it basically results in a continuum of solutions encoded in the sampled posterior densities. Note that a hyper-parameter optimization could have been performed to optimize λ, but because of the low amount of computation time to produce the data of figure 12, an algorithm like the TPE was not necessary.

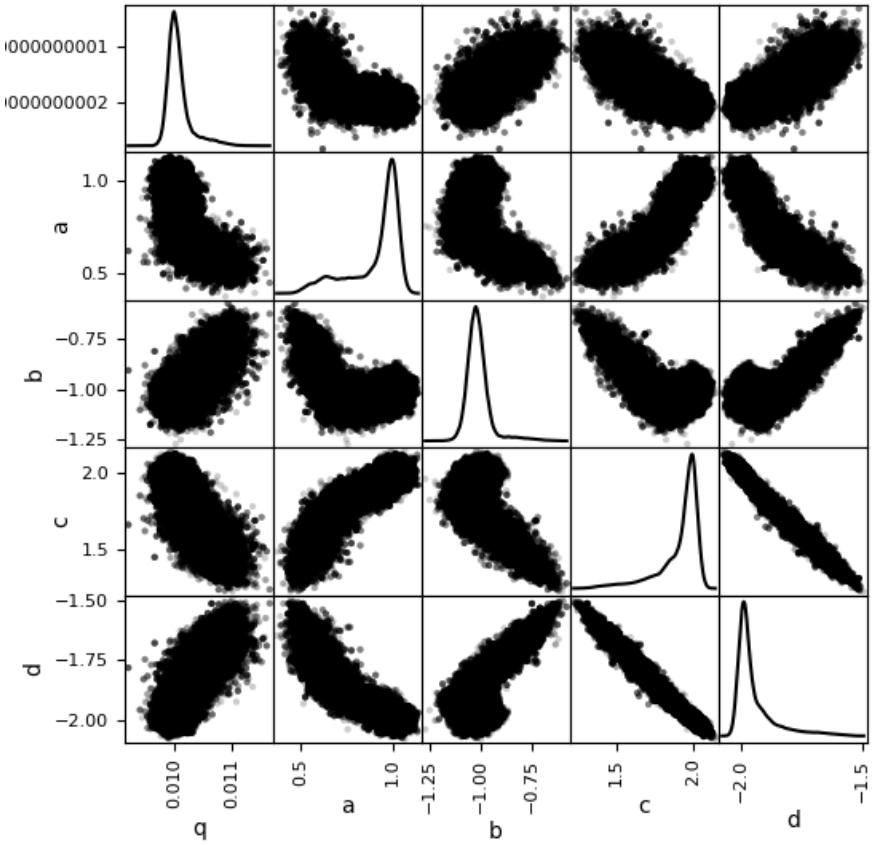


Figure 15: The same scatter plot as in figure 14 but with a higher burn-in number.

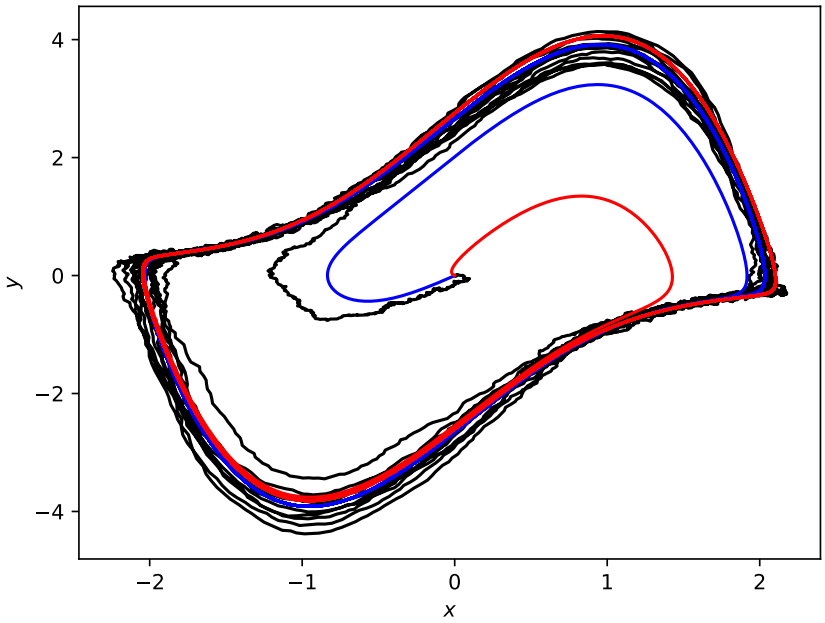


Figure 16: The same scatter plot as in figure 13 with the optimal mean-based estimation in red and the optimal median-based estimation in blue, but with a higher burn-in.

5 Conclusion and Outlook

6 References

[1] *Data Analysis - A Bayesian Tutorial*, D. S. Sivia and J. Skilling, second edition, Oxford University Press, 2006.

[2] *Approaching complexity by stochastic methods: From bio-logical systems to turbulence*, R. Friedrich et al. in Physics Reports 506.5 (2011), pp. 87–162.

[3] *Maximum likelihood estimation of drift and diffusion functions*, David Kleinhans and Rudolf Friedrich in Physics Letters A368.3-4 (Aug. 2007), pp. 194–198.

[4] *Bayesian differential pro-gramming for robust systems identification under uncertainty*, Yibo Yang, Mohamed Aziz Bhouri, and Paris Perdikaris in arXiv:2004.06843 (2020).

[5] *Statistical inference for a multi-variate diffusion model of an ecological time series*, Melvin M. Varughese and Etienne A. D Pienaar in Ecosphere 4.8, 2013.

[6] *Bayesian Probability Theory - Applications in the Physical Sciences*, Wolfgang von der Linden, Volker Dose and Udo von Toussaint, Cambridge University Press, 2014.

[7] *Mathematische Statistik* (lecture notes), Matthias Löwe, Münster, 2018.

[8] *Wahrscheinlichkeitstheorie - einschließlich Grundlagen der Maß- und Integrationstheorie - Skripten zur Mathematischen Statistik Nr. 40*, Gerold Alsmeyer, Münster, 2016.

[9] *The Lady Tasting Tea*, D. Salsburg, 2002.

[10] *The 1986 CODATA Recommended Values of the Fundamental Physical Constants*, E. Richard Cohen and Barry N. Taylor in Journal of Research of the National Bureau of Standards, 92(2), March-April 1987.

[11] *Mathematical Foundations of Information Theory*, Aleksandr Iakovlevich Khinchin New York, 1957.

[12] *An introduction to information theory and entropy* (lecture notes), Tom Carter, Santa Fe, 2013; cf. `http://astarte.csustan.edu/~tom/SFI-CSSS/` (status as of October 2019).

[13] *Probability and Symmetry*, Paul Bartha and Richard Johns, University of British Columbia, 2001.

[14] *Machine Learning*, Tom M. Mitchell, 1997.

[15] *Maximum entropy sampling and optimal Bayesian experimental design*, Paola Sebastiani and Henry P. Wynn, The Open University and University of Warwick, 1999.

[16] *Bayesian Adaptive Exploration*, Thomas J. Loredo, Cornell University, 2004.

[17] *Algorithms for Hyper-Parameter Optimization*, James Bergstra et al., 2011.

[18] *Bayesian Optimization Primer*, Ian Dewancker, Michael McCourt and Scott Clark, SigOpt, 2015.

[19] *A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning*, Will Koehrsen, `https://towardsdatascience.com/https://tex.stackexchange.com/questions/129051/two-subfigures-in-two-rowsa-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f` (status as of October 2019).

[20] *On Estimation of a Probability Density Function and Mode*, Emanuel Parzen, Stanford University, 1962.

[21] *Generalized Accept–Reject sampling schemes*, G. Casella, C.P. Robert and M.T.Wells, in *A Festschrift for Herman Rubin*, 2004.

[22] *Regression Shrinkage and Selection via the Lasso*, Robert Tibshirani, 1996.

[23] *Gaussian Processes for Machine Learning*,Carl Edward Rasmussen and Christopher K. I. Williams, Massachusetts Institute of Technology, 2006. Available at `http://www.gaussianprocess.org/gpml/chapters/` (status as of November 2019).

[24] *Bayesian optimization*, Martin Krasser, 2018, `http://krasserm.github.io/2018/03/21/bayesian-optimization/` (status as of November 2019).

[25] *Stability selection enables robust learning of partial differential equations from limited noisy data*, S. Maddu et al., 2019.

[26] *Discovering governing equations from data: Sparse identification of nonlinear dynamical systems*, Steven L. Brunton et al., 2015.

[27] *Nonlinear Dynamics And Chaos*, Steven Strogatz, Westview Press, 2001.

[28] *Practical Bayesian Inference - A Primer for Physical Scientists*, Coryn Bailer-Jones, Cambridge University Press, 2017.

[29] *Bayes Factors*, Robert Kass and Adrian Raftery, Journal of the American Statistical Association, 1995.

[30] *Choosing Models for Cross-Classifications*, Adrian Raftery, American Sociological Review, 1986.

[31] *Philosophy and the practice of Bayesian statistics*, Andrew Gelman and Cosma Rohilla Shalizi, arXiv 1006.3868, 2011.

[32] *A Comparative Study of Black-box Optimization Algorithms for Tuning of Hyper-parameters in Deep Neural Networks*, Olof Skogby Steinholtz, Luleå, 2018.

[33] *Hyperparameter optimization for Neural Networks*, Yurii Shevchuk, 2016. Available at `http://neupy.com/2016/12/17/hyperparameter_optimization_for_neural_networks.html` (status as of January 2020).

[34] *Skript zur Vorlesung Mathematische Statistik* (lecture notes), Zakhar Kabluchko.

[35] *Information Criteria and Statistical Modeling*, Sadanori Konishi and Genshiro Kitagawa, Springer, 2008.

[36] *The Fokker-Planck Equation*, Hannes Risken and Till Frank, Springer, 1996.

[37] *Einführung in die Wahrscheinlichkeitstheorie als Theorie der Typizität*, Detlev Dürr, Anne Froemel and Martin Kolb, Springer, 2017.

[38] *Stochastik - Struktur im Zufall*, Matthias Löwe and Holger Knöpfel, Oldenbourg Wissenschaftsverlag, 2011.

[39] `https://emcee.readthedocs.io/en/stable/user/install/`, status as of May 2020.

[40] *Time Series Analysis*, J.D. Hamilton, Princeton University Press, 1994.

[41] *Nonlinear Time Series Analysis*, H. Kantz and T. Schreiber, Cambridge University Press, 2003.

[42] *Time Series Analysis and Its Applications*, R.H. Stoffer and S. David, Springer, 2006.

A Further Mathematics

Further mathematical theorems are discussed in this section.

A.1 Information Theory

Information theory is connected to probability theory and its applications, because it can e.g. allow a probabilistic estimation of the expected improvement of a general property. This has major applications in the Bayesian experimental design, but is probably enough to fill a thesis on its own. Instead, the basics of information theory will be discussed here in order to derive two of the most prominent a priori distributions used in Bayesian statistics: the uniform and the Gaussian distribution. Their ubiquity in (Bayesian) statistics is usually taken for granted, but the concept of entropy in the context of information theory justifies their use.

A.1.1 Deriving the Shannon Entropy

The first well-known axiomatic approach at deriving the Shannon entropy is given in pp. 9-13 of [11]. Unfortunately, there seems to be a disagreement within the literature with respect to the question, what entropy actually is. Within this thesis, we will outline the approach to treating entropy as a measure of impurity. Then, information will be regarded as negative entropy like in [14]. Although [12] chooses to regard entropy as equivalent to information, we will adapt the intuitive derivation of the entropy in [12] for this thesis.
For events with probabilities $p \in [0, 1]$, three axioms are assumed for the information $I(p)$ obtained by observing such an event:

1. A certain event will not provide us with relevant information: $I(1) = 0$
2. Information due to independent events (i.e. their probabilities factorize) is additive: $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$
3. $I(p)$ should be continuous.

From these axioms, one can conclude that $I(p) = a \log_b(p)$ for some $b > 1$ and $|a| = 1$. The choice of b is irrelevant and only reflects the used unit of information, as e.g. $-\log_2(p)$ corresponds to measurements in bits. But the question of $a = 1$ or $a = -1$ seems to be not unanimously answered by

the literature. In these sections, we will give proof to support the idea of $a = 1$. For now, we will suppose that $I(p) = \log_b(p)$ and we will calculate some examples to show that this approach is justified.

For a discrete ensemble of events $a_1, a_2, ...$ with probabilities $P = (p_i)_i$, or for a continuous probability density $P(x)$ the Shannon entropy H is defined as the expectation value of the negative information as

$$H_d(P) = -\sum_i p_i \log_b(p_i) \tag{92}$$
$$H_c(P) = -\int P(x) \log_b(P(x))\, \mathrm{d}x$$

with any $b > 1$. As the negative information, it is intuitive to regard H as a measure of impurity, similar to entropy in classical thermodynamics.

A.1.2 Maximum Entropy for Discrete Distributions

Usually, it is desirable to observe an ensemble in such a state that the information is maximized, since science revolves around maximizing the information obtained via observations of ensembles. For simplicity and to take into account the discrete nature of computationally solved problems, the discrete $H_d(P)$ is often regarded. Maximizing the information is according to equation (92) equivalent to $H(P) \stackrel{!}{=} \min$ under the condition $\sum_i p_i = 1$ and $p_i \geq 0 \forall i$. This is achieved by having one i with $p_i = 1$ as a certain event and all other $p_{j \neq i} = 0$, since in this ensemble, $H_d = 0$, whereas any deviation from this choice of P results in $H > 0$. The minimum entropy and therefore maximum information is obtained by transforming the probabilistic into a deterministic problem, as with this setup, event i will occur with certain possibility. This makes sense, since in such a setup, we know exactly how the system will behave, because it is now a deterministic problem: Our knowledge about the system is maximized.

On the other hand, it is also instructive to regard the maximized entropy. Maximizing the entropy $H(P) = \max$ under the condition $\sum_i p_i = 1$ can be solved via Lagrange multipliers. Then, the maximum entropy leads to a uniform distribution $p_i = p \; \forall i$, if no other restriction is applied to the system: The uniform distribution provides the highest level of irregularity

and therefore the lowest amount of information, because there is no "structure" (e.g. a peak) encoded in the probability density. This makes sense and had even been pointed out long ago by Laplace in his Principle of Indifference, according to which it makes sense to assign a uniform distribution to a problem, if there is no additional information given.

Interestingly, this provides us also with an analogy to the most "natural" distribution: Intuitively, it is highly unnatural to see a sharp concentration of mass (or probability) in a very small area, unless there is a strong constraint causing this. Like the entropy of classical thermodynamics, the maximized Shannon entropy demands an even spread of the probability and therefore disencourages unnaturally concentrated distributions. Hence, it might be reasonable not to regard entropy as a measure of impurity, but as a measure of naturalness. This requires the interpretation of naturalness being the opposite of information, which makes sense if information is considered to be *purposely* encoded in a system.

A.1.3 Maximum Entropy for Continuous Distributions and under Constraints

In the general case, the entropy might be computed with respect to continuous distributions and under more restrictive constraints than the normalization $\int \mathrm{d}x p(x) = 1$ or $\sum_i p_i = 1$. Since the discrete case can be derived from the continuous one via a suitable $p(x) = \sum_i \delta(x_i) p_i$, we will focus on the general, i.e. continuous case in this paragraph. A constraint might be a macroscopic observation F_i, which the observer knows to be linked to microscopic and probabilistic features via a function $f_i(x)$, resulting in

$$F_i = \int \mathrm{d}x f_i(x) p(x) \tag{93}$$

as the i^{th} constraint. Such a problem is solved with a Lagrangian method. The Lagrangian $\mathcal{L}$ is then given as

$$\begin{aligned}\mathcal{L}(p(x), \lambda_0, \lambda_{1,\dots,n}) = &- \int_X p(x) \log(p(x)) \, \mathrm{d}x \\ &+ \lambda_0 \left(\int_X p(x) \, \mathrm{d}x - 1 \right) + \sum_{i=1}^{n} \lambda_i \left(\int_X f_i(x) p(x) \, \mathrm{d}x - F_i \right)\end{aligned} \tag{94}$$

with the 0^{th} condition being the normalization, X the support of the density $p(x)$ and the Lagrangian multipliers λ_i.

A.1.4 Examples: Deriving Laplace's Principle of Indifference and the Gaussian Normal Distribution

The continuous case of maximum entropy provides us with two very interesting examples to derive "natural" distributions, which will be discussed in this section.

Example 1: Deriving Laplace's Principle of Indifference
If we regard a continuous distribution without any constraints but the normalization, equation (94) changes to

$$\mathcal{L}(p(x), \lambda_0, \lambda_{1,\dots,n}) = - \int_X p(x) \log(p(x)) \, \mathrm{d}x + \lambda_0 \left(\int_X p(x) \, \mathrm{d}x - 1 \right). \tag{95}$$

Maximizing this leads to

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{p(x)} = - \log p(x) - 1 + \lambda_0 \tag{96}$$

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{\lambda_0} = \int_X p(x) \, \mathrm{d}x - 1 \tag{97}$$

with the integral disappearing in (96) due to the functional derivative. (96) implies $p(x) = \exp(-1+\lambda_0)\mathbb{1}_X(x)$, meaning that $p(x)$ is uniform. Under the normalization constraint (97) and assuming for simplicity that $X = [a, b]$ is an interval, this results in

$$p(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x) \tag{98}$$

with a suitable λ_0. This is an analogy to the discrete uniform distribution

derived in section A.1.2 and is a formal derivation of Laplace's Principle of Indifference: without further knowledge, it is reasonable top assume equal probabilities for all possible outcomes. This principle is highly relevant for Bayesian approaches, since this justifies the choice of a uniform prior over any arbitrarily complicated prior, if no further knowledge about the system is present. For a further discussion about how the Principle of Indifference allows scientist to deal with seemingly ill-defined probabilities, cf. [13].

Example 2: Gaussian Errors

In experimental measurements, the errors x are usually regarded as random variables with a mean $\mathbb{E}[x] = 0$, since a nonzero mean would indicate a systematic error, and with a known variance $\mathbb{V}[x] = \sigma^2$ derived from the knowledge about e.g. the accuracy of the measuring instruments or the resolution of the measurement scale. Regarding these two constraints, the principle of maximum entropy can now derive the most reasonable probability distribution for the error via maximizing the Lagrangian

$$\begin{aligned}\mathcal{L}(p(x), \lambda_{0,1,2}) = &- \int_X p(x) \log(p(x)) \,\mathrm{d}x + \lambda_0 \left(\int_X p(x) \,\mathrm{d}x - 1 \right) \\ &+ \lambda_1 \left(\int_X x p(x) \,\mathrm{d}x - 0 \right) + \lambda_2 \left(\int_X (x-0)^2 p(x) \,\mathrm{d}x - \sigma^2 \right).\end{aligned} \tag{99}$$

From the first Lagrangian condition $\partial\mathcal{L}/\partial p(x) = 0$, it follows that

$$0 = -\log p(x) - 1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2. \tag{100}$$

This is solved by $p(x) = \exp((1-\lambda_0) - \lambda_1 x - \lambda_2 x^2)$. Applying the other three Lagrangian condition $\partial\mathcal{L}/\partial\lambda_i = 0$, the shape of $p(x)$ is determined:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} \right) \tag{101}$$

According to the principle of maximum entropy, the Gaussian normal distribution is the most natural one for this set of constraints. This is a formal justification of the standardized approach in many quantitative sciences to treat their experimental errors as Gaussian variables, which is usually omitted in literature on interpreting experimental results.

A.1.5 Maximum Entropy: an Intermediate Conclusion and Discussion

These two examples illustrate that the principle of maximum entropy is by no means an abstractly mathematical concept, but very well-suited to describe real life phenomena. It is deeply rooted in basic ideas of statistical sciences, as Laplace's Principle of Indifference and the typically Gaussian errors can both be derived from it. The interpretation of entropy as a measure of naturalness is therefore justified and, in accordance to the arguments in section A.1.2, allows us to measure the information as $-H$. Two possibly controversial aspects about these definitions of entropy and information will be discussed in this section.

Information is not Symmetric

A potentially puzzling or contradictory aspect of the Shannon entropy, which the maximum entropy principle is based on, may be encountered in the following line of thought: Regarding an event A and its complement A^C, one might be inclined to assign them the same level of information. Since via symmetry, the occurrence of A and its non-occurrence should provide us with the same level of information. Yet, this obviously only works with respect to the individual information function $\log(p(A))$, if $p(A) = 1/2$, which will usually not be the case. The error in this reasoning is as follows: The information for an event A described by the axioms A.1.1 is not the information that A occurs, but the information encoded within the system as a whole.

Information as Negative or Positive Entropy?

As already mentioned before, the Literature is split on the definition of information I as $I = -H$ or $I = H$. Here, we have interpreted entropy as a measure of impurity and naturalness and outlined that it makes sense to regard a natural system as the opposite of a system with (purposely) encoded information. The most striking example to support this interpretation is that a deterministic system, as described in section A.1.2, minimizes the entropy, whereas it allows the observer to perfectly predict the system behaviour and therefore provides a maximum amount of information.

However, this does not mean that the approach of $I = H$ (cf. [12]) should be dismissed immediately. In this framework, the information is

maximized by a uniform distribution, whereas the deterministic distribution provides the observer with a minimum of information. This perspective makes sense, if you consider that the deterministic system with $p_i = 1$ will only provide the observer with knowledge about a single event i, whereas the system with uniform distribution will, within a reasonably low amount of experiments, show the occurrence of *all* $i = 1, ..., n$ and the consequences of their occurrence. And how could observing only one event again and again be more informative than observing many events and their consequences? This is an important distinction that has to be kept in mind while dealing with Shannon entropy.