Model Selection to detect growth paths : An inspection using synthetic data

Alassane-Anand Ndour

Abstract

This study aims to understand and evaluate the use of model selection strategies in the detection of growth models in different signal-to-noise levels. To study this, we run in a consistent manner, a series of experiments at several noise levels on simulated data. Each experiment involved classifying a set of datasets according to their underlying functional form with different model selection methods. The main selection strategies considered were divided into Bayesian and Frequentist methods which each included of traditional measures (e.g. AIC, BIC) as well as modern ones (e.g. WAIC). We find that WAIC can be used confidently for low to medium levels of noise (with an additive Gaussian error with mean 0 and up to a variance of 0.5) and that enhancing AIC and BIC with entropy as proposed in a recent study, offers a more robust selection process.

Contents

1 Introduction and Objectives

1.1 Background to the problem, project choice and beneficiaries

Often times, researchers aim to detect growth trends within datasets in order to understand if the subject of the study was altered by an object. For example, a researcher in ecological science might aim to understand and quantify the effect of communities on populations (Strauss 1991). There have been different ways of tackling this problem. One such example is the use of power analysis to detect trends proposed by Gerrodette (1987). Here we propose to contribute to this literature by using model selections to detect growth processes. This is particularly important in certain fields that shy away from using model selection methods: in ecological publication, a field that commonly deals with population growth, between 1993 and 2003, only 2% of publications mentionned them (Aho, Derryberry, and Peterson 2014). Growth model selection is a large literature in many domains which often takes into account the specificity of the subject matter. For instance, in biostatistics, bioassay data can be modelled using different logistic like functions - e.g. 4 Parameter Logistic or 5 Parameter Logistic as shown by Gottschalk and Dunn (2005) - that are tested against each other.

Generally, model selection is a basic scientific requirement that answers what functional form a set of data corresponds to. There are different requirements for a function to be chosen in a selection problem such as the simplicity of the model (Occam's razor states that the simpler model should prevail), the estimation of parameters or even the "certainty" of the selection. For instance, in economics different models can be tested against each other to verify how well they explain GDP growth as demonstrated by Sala-i-Martin, Doppelhofer, and Miller (2004). In such a case, the main requirement of the models is interpretability as it is required for policy making. In contrast, for prediction interpretability of neural nets for example is not always necessary. Naturally, depending on the use case, it is crucial to select the best functional form as the real one is often unknown, and a wrong selection invalidates any following inference (Nguimkeu 2014).

Most domains use statistical foundation to select models. It is often done by comparing the relative likelihood that the data's underlying generating process was given by a specified function (Claeskens and Hjort 2008). It is a well-known and documented

problem. However, there is a specific context that is not tackled often by the current literature: how would one classify a large pool of different datasets using model selection processes to determine their appropriate underlying functional form? This classification problem is the subject of matter. Therefore, by comparing several relevant model selection strategies, this project aims to obtain a growth model classification method that by construction classifies datasets accurately but also demonstrates the uncertainty level by which it is doing so and provides the estimated parameters.

The project was chosen for its application in a wide variety of domains. The analysis therefore contributes to the literature of model selection by offering an experimental evaluation of several selection processes. Researchers studying growth models (in ecological science for instance) who wish to have empirical evidence on detecting growth paths based on model selection can benefit from this work. This echos Evans (2019) who undertakes a similar task in psychonomic studies by evaluating evidence accumulation models and experinmenting on them with different model selection strategies. Additionally, the present study extends on the work of Murari et al. (2019) by taking the authors' suggestion and comparing their novel selection process to other Bayesian methods. Therefore, practitionners who wish to have an insight in the empirical performance of novel selection strategies can also refer to this study. Finally, by careful experimentation of noise-to-signal, this study contributes empirically to the exploration of noise-robust selection strategies.

1.2 Objectives, metrics and outline

The aim of the project is to construct a dataset classification method that can accurately identify the functional form the dataset is following. The challenge of this task arises when the dataset to be classified has a weakening signal-to-noise ratio. Consequently, it is essential for the classification to be able to detect any underlying patterns in a noisy setting. Thus, a large part of the study will focus on how well the classification performs as the noise in the data increases. We use F1 score and accuracy to measure the success og a model selection classifier. A classifier with a high score and accuracy (above 95%) can be considered a suceesful in identifying the underlying functional form. Ideally, the classifier could also quantify the noise level for a given dataset. Furthermore, our model should offer an estimation that pinpoints its classification "certainty" and that of the model parameters. Therefore, when possible, correct detection of noise level

and identification of parameters within a confidence interval are measured. Here it is difficult to pinpoint what is considered a successful classifier since parameter estimations can greatly vary depending on estimation methods. Nonetheless, this study aims to measure and quantify parameter classification. Finally, as the selection process must scale to a large number of datasets, we record the computation cost of each method and will favour a less expensive but accurate classifier.

Section 4 provides more details on the methodology used. Here we offer a brief overview of the ways the classification task was handled. There are broadly speaking two ways of thinking of how one might choose a model: a frequentist one which aims to compute a statistic on the data for which we know the distribution VanderPlas (2014) and a Bayesian approach which compares the posterior of different models using methods such as Bayes factor (BF). Here we apply both views on a generated dataset and compare them using the based on performance accuracy, parameter estimations and speed of computation. The rest of the discussion is organised in the following manner : section 2 aims to review the relevant literature of the study particularly in the fields of data simulation and model selection; section 3 describes the data used; section 4 provides a deep dive into the methodology used in the study; sections 5 and 6 provide the findings along with a discussion around them and section 7 offers closing remarks.

2 Context and relevant literature

The analysis carried out in this project has two fundamental building blocks which are data simulation and model selection strategies. For this reason relevant literature on both these topics will be presented in this section.

2.1 Data simulation background

Data simulation simply refers to the process of generating random numbers from a distributional statement. It a common statistical technique to understand or forecast a phenomenon that might occur in observational data, all so in a controlled context. Kéry and Royle (2016) provide an insightful outline of the advantages of simulations. Here we highlight some of them as they were the main reasons for simulations used in this study. According to them data simulations allow researchers to know the truth behind the data, in a parametric context for instance. This is particularly important as the

researcher has less insight on the internals of certain black box algorithms. The authors point out that having knowledge of parametric values before applying a Markov-Chain Monte Carlo (MCMC) simulation as is done in the present study, is necessary because it can provide evidence that the simulation is not going astray. Additionally, it can be noted that MCMC is by construction a simulation and represents a corner stone of Bayesian statistical techniques (Brooks et al. 2011) emphasizing the importance of simulations. Kéry and Royle (2016) also highlight that simulations help calibrate model parameters. More generally, it can be said that simulations have also been used in order to study, compare and contrast particular models in many fields. For instance, in Van Der Ploeg, Austin, and Steyerberg (2014) the authors study the effect of small sample size on different modelling techniques in patient survival rates; the data they use was simulated based on observed data. In Murari et al. (2019) the authors compare different information criterion in model selection problems using simulations. These examples not only show that simulations are present in many domains but also serve as a solid control to study intra and inter model specificities. Finally, Kéry and Royle (2016) also point out that simulations allow researchers to include errors that can then be accounted for and studied as one should do with observed phenomenon. This principle is important in our context and can be illustrated by Harris et al. (2016) who use Bayesian techniques to account for sampling errors in a simulated dataset.

Although, simulation is undeniably an important and effective tool in the researchers arsenal, it must be noted that there exist different forms of data simulations. Since this research project originally stems from the biology literature (Harris et al. 2016), examples in biology and biostatistics will be emphasized here. Broadly speaking, simulation methods can be divided between Simulation Optimization methods and algebraic methods as described by Amaran et al. (2016). The former refers to techniques used to optimize stochastic simulations that cannot be described algebraically. These are more black-box type simulations and can be found in projects such as Montagna and Omicini (2017) which proposes a framework to optimise parameters in biological system development simulations. It can be stated that these models are required in specific settings but can lack generalisation as pointed out by Fu et al. (2000). In contrast algebraic solutions are more general but they cannot explicitly tackle more precise stochastic tasks. For instance, we know that population growths in ecological studies follow a logistic growth process (Snider and Brimlow (2013)). This is a general result and observed data can be seen to fit this pattern as demonstrated when Buehler

et al. (1991) adapt logistic growth functions to take into account the human activity impact on the bald eagle population in the US. However, to model stochastic differential equations as is necessary in the separation of DNA molecules - Cho and Dorfman (2010) - stochastic simulations are necessary. In this study, since generalisation was necessary the algebraic route described in section 3 was taken.

2.2 Model selection background

The second building block of this research is statistical model selection which refers to the step of selecting the most appropriate statistical model among a set of candidates for a given dataset as described by Ding, Tarokh, and Yang (2018)). This can be identifying the number of regressors required in a linear regression or selecting the type of neural network necessary for a given task. This step is performed in an array of domains: in ecological science where researchers use mark-recapture (marking an animal and recapturing it in a later period) in order to estimate the population of a species and its probability of survival it is common to have multiple statistical models compete and use the best one for inference (Johnson and Omland 2004); in cosmology, researchers estimate models, which compete to describe phenomenon such as the geometry of the expansion of the universe (Liddle 2007). Therefore, the overarching importance of model selection is not contested; as pointed out by Nguimkeu (2014) a wrong statistical selection invalidates any following inference. In addition to selection, as described by Claeskens and Hjort (2008), model averaging is a closely linked problem as researchers might wish to combine relevant competing models. Even though the importance of model selection is not contested it remains an open problem in statistics that often requires the combination of multiple methods. Furthermore, within the literature there are different ways of approaching the selection problem. Dormann et al. (2018) outlines the schools of thought that are prevalent in model selection and averaging issues. This discussion is key to understand and compare the different paths available.

On one hand, empiricists base model selection on the data and make fewer assumptions (Fernandez 2015): these techniques have proven effective and are extensively present in the machine learning literature – Bishop (2006). Popular methods in this line of thought include algorithms such as bootstrap aggregations (Breiman 1996) or cross-validation (CV) (Bem and Allen (1974); Stone (1974)) which have prominent supporters such as Lambert (2018) or Bishop (2006). These methods are often relatively computationally

expensive but have proven very effective (Dormann et al. 2018). In general, the algorithms in this school of thought repeatedly and consistently sample data points and then compute an average metric or use more brute force methods such as grid searches. Interestingly, even though these methods are widely used, some of their properties are still unknown today; most research in the field has focused on optimizing parameters or hyperparameter values (e.g. number of folds, kernel smoothing) Yang (2007). This is done for a good reason since they are a critical component of empirical methods as they oversee computation cost and the overfitting aspects of a model. This includes studies such as Arlot and Celisse (2010), who survey the use of different cross-validation works and advise to select a large fold size if the ratio of noise to signal is large due to the bias in error estimation that is likely to occur. Furthermore, studies such as Kim (2009), which compares CV and bootstrapping simulations for model selection, are useful to know which technique to use in a given scenario - here the author concludes that .632+ bootstrap can suffer from a bias on large and small samples.

Although most empirical research focuses on the practical applications of these methods there are known theoretical results that are worth discussing in our context. First, it is important to know that CV for instance has multiple uses which all fall under the umbrella term of model selection : Zhang and Yang (2015) demonstrate that CV can be used for parameter tuning or selection between different models and confusing these tasks may lead to errors as for instance Leave-One-Out Cross-validation (LOO CV) is asymptotically optimal for non-parametric order of nesting selection but does not necessarily lead to the best model. This differentiation of tasks is important in our context: here we decide to focus on finding the best model between competing ones and not on the optimization of parameters. It is also important to understand that the methods presented are not mutually exclusive and are sometimes equivalent; for instance LOO CV is asymptotically equivalent to Akaike Information Criteria (AIC) (Akaike 1974).

AIC is part of the of the information theoretic (IC) family of model selection methods; in fact, it is its earliest member and is still widely used today (Cavanaugh 1997). The IC approach to model selection aims to compare the distance to the "truth" of each candidate model and select the one with the smallest distance (Blankenshipa, Perkinsb, and Johnsonc 2002). In the case of AIC this distance is the Kullback and Leibler (KL) divergence (Kullback and Leibler 1951). As the "truth" that describes the data is not known, it is estimated solely with the data using the principles of maximum likelihood

(Cavanaugh 1997). Here already it is important to note that IC approaches cannot generally be used across datasets but are only valid for the specific dataset they are computed on (Park 2018). This is for instance different from CV where the underlying assumption is that the distribution of each fold is similar, and the selected model could be applied to similar datasets. The fundamental purpose if IC methods is highlighted when comparing them to a typical machine learning methodology; in model selection there exists a fine balance between the complexity of the model and its fit to the data. A model with high complexity might overfit the data which comes at the cost of loss of degrees of freedom and lack of generalization. To strike this balance in machine learning, researchers typically hold-out one or multiple sets and test the model on the held-out data (e.g. CV). However, this is highly dependent on the quality of the held-out data determined by hyperparameters. IC methods are free from these issues and are aimed to penalize the cost of complexity while including the benefit of higher fit.

The BIC (unlike its name suggests) is neither Bayesian (as it is mainly based on maximum likelihood principles) nor strictly information theoretic as it does not use KL divergence (Park 2018). It solves selection issues with approximation of the marginal likelihood of the model. Often times AIC and BIC tend to agree as they are computationally similar, but they serve slightly different principles: it is advised to use the AIC to tie a metric towards the out-of-sample fit whereas the BIC can be used for strict model selection within the sample. This distinction boils down the researcher's reason to engage in the selection process (Park 2018). In information theoretic model selection, AIC and BIC are often put in competition. For instance, they are often compared in terms of asymptotic optimality under parametric and non-parametric assumptions (Shao 1997). Furthermore in terms of selection under low signal to noise ratios, they behave differently: BIC performs better when signal-to-noise is low or high whereas AIC performs better for more balanced datasets (Liu and Yang 2011). There have been different attempts to ensure better performance under noisy conditions such as the AICu proposed by McQuarrie, Shumway, and Tsai (1997). One such recent attempts was undertaken by Murari et al. (2019) who include Shanon entropy (Shannon 1948) in AIC and BIC. This method is implemented here and discussed further in section 4. In an interesting discussion entitled "Stop the war between AIC and BIC by CV", Zhang and Yang (2015) show that under specific CV settings the conflict between AIC and BIC in terms of asymptotical efficiency can be solved (in a homoskedastic setting). Within IC the main candidate metrics are AIC and BIC however in the past

few years there has been a shift to steer away from the former because the penalty term it includes is arbitrary - Lambert (2018)). Instead of the AIC researchers use the Deviance information criterion (DIC) - Gelman et al. (2004); Gelman, Hwang, and Vehtari (2014) - which is more Bayesian in nature as it uses the sum of the variance of MCMC posterior draws to penalize for complexity. BIC is also closely linked to the third main model selection methodology : BIC is a computational simple way of obtaining a conservative approximation of Bayes factor (BF) in the unit information space (Kass and Raftery (1995); Raftery (1999)).

Bayes factor is one model selection method in a purely Bayesian framework. It is intuitively straightforward but its difficulty lies in its analysis and computation (Chipman, George, and McCulloch 2001 ; VanderPlas 2014). In general, a Bayesian framework is constructed around Bayes rule which revolves around the likelihood, the prior and the posterior (Downey 2012). In a modelling setting the likelihood contains the description of the data given the model and the prior withholds the information that the researcher knows regarding the model. The posterior describes the information of the model given the data. BF then boils down to the ratio of marginal likelihoods (assuming constant priors) (Kass and Raftery 1995). Although BF is simply explained, it comes with two difficulties. The first one is analytically deriving the models that need to be estimated: as modelling gets more complex obtaining an analytical form in order to estimate them becomes harder (Vajpeyi Avi, Smith Rory 2016). The second is the computation costs associated to the models: to estimate a posterior, most models first require data simulations such as MCMC which adds to the computation cost of marginal likelihood integrations over the parameter space. This can make conputations intractable or impossible for complex models (VanderPlas 2014). However, when computed, Bayes factor can be robust even under noisy conditions. For instance, Vajpeyi Avi, Smith Rory (2016) managed to improve the study of binary black hole systems, which inherently entails noisy datasets, by using Bayes factor instead of previous methods that relied on Maximum likelihood estimations. It is important to note that Bayes factor is strongly criticized by prominent figures in the Bayesian statistics literature such as Gelman and Rubin (1995) who highlight that these methods do not make full use of the broad range of procedures allowed in a Bayesian setting and do not take into account the difference between model selection and model averaging. Instead Vehtari, Gelman, and Gabry (2017), propose to evaluate models combining the Watanabe Information Criterion (WAIC) - which is completely Bayesian since it includes all the values of the posterior

draw along, Watanabe (2013) - along with LOO. Building on this framework, to reduce computation time and take into account the time dimension of growth models, Paul-Christian Bürkner, Jonah Gabry (2019) proposed the Approximate leave-future-out CV for Bayesian time series.

3 Data

The data used in this project was generated data since there are several practical and methodological reasons for doing so. First, methodologically, as advised by Kéry and Royle (2016), generating data offers an ideal control environment under which parameters and hyper-parameters are known. Furthermore, in growth cell literature, from which this project stems (Harris et al. 2016) synthetic data is standard practice. Second, data from the commercial partner that was meant was to be analysed here was unavailable due to legal restrictions and no open source equivalents were found. As the synthetic data is at the heart of the analysis this section will describe in greater detail the data meant to be mimicked and the process and tools used to do so.

The type of generated process in this project is similar to a cell counting process proposed by CompuCell3d (Cickovski et al. 2005) with certain restrictions which led to custom data generation. The enforced restrictions are as follows:

At any time x we must be able to estimate the number of a given cell count. We have knowledge of the growth function that the cells take (i.e. $f(x)$). We also know that introducing an agent in our cell sample alters the growth path that the sample follows to another process (say $g(x)$). Given this information, we should be able to obtain the number of cells for any given time on the condition that we have knowledge of the presence or absence of the agent. However, if we do not know if the agent has been introduced in the sample then we must choose whether we estimate the numbers of cells using $f(x)$ or $g(x)$ based on the count. At this point, a simple model selection is sufficient to capture the correct model or even combine the two models if necessary. However, the primary difficulty with this is that the counting process is subject to a large amount of noise. Therefore, the problem at hand is to find the ideal model selection method under noisy conditions. In this case, the ideal model selection would favour a growth function that is able to identify the true growth process along with the corresponding parameters. Furthermore, it is interesting for the researcher to be able to

gauge the uncertainty surrounding the selected model and its parameters. The growth functions ($f(x)$ and $g(x)$) used in the study are described further in this section.

Within this context, the data generated is meant to mimic a growth process through time in which x represents the time through which the count y increases. To bind the problem the count was generated and then normalized using a simple min-max normalization. Therefore, we have:

$$x \sim U(0, 1000)$$

$$y \in [0, 1]$$

Although the normalization is clearly not realistic, it is ideal to bind the problem and does not undermine generalization.

The two growth functions used to generate the data are a simple linear function (1) and a logistic function (2) of the following forms :

$$f(x) = \alpha + \beta \times x \tag{1}$$

where α is the intercept and β the the coefficient of x and :

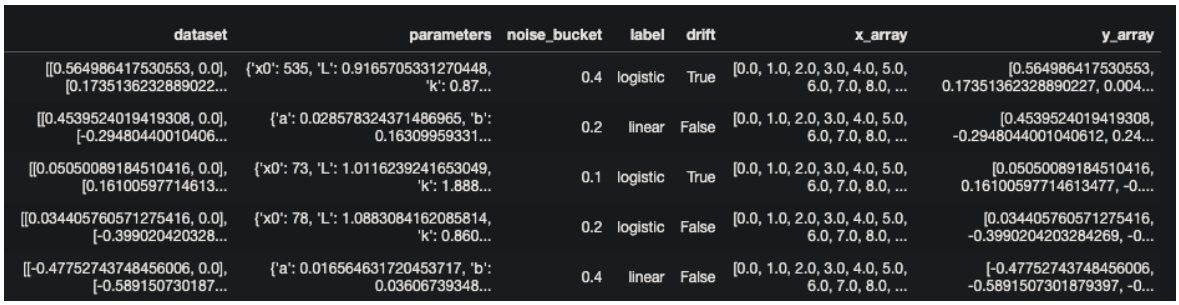
$$g(x) = \frac{L}{1 + e^{(-k(x-x_0))}} \tag{2}$$

where L describes the maximum value the curve could take, k describes the growth rate of the logistic function and x_0 represents the sigmoid's midpoint.

For all these parameters, the following values were uniformly drawn :

- $\alpha \sim U(0, 0.05)$
- $\beta \sim U(0, 0.2)$
- $L \sim U(0.9, 1.1)$
- $x_0 \sim U(\frac{max(x)}{4}, \frac{3max(x)}{4})$
- $k \sim U(0.5, 2)$

In order to simulate the noise in the problem and analyse it in a coherent manner, different levels of additive Gaussian errors are introduced. The Gaussian errors all have a mean of 0 and a variance σ ranging from 0.1 to 1 by intervals of 0.1. We refer to each



dataset	parameters	noise_bucket	label	drift	x_array	y_array
[[0.564986417530553, 0.0], [0.1735136232889022...	{'x0': 535, 'L': 0.9165705331270448, 'k': 0.87...	0.4	logistic	True	[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, ...	[0.564986417530553, 0.17351362328890227, 0.004...
[[0.4539524019419308, 0.0], [-0.29480440010406...	{'a': 0.028578324371486965, 'b': 0.16309959331...	0.2	linear	False	[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, ...	[0.4539524019419308, -0.2948044001040612, 0.24...
[[0.05050089184510416, 0.0], [0.161005977146 13...	{'x0': 73, 'L': 1.0116239241653049, 'k': 1.888...	0.1	logistic	True	[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, ...	[0.05050089184510416, 0.16100597714613477, -0....
[[0.034405760571275416, 0.0], [-0.399020420328...	{'x0': 78, 'L': 1.088308416208581 4, 'k': 0.860...	0.2	logistic	False	[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, ...	[0.034405760571275416, -0.3990204203284269, -0...
[[-0.47752743748456006, 0.0], [-0.589150730187...	{'a': 0.016564631720453717, 'b': 0.03606739348...	0.4	linear	False	[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, ...	[-0.47752743748456006, -0.5891507301879397, -0...

Figure 1: Sample of data used in tabular format

of these noise levels as noise buckets. Each noise bucket is comprised of 100 synthetic datasets. To generate the data, custom `numpy` functions were created and called. The meta-data regarding the created dataset is stored along with the data in nested `pandas` dataframes.

Table 1: Meta data of generated datasets for experiments

	Logistic examples	Linear examples	Logistic drift examples	Linear drift examples
Additive error var = 0.1	100	100	100	100
Additive error var = 0.2	100	100	100	100
Additive error var = 0.3	100	100	100	100
Additive error var = 0.4	100	100	100	100
Additive error var = 0.5	100	100	100	100
Additive error var = 0.6	100	100	100	100
Additive error var = 0.7	100	100	100	100
Additive error var = 0.8	100	100	100	100
Additive error var = 0.9	100	100	100	100
Additive error var = 1.0	100	100	100	100

To understand the robustness of any of the selection processes employed in this study, we add to the generated data a drift term which distorts the growth functional. The drift was added to a rescaled dataset and was uniformly distributed $drift \sim U[0.5, 1]$. From the synthetic data we record the x and y values (which are referred to as the datasets) as well as the label (i.e. "linear" or "logistic"), the set of corresponding parameters and the associated noise bucket. A subset of the data used is presented in figure 1 to provide clarity on the data used.

To clarify the problem we face in the selection process, figure 2 illustrates the distortion that occurs as we increase the noise in the data. The graphic on the left shows very little noise (smallest noise bucket in the data). Here one could easily eyeball the functional form associated to each dataset. However, the plot on the right demonstrates that as

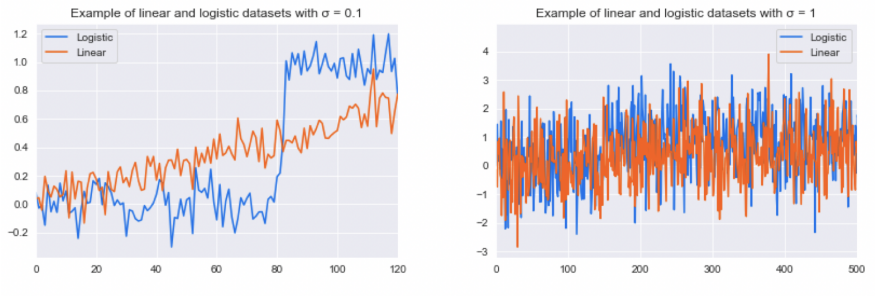


Figure 2: As noise increases the underlying process becomes harder to spot

we add noise in the data this task becomes less evident and requires a methodological identification process.

4 Methods

This section describes the different estimation strategies used for model selection. To compare how well the different methods performed, we observe the classification accuracy of the method (i.e. how well the datasets are classified as well and how confident the classification is in its choice). Furthermore, as the systematic error and the parameters are of interest, we also focus on the estimated parameter values for different noise levels as well as the estimation of the error itself when possible. Finally, the time/complexity required for the estimation is also an important aspect of the study. To structure this discussion, we first focus on selection strategies that are more often regarded as Frequentist methods [describe more] and then we highlight Bayesian model selection processes.

4.1 Frequentist model selection

As discussed in section 2, there are a battery of different methods to perform model selection, some of which are used in this study. However, before performing any selection it is important to estimate our models. Fitting the models to the data or curve fitting refers to the process of obtaining a mathematical function that can approximate a data. There are many approaches to solve such a problem but the most common one is to solve the least square shown in formally shown in equation (3).

$$\min_{\theta} \sum_{n=1}^{N} (y_n - \hat{y}(\theta, x))^2 \tag{3}$$

where $y(\theta, x)$ is $f(x, \theta)$ or $g(x, \theta)$ depending on the functional form chosen and θ is the vector of parameters. Least square aims at minimizing the sum of the distances between the fitted curve and the data points. Here a noticeable difference has to be highlighted between a linear functional form and a logistic one: The former represents an unbound optimization problem whereas the logistic function is by construction bound. This implies that different algorithms must be used to fit the two functions.

(i) Solving the linear least square problem: Solving the linear regression problem is straightforward and a common result. As such the details of solving it are not expanded on here. If needed, readers can refer to Wooldridge (2003) for mathematical derivation of ordinary least square problem.

(ii) Solving the bounded non-linear least square problem : To solve the logistic curve fitting problem, we employ the Trust Region reflective algorithm (Nocedal and Wright 2006) which given bounds, subsets the region of the objective function (in this case the equation ...) and gradually expands it each time an adequate model fit is obtained. In our case, the normalization of the data was key as the bounds given to the algorithm were [0,1]. Taking a step back from the synthetic data framework, in general binding the problem with known bounds is often valid since researchers would normally have an idea of the growth they are evaluating and can often determine an upper and/or a lower limit of the growth process.

Both (i) and (ii) were solved using `python`'s scientific library `scipy`.

Once each of the datasets were fit with a logistic and a linear regression, the model selection process can take place. We use a panel of different selection metrics and evaluate them based on classification accuracy, parameter estimation and computation cost.

Mean Squared Error and Mean Absolute Error: the naive approach

We begin the analysis with a naive approach to model selection by using the Mean Squared Error (MSE) and the Mean Absolute Error (MAE - defined as the average absolute value of the error). We do so as these are popular metrics in empirical machine learning. Both of these evaluate the average error that the model prediction would

generate and are naturally meant to be minimized. The constitute moments of the error as they encompass its variance and bias. A dataset is classified as linear if the MSE/MAE of the linear model is lower than the MSE/MAE of the logistic model (and vice-versa). However, we only use these metrics as a starting point: MSE and MAE are not the most suitable for selection outside of a CV process (Bishop 2006) as they do not take into account any model complexity. Consequently, a model with more parameters will by construction tend to cause less error but can break the rules of an appropriate model which aim to make a selection which would not overfit and is as simple as necessary (i.e. Occam's razor) – Cosma (2015). With this in mind we use information criteria which are more appropriate tools here.

BIC, AIC, and entropy enhanced BIC and AIC

In order to penalize the complexity of a model the most popular metrics used are the Bayesian Information Criteria (BIC) and the Akaike Information Criteria (AIC). They both aim at estimating the likelihood of a model to predict future values (ScienceDirect 2019) while balancing the benefit of good fit with the model's complexity. They are defined as:

$$AIC_a = -2ln(L) + 2k \tag{4}$$

$$BIC_a = -2ln(L) + 2ln(N)k \tag{5}$$

where L is the likelihood of the model, k is the number of parameters and N is the sample size. These measures are meant for selection problems such as the one at hand. However, in empirical work, as the likelihood is often difficult (if not impossible) to obtain, workarounds exist (often by making assumptions on the error term's distribution) such as the one applied here where using the `RegscorePy` package:

$$AIC_b = N \times ln(MSE) + 2k \tag{6}$$

$$BIC_b = N \times ln(MSE) + k \times ln(N) \tag{7}$$

This is done because the MSE is an estimate of the error's variance and since the error has mean 0, given a constant that can be dropped (since we compare Information Criteria on the same samples) we can replace the likelihood by the MSE. Regardless of the minor definition changes, the rule for model selection using AIC/BIC is to make a decision based on the lower Information Criteria value. Consequently, a similar classification rule as the MSE/MAE can be applied here. Since the problem at hand is to make appropriate model selection choices with respect to different noise levels in the data, we make an addition to our Information Criteria as suggested by Murari et al. (2019). In their study, the researchers demonstrate that including Shannon Entropy into the BIC and AIC can enhance the criteria, especially when the data is subject to a high amount of noise. The reasoning to this is, holding everything else constant, models which have a more uniform distribution of error should be favoured because for a perfect model, noise would only be coming from the data. To quantify the degree of uniformity of the error, Entropy is added by the authors in the following manner:

$$BIC_c = N \times ln(\frac{\sigma_e^2}{H}) + k \times ln(N) \tag{8}$$

$$AIC_c = N \times ln(\frac{MSE}{H}) + 2k \tag{9}$$

where σ_e^2 is the variance of the error and H is the Shannon entropy. Using our definition of BIC (equations (6), (7)) and combining it with Murari et al. (2019) we have :

$$BIC_H = BIC_b - Nln(H) \tag{10}$$

$$AIC_H = AIC_b - Nln(H) \tag{11}$$

which we estimate in this work since model selection in low signal to noise ratio is the subject of study. Note that we can safely meet the assumption of error normality of Murari et al. (2019) by checking the distribution of the errors. One such check is presented in figure 3.

The study at hand also serves as an extension to Murari et al. (2019) since the authors concluded that comparing their entropy enhanced AIC/BIC measures to Bayesian

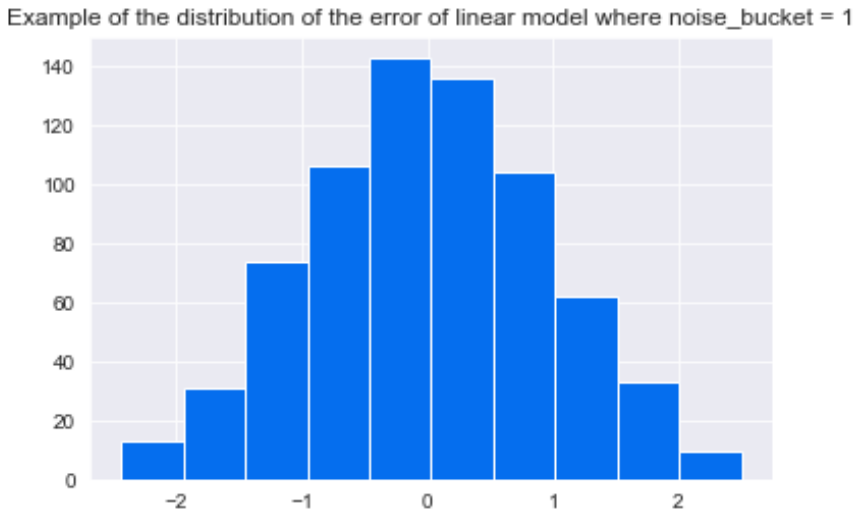


Figure 3: The assumption of normality of error has been checked - it is not a strong assumption in this case

selection approaches are an unexplored territory in the current literature. The main packages used in this part of the analysis were `RegscorePy`. Furthermore, since no `python` implementation of Murari et al. (2019) is currently available custom `numpy` functions were created.

χ^2 Selection to estimate uncertainty

In order to have a test that better quantifies the degree through which we select one model over another, a hypothesis test is required. In a frequentist context, to do so the goal is to calculate a statistic that relates to a distribution of which we know the properties. In our case, we choose a χ^2 distribution. This part of the discussion follows VanderPlas (2014) who describes details χ^2 model selection process. We assume that the errors are independent and normally distributed which would mean that the normalized sum of errors follows a χ^2 distribution. As outlined above, this assumption is not too strong for most of the fitted models and holds particularly true as the signal to noise ratio increases. Thereon we compute the χ^2 statistic which is the normalised sum of errors and follows a χ^2 distribution with the degrees of freedom related to the number of parameters in the model. From there we obtain the χ^2 likelihood (by referring to the values in distribution table). This number can be interpreted as the likelihood of observing the error values given our model.

This selection methodology is a useful addition to the methods outlined above because it can quantify the certainty of the classification made using hypothesis testing: by formulating a hypothesis and testing it on the difference of the χ^2 likelihoods as demonstrated by VanderPlas (2014). The only necessary condition is that the models must be nested which is the case here as we can write:

$$g(x) = S(f(x)) \tag{12}$$

where :

$$S(u) = \frac{L}{1 + e^u} \tag{13}$$

and $\beta = -k$ and $\alpha = kx_0$.

Consequently, we formulate our null hypothesis as the data following a linear generated process and find the p-values related to $\chi^2_f - \chi^2_g$.

It is noteworthy to mention that there may be caveats in this process : Schulze-hartung and Melchior (2014) point out that noise and non-linearity may adversely affect a χ^2 test. These results should be kept in mind for interpretation. The computations were done using `scipy`'s `stats` module.

4.2 Bayesian model selection

Another set of approaches in the literature use Bayesian methods to estimate models and the corresponding parameters. This methodology is more complex to implement which often hurdles practitioners – VanderPlas (2014). Here we provide an overview of a Bayesian approach and the parameter settings used in this study.

As in the section 4.1 we first contextualize and estimate the model before outlining the selection process. In general, a Bayesian model contains a set of parameters (and hyperparameters) θ. In the case of the logistic form $\theta = \{L, k, x_0, \sigma\}$ and for the linear model $\theta = \{\alpha, \beta, \sigma\}$. The modelling goal is to obtain the probability distribution of θ given the data (i.e. $P(\theta|D)$). Equation (14) provides the standard Bayes rule approximation where D is the data, the right-hand side is the posterior, the first term on the left-hand side is the likelihood and the second one is the prior.

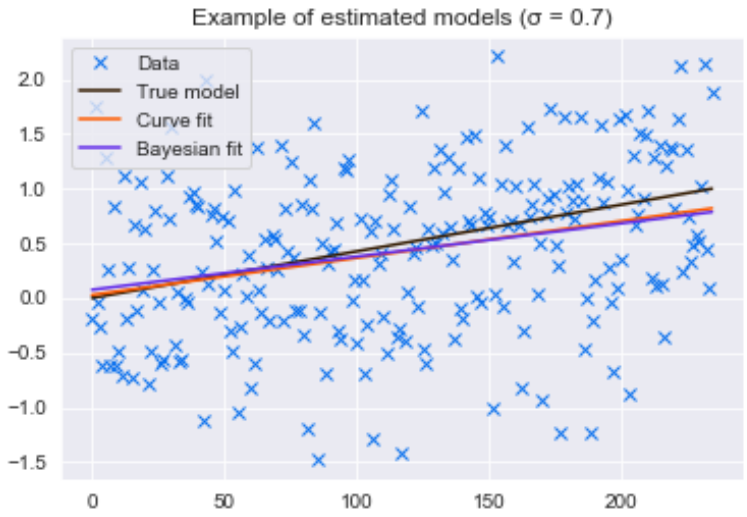


Figure 4: Example of estimation of Bayesian regression vs Least Squares fit

$$P(\theta|D) \propto P(D|\theta) \times P(\theta) \tag{14}$$

One particularity with a Bayesian model is that the prior distribution assigned to the parameters plays a crucial role in the obtained model. In our case, the set of priors considered are all bounded flat (i.e. the values of the parameters are uniformly drawn within given bounds) for the parameters and a Gaussian for the nuisance hyperparameter (σ). Different bounds were tried but as our expectation in a growth model is positive growth with our count unable to be negative, the largest bounds chosen were all positive real numbers. These priors are similar to those used in the literature and are quite general. Furethermore, as pointed out by Harris et al. (2016) they can be seen as uninformative. Note that flat priors are not necessarily uninformative, and Jeffrey's prior can also be used (they were also tried in certain experiments here). As posterior distributions are difficult to express analytically, as generally done in Bayesian problems we turn to MCMC.

MCMC is a numerical simulation that samples data from a given distribution where each future chain is only dependent on the present and not on all past chains. In our context these simulations are important as convergence of MCMC to the target distribution is a known result and by sampling enough data points from the posterior

we can estimate it (Ravenzwaaij, Cassey, and Brown 2018). The parameters required to run the MCMC simulation consists of the number of "walkers" (the number of chains used), the "burn-in" amount (the number of steps to discard from each chain) and the number of points sampled per chain. These parameters were set according to guidelines from documentation.

To implement MCMC, different `python` packages were tested: although `PyMC3` (Patil, Huard, and Fonnesbeck 2010) was the first option due to its popularity in the `python` community, it was too computationally expensive for the task at hand – it seems better suited when the number of datasets is small. Instead, we use the `emcee` package which is an implementation of Foreman-Mackey et al. (2013)'s affine-invariant ensemble sampler for MCMC. It proved quick and reliable in the tests conducted likely due to the fact that it is written originally in `python` which speeds up sampling and compilation process. To use `emcee` well, it is important to express the posterior in log form. Hence equation (15) we have:

$$log(P(\theta|D)) = log(P(D|\theta)) + log(P(\theta)) \tag{15}$$

We then assume that the data is independently and identically distributed and following $y \sim N(\hat{y}(\theta, x); \sigma^2)$. Therefore the log-likelihood function is given by :

$$log(P(D|\theta)) = -\frac{1}{2} \times \sum_{i=1}^{N} log(2\pi\sigma^2) + \frac{(y_i - \hat{y}(\theta, x_i))^2}{\sigma^2} \tag{16}$$

It is noteworthy to mention this is type of log-likelihood functions typically used in population research as demonstrated by Firat et al. (2016) which uses a similar Gaussian likelihood coupled with a logsitic model to study the growth of Japanese Quail.The flat prior terms are set as $log(P(\theta)) = 0$ for all positive parameters in θ. Note that σ represents the noise parameter - sometimes called the nuisance parameter and is also estimated here. One of the advantages of a Bayesian model is that parameters are obtained as distributions which allows us to make decisions on the preferred model in different ways and model uncertainty more accurately. Furthermore, plots such as figure 5 are used for closer inspection of parameters spaces and distributions.

Once a posterior is estimated for each model, classifications can be made using model selection techniques.

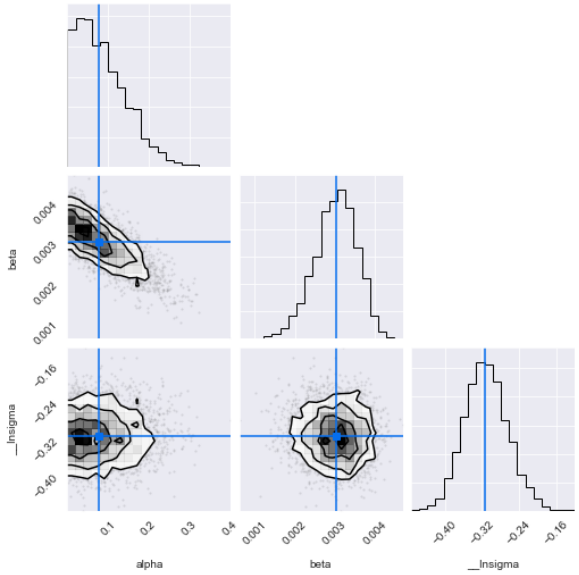


Figure 5: Example of MCMC samples for linear model

Bayes factor : Classic Bayesian model selection

One common selection method in Bayesian approaches is to calculate Bayes factor (BF) and use the table described by Raffety(1995) to select the better model. BF is described as the ratio of the likelihoods for different models. For instance, if we define our hypothesis that the data D is generated by $f(x)$ as H_0 and the alternative H_1 that the data is generated by $g(x)$ then Bayes factor is defined as :

$$BF = \frac{P(\theta_{H_0}|D)}{P(\theta_{H_1}|d)} \times \frac{P(\theta_{H_0})}{P(\theta_{H_1})} = \frac{P(D|\theta_{H_0})}{P(D|\theta_{H_1})} \tag{17}$$

Since there is no prior evidence favouring one model we set $\frac{P(\theta_{H_0})}{P(\theta_{H_1})}$ to 1. We can then compute BF by taking the ratio of the posterior distributions. From the MCMC computation it is then necessary to obtain the posterior. There are several ways to calculate the posterior values such as computing a harmonic mean of sampled values. However, this has been shown to render values that can stray away from the true distribution as shown by [citation]. Instead, we compute the integral over the parameter space of the marginal likelihoods given by :

$$P(D|\theta_{H_i}) = \int_{\theta} P(D|\theta_{H_i}) \times P(\theta_{Hi}) d\theta_{H_i} \tag{18}$$

where i corresponds to the hypothesis.

Note that for more complex models this computation is not possible as the number of integrations increases with the number of model parameters. Here for computation purposes we simplify the models by setting the value of L to 1 and during the calculation. This assumption is not strong because we know that the true value of $L \in [0.9, 1.1]$ - recall that L corresponds to the upper-limit of the logistic function. Also, in practice, researchers could either estimate this parameter or set it equal to a known upper bound.

Once the computation of BF complete, we compare the posteriors and select the highest one. Note that for practical reasons the scale given by Kass and Raftery (1995) could not be used since the models were too close one to another. In experiments, following Porciani (2012) we find that the median value of the posterior distributions is helpful fits well to the data. These values are recorded as well as the distributional properties of the error term and all corresponding computation times.

Modern approaches to Bayesian IC

A more recent and popular approach to Bayesian selection involves the principles of IC while taking advantage of the distributional nature of Bayesian models. Recall that IC aims to balance complexity with model fit. To do so it includes a term based on the likelihood which is subtracted by a complexity penalty term. Within a Bayesian framework, since parameters are distributions based on MCMC draws, a likelihood term and a complexity penalty can be created. This is the underlying principle of the DIC and WAIC. Since the latter is more robust we focus on it using the work outlined in Gelman, Hwang, and Vehtari (2014). Hence, from Gelman, Hwang, and Vehtari (2014), WAIC is defined as:

$$WAIC_a = -2(lldp - Pwaic) \tag{19}$$

where $lldp$ corresponds to the log predictive accuracy of the model computed as:

$$lldp = \sum_{i}^{n} log(\frac{1}{S} \sum_{s}^{S} P(y_i|\theta^s)) \tag{20}$$

where S corresponds to the number of MCMC simulation draws. Essentially, $lldp$ correponds to the log average fit for each draw at each datapoint. $Pwaic$ is computed

as :

$$Pwaic = \sum_{i}^{n} Var_s^S(log(P(y_i|\theta))) \tag{21}$$

which is the sum of the sample variances log. Since no `python` implementation of WAIC is currently available, its computation was done using custom `numpy` functions. As with other IC methods the decision rule is based on the model with the lowest WAIC.

5 Results

To discuss the results, we proceed by evaluating the different aspects of the classifiers study.

5.1 Classification and Noise

The first aspect we focus on is the quality of our classifiers: how do the different model selection strategies perform as we increase the amount of noise in the data. This is of course the most crucial aspect of a good classifier. Furthermore, inspecting accuracy can be a clear medium to evaluate and understand the consequence of model selection strategies. Before inspecting results a few hypotheses can be outlined. First, we naturally expect the quality of the classification strategy to decrease as the signal to noise ratio weakens. However, since the entropy-based measure is meant to better the performance of the IC under noisy conditions, we postulate that it should perform higher than the other frequentist IC measures. Additionally, since WAIC is a stronger measure than BF as it takes into account all posterior draws from the MCMC computation, we expect it to outperform BF (and BIC since they are closely linked). Also, since we must make an assumption on the value of σ in the frequentist framework whereas it is estimated in the Bayesian one, it can make an hypothesized that Bayesian values can outperform Frequentist classifiers. On the other hand, there are clear weaknesses in the Bayesian methods used - particularly BF. Since the BF values were too close to each other and we do not want datasets unclassified, it is not possible to use Kass and Raftery (1995)'s scale : said otherwise, there was no acceptance threshold to ascertain one model was better than another. Although the scale has been criticized as being arbitrary, it is still viewed in a similar light as p-values. Furthermore, Bayesian

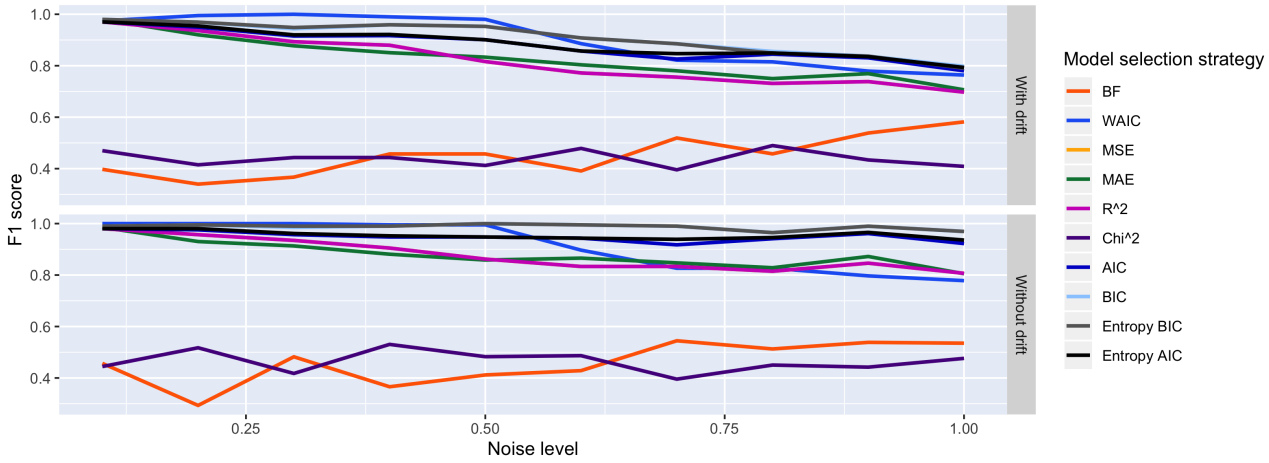


Figure 6: Certain model selection procedures have proven to be strong classifiers - zoomed in version in appendix

models are dependent on the priors assigned. Although the priors given were flat, as demonstrated by VanderPlas (2014), these can still hold information which can lead to wrong estimations. Another simple expectation is that the classifiers perform better in datasets without drifts since including the drift term guarantees that the underlying data generating function is not the same as the ones evaluated; nonetheless, in terms of model selection the closest functional form should still be identified. Finally, being able to quantify the certainty of selecting a model over the other is also an aspect of the analysis of interest here. To do so, the χ^2 test and the deviances the WAIC measures can provide insight. We expect naturally a negative correlation between certainty of model selection and signal to noise ratio. To study the classification quality, we look at classic classification measures such as the F1 score and accuracy for different noise levels (figure 6) as well simple confusion matrices (figure 7) for more closer inspections.

Figure 6 presents the F1 scores at each noise bucket for datasets with drifts and without drifts separately - note that a figure with the same analysis for accuracy and plots for with only the top classifiers are presented in the appendix. The results are in line with many previous findings outlined in sections 2 and 4 with a few surprises. First, as expected the F1 scores and accuracies (cf Appendix figure) decrease with the level of noise. This is naturally even more pronounced for the datasets that contain a drift term as the distortion is increased. Additionally, we note that most of the model selection methods perform well (over 95% accuracy) when the variance of the nuisance parameter is under 0.5. The most accurate overall frequentist technique was the entropy enhanced

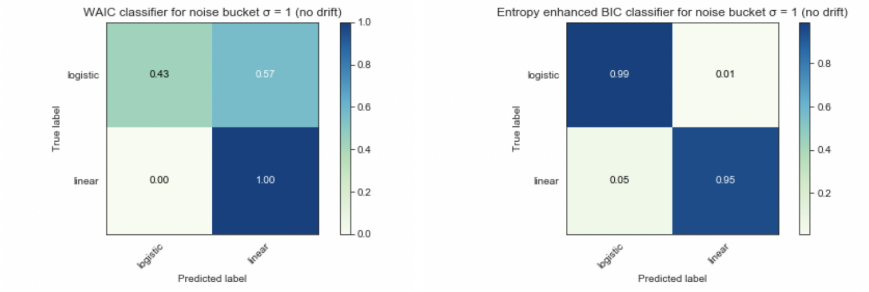


Figure 7: Each model selection method weights models differently

BIC with an overall accuracy of 99% without drift (91% with drift) whereas for the best Bayesian method it was the WAIC with an overall average accuracy of 89% without drift (88% with drift). This already demonstrates that although the selection is on generally more accurate for the frequentist method, it is less robust and vulnerable to small changes. Furthermore, another a priori surprising insight is the weakness of the selections using BF and χ^2 which on average respectfully classified models correctly 51% and 49% of the times. These scores are very low but can easily be explained by the fact that these measures trickle from formal statistical test that imply levels of confidence. Said otherwise, the confidence level is not high enough to fail to reject our null hypothesis whereas the other measures only select a model if the metric is smaller (or larger) than the competing value without taking confidence into account.

One very interesting finding is the confirmation that entropy enhanced ICs proposed by Murari et al. (2019) are more robust than simple ICs. Furthermore, the criticism of BF by Gelman and Rubin (1995) and his encouragement to adopt WAIC can be demonstrated by these results as WAIC is not only better in selection but is alse more robust to change than BF. This robustness is likely due to the fact that it to takes into account the MCMC draws which are not used to their fullest with BF. However, surprisingly WAIC is very stable and robust with low to medium noise levels (F1 score close to 99%) but quickly drops after that. This is in line with work such as Evans (2019) who shows that among the different metrics AIC and BIC are the most stable. Finally, in line with the literature, BIC performs well with low signal-to-noise ratios (under $\sigma = 0.5$) but then AIC seems to be more stable.

One surprising finding is presented in figure 7: The classification type of error differs

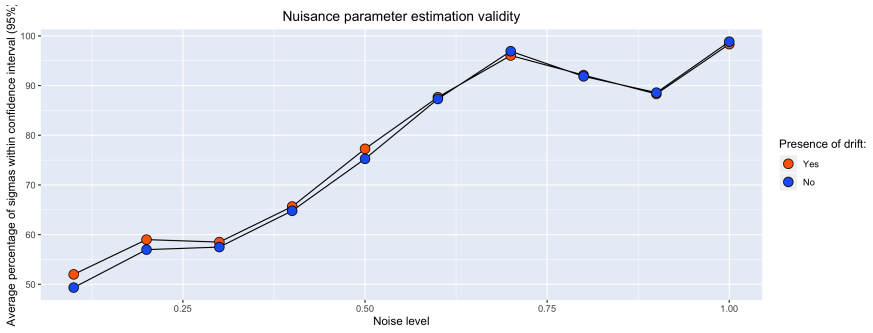


Figure 8: Each model selection method weights models differently

depending on the model selection method used. If we compare the most accurate Bayesian selection method (WAIC) and its frequentist counterpart (entropy enhanced BIC) then we notice that the classifiers' shortcomings are different. When mistaken, WAIC seems to misidentify logistic models for linear ones (false positives) whereas the BIC method has a higher rate of false negatives. This finding can be explained by the priors established within the Bayesian estimation. It is likely that the linear priors do not affect the linear model and the logistic one in the same manner. This is difficult to establish clearly but if indeed the case, this could place emphasis on one model more than on another. Another possible explanation would be that the same number of MCMC draws are not equal in estimating different models. The latter explanation would be in line with literature such as Liu, Nordman, and Meeker (2016) which cautions on the number of draws required by MCMC.

In terms of identifting the nuisance parameter, the Bayesian estimation can provide insights. Within the estimated parameters θ we include the variance of the error. Consequently, the posteiror drwas include a value corresponding to the σ - the measure of vairance of the additive Gaussian noise. Estimating this value can help researchers understand to what degree their data is affected by noise. To do so, report the findings around this term, we create the 95% confidence interval (assuming Gaussian ...) of the median MCMC chain and evalute at each noise level, what proportion of σs are within the confidence interval. The result is reported in figure 8. Here, the trend is clear, as the signal-to-noise decreases, the better the variance of the error is estimated. This is intuitive since as the errors become larger, their standard errors minght increase which leads to larger bounds to capture estimate the variance of the error. This result is in line with findings in section 5.2 that favour choosing Bayesian estimationa and selection when the researcher wishes to focus on the study of parameters.

5.2 Parameter Estimations

Parameter estimations in this study can be analysed in two ways: how close estimated parameters are from the true parameter values and how much they differ from each other between models. In order to study this matter for the frequentist method each fitted model parameter values and their standard errors were reported. For the Bayesian selections, we select the median value of the parameter distributions given the MCMC draws as these have been proven to be good point estimates (Porciani 2012) along with their standard errors. All standard errors are used assuming a Gaussian distribution of estimate and the confidence interval is set at 95%. We can then evaluate among the true positive and true negatives classifications the proportion whose parameters are within the confidence (or credible for the Bayesian view) intervals (we assume normally distributed parameters). This can be done at different noise levels in order to gain insight on how much confidence changes as the noise increases - expectations here are a negative correlation between noise and estimation quality. It is important to note that since estimation is done before selection, the findings here merely present whether or not the selected models are have confident estimations - and not whether the selection produces confident estimations.

First, we focus on the parameter estimations of the linear model. We pay particular attention to β rather than α since it is the main parameter of interest. Figure [appendix] shows for the two best estimation methods in terms of accuracy the percentage of β parameters that lie within the confidence/credible intervals outlined above. It can be interpreted in the following manner: among the accurate selection of the Entropy BIC for $\sigma = 1$, 73% capture the true value of β. The surprising result here is the clear performance difference between the two selection methods : BIC selected models performed better on estimation for any noise level. This is surprising since the WAIC also favoured linear models, one might expect better parameter estimation. These results were similar for comparing all different Bayesian metrics against the frequentist ones. Although this might be surprising, it is important to bear in mind that Bayes selection are strongly dependent on MCMC samples. In practice, MCMC is not scaled for different datasets in this fashion which implies MCMC parameters and the estimations that follow are usually tailored to a specific dataset. Furthermore, here median values of the parameter distribution are reported and used for credibility intervals. Upon closer inspection of the parameters it seems also seems that the parameter estimations are finer tuned for the Bayesian model as the standard errors are much smaller. Moreover,

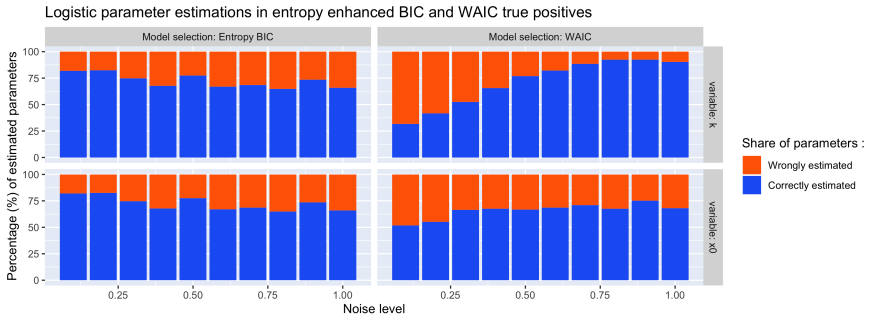


Figure 9: Parameter estimations are better using Bayesian methods in logistic case

the estimation parameters are mostly off by very small margins - the estimated bounds would need to change by less than 0.1% on average for the true parameters to be within the confidence intervals. Since this estimate was done for the median a small change in the point estimate such as the mode or the mean of the posterior would have altered this finding. Another insight in line with theory is the better performance of the χ^2 selection : at the highest noise level 79% of parameters were within the estimated bounds. This is expected since the χ^2 test was much more severe in its selection.

Now, we look more closely at the parameter estimations of the logistic model. Figure 9 plots the share of parameters within confidence bounds for the true positives chosen by the Entropy BIC and the WAIC. Here there are many interesting findings that can be useful for practitioners. First, we notice that on average the entropy BIC performs reasonably well with the parameter k being correctly estimated 71% of the time. This is comparable to the WAIC model which correctly estimates it 72% of the time. However, the greater difference lies on their respective performance as the noise increases. We notice that as the signal-to-noise ratio increases, the Bayesian model performs better which demonstrates the clear advantage of obtaining parameters as density functions instead of point estimates. Furthermore, the standard errors of the Bayesian models are still on average smaller than those in the frequentist framework. Note that when BF is chosen as the selection process this trend does not change. Although this trend is not clearly visible for the parameter x_0 the standard errors remain smaller as the noise increases. These findings suggest that for non-linear models, if parameters are of interest, Bayesian model selections might be better suited if there is a particular interest in the estimated parameters.

5.3 Computation cost

In terms of computation costs there are little surprises expected as the methods used are well documented. Naturally, computation costs are strongly dependent on the specific implementations but in this case, it is unlikely to impend generalization due to the fact that the methods used are straightforward. We expect that frequentist methods of model selection are quicker in computation cost since they do not require MCMC. For comparability we did not take into account the MCMC computation costs in the measurements. Another expectation is that the costliest method would be BF since it requires integrations whereas the other are relatively quick since they are only composed of vectoral sums.

The expected results were indeed confirmed: BF was the longest in terms of computation taking on average 5 seconds per dataset computation (excluding MCMC computation cost). This is quite slow considering that the computation costs of all other strategies were close to insignificant (all smaller than 1ms on all dataset computations). Figure 10 describes the proportional computational costs of the different strategies used (we exclude BF in figure 10 for clarity). The main interesting finding here is that once the MCMC has been run it not only offers a more versatile result in terms of parameter estimation but WAIC is also within the computation cost other frequentist methods (in fact in this study it was faster than other methods).

6 Discussion

This section aims at synthesising the findings highlighted in section 5 and offering a discussion around them. It is first insightful to notice the similarities and differences the findings have with some previous results. In line with expectations, AIC and BIC were accurate at their choice of selected model with an average accuracies of 92% and 95% respectivelly over all noise levels. Furthermore, all selection strategies became less accurate with noise but at different rates. The methods that are not advised to be used for selection such as MAE, MSE or adjusted R^2 decreased the fastest whereas more robust processes like IC performed better. On the other hand, formal statistical tests did not work well as selection methods, likely due to the certainty of classification. Furthermore, the addition of a drift term as a robustness check confirmed the accuracy trends as the order of the dataset classifiers remained the same. Finally, there were

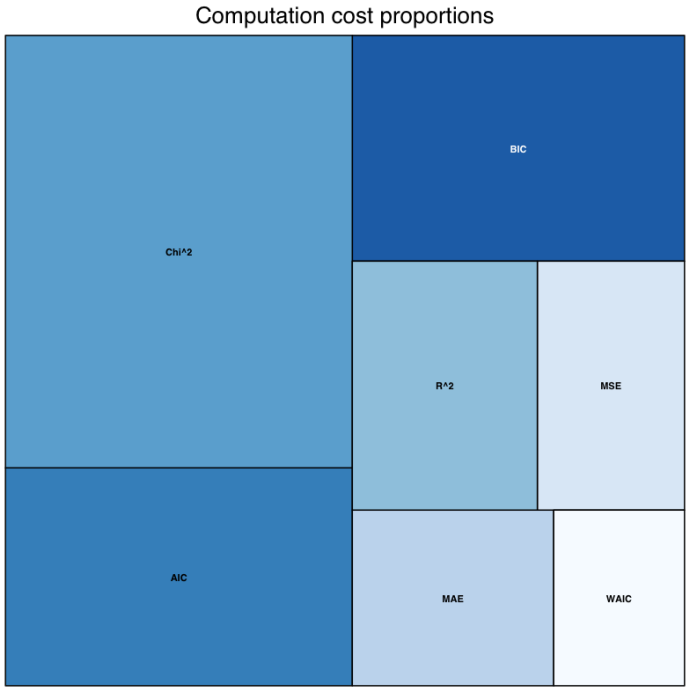


Figure 10: WAIC is the fastest model selection method if we exclude MCMC costs

interesting insights on the more recent selection methods such as entropy enhanced ICs or WAIC: the former confirmed the results demonstrated by Murari et al. (2019) and proved that entropy can make selection more robust under noise; the latter demonstrated results consistent with Evans (2019) who finds that WAIC does not always match BIC and AIC results.

One novel finding is the accuracy that WAIC has under low to medium noise levels (up to 98% at medium noise levels). This suggests that WAIC can be used with a high level of confidence for selections under $\sigma = 0.5$. Moreover, the consistency of entropy enhanced methods would suggest that practitioners unsure about noise levels in the data might consider entropy AIC or BIC metrics. This can also be done quickly since computation costs are fairly low compared to BF, for instance. On the other hand, if researchers are more interested in studying the parameters of the model for inference, using Bayesian estimations coupled with WAIC might be more appropriate as they provide distributions which can be explored in different ways and the selection method is accurate in parameter estimation. Moreover, with contextual knowledge the estimation can be enhanced with informative priors although all this is at the expense of a high computation cost due to MCMC.

While these insights can be helpful for a researcher aiming to detect a growth path using model selection, certain caveats of the study should be considered. First, although the underlying functional forms correspond to some real-world examples such as population growth, they remain simplistic and more experiments are necessary to generalise results better. This can be done by considering other functional forms such as polynomial or exponential functions along with various distortions and noise types.

Second, it was naturally not possible to include all the model selection strategies that exist in the literature and a choice of the most prominent was done. Some notable missing example is DIC for instance which was excluded as it is theoretically more restricted than WAIC. There exists a large palette of selection strategies and testing the ones present in this study against others can only contribute further to the field.

Third, within each topic, a dataset is often subject to a large body of literature that can inform the researcher of the most appropriate ways to tackle selection depending on the goal of the study. In economics, interpretability of coefficients is a crucial: fitting a high degree polynomial for instance would not make sense regardless of the model selection choice. Therefore, model selection is meant to be kept as a complement of theory, not a replacement.

Finally, we note that selection strategies were compared between different estimation methods - Bayesian vs frequentist. An argument can be made that due to different estimation strategies AIC and WAIC for instance cannot be compared. In fact, the reason this was not discussed earlier in the study is because under these estimation differences lies a philosophical difference between a Bayesians and frequentists. Bayesians consider probability as a degree of belief whereas frequentists see probabilities as a frequency. Since this is not the subject matter and a large literature on this topic exists, readers are encouraged to consult VanderPlas (2014) which provides a good overview of the differences between the two schools of thought. In terms of estimation, comparing metrics between the two methods is often done and these differences in estimations cannot be avoided.

7 Evaluation, Reflections, and Conclusions

This section aims to evaluate the overall study conducted and reflect upon its process and outcomes. The objective of this study was to evaluate empirically the use of model

selection in order to detect different forms of growth functions in a noisy setting. To do so, experiments using synthetic data were conducted and evaluated at different signal-to-noise levels. Traditional model selection strategies such as IC or BF as well as more cutting-edge methods such as WAIC or entropy enhanced ICs, were evaluated to understand their robustness in the selection process. Overall, this objective was attained and the findings presented are insightful for empiricists attempting to select a growth model in a noisy setting. Three main findings can be outlined as novel and an addition to the literature. First, WAIC - a more recent Bayesian selection method - can be used confidently for low to medium levels of noise (with an additive Gaussian error with mean 0 and up to a variance of 0.5). With higher noise levels, WAIC selection accuracy declines quickly. Second, if parameters are of interest in the selected models, using Bayesian selection methods as the noise increases might lead one closer to the true parameter values and therefore decreases chances of consequent wrong inferences. Third, as suggested in a recent finding, enhancing AIC or BIC with a measure of entropy can create robustness in the selection process, even with a high level or noise. This is a particularly useful result for empiricist who do not know the amount of noise present in the data and still wish to select the best model.

In terms of planning, the project was successful as most the initial commitments (outlined in the project proposal - Appendix A) were delivered. However, a few caveats and notable differences can be noted. The most major problem was the absence of a real-world set that would have been able to provide stronger evidence to the results provided. This was unfortunately a hurdle that we were unable to surpass due to most dataset in this field either being proprietary or too small in size. Additionally, originally more work was meant to be done on obtaining posterior distributions based on the work of Harris et al. (2016). However, after closer inspection of the results of their study, it seemed that the posteriors they obtained were difficult to generalise: in their work, the framework was set on pipetting and measurement errors that occur in biological growth. Adding this type of error in the present work would be the equivalent of adding a bivariate distribution (error or no error) to the likelihood which would not have been sufficient to achieve the goal of the study. Another smaller difference from the original plan was the absence of a noise-filter. Removing the noise filter from the selection process was a decision the closer inspection of model selection with respect to noise. This project was particularly useful as it provided an opportunity to explore the area of model selection and connect with its most recent advances. There are several ways

to expand this work that can be considered. First, using observed data to confirm the experiments dons here is important. Second, some of the open source tools used can be contributed to be the addition of the functions used in this study. This can be done by expanding them and adding them to existing packages for instance. Third, more model selection methods need to be considered. A notable one was outlined recently by Paul-Christian Bürkner, Jonah Gabry (2019) who proposed the Approximate leave-future-out CV for Bayesian time series. Comparing its robustness to the methods used in this study should help in the understanding of its robustness and build a bridge between empirical selection techniques, and the Bayesian ones demonstrated here.

8 Analysis Appendix

Table 2: F1 score by model selection

Noise level	Drift	BF	WAIC	MSE	MAE	R2	Chi2	AIC	BIC	Shannon BIC	Shannon AIC
0.1	FALSE	0.46	1.00	0.98	0.99	0.98	0.44	0.98	0.99	0.99	0.98
0.1	TRUE	0.40	0.97	0.97	0.98	0.97	0.47	0.97	0.98	0.98	0.97
0.2	FALSE	0.29	1.00	0.96	0.93	0.96	0.52	0.98	1.00	1.00	0.98
0.2	TRUE	0.34	0.99	0.94	0.92	0.94	0.41	0.95	0.97	0.97	0.95
0.3	FALSE	0.48	1.00	0.93	0.91	0.93	0.42	0.96	0.99	0.99	0.96
0.3	TRUE	0.37	1.00	0.89	0.88	0.89	0.44	0.92	0.94	0.95	0.92
0.4	FALSE	0.37	1.00	0.90	0.88	0.90	0.53	0.95	0.99	0.99	0.95
0.4	TRUE	0.46	0.99	0.88	0.85	0.88	0.44	0.92	0.96	0.96	0.92
0.5	FALSE	0.41	1.00	0.86	0.86	0.86	0.48	0.95	1.00	1.00	0.95
0.5	TRUE	0.46	0.98	0.82	0.83	0.82	0.41	0.90	0.95	0.95	0.90
0.6	FALSE	0.43	0.90	0.83	0.87	0.83	0.49	0.94	1.00	1.00	0.94
0.6	TRUE	0.39	0.89	0.77	0.80	0.77	0.48	0.86	0.91	0.91	0.86
0.7	FALSE	0.54	0.83	0.83	0.85	0.83	0.40	0.92	0.99	0.99	0.94
0.7	TRUE	0.52	0.82	0.76	0.78	0.76	0.40	0.83	0.89	0.89	0.85
0.8	FALSE	0.51	0.83	0.81	0.83	0.81	0.45	0.94	0.96	0.96	0.95
0.8	TRUE	0.46	0.82	0.73	0.75	0.73	0.49	0.84	0.86	0.85	0.85
0.9	FALSE	0.54	0.80	0.85	0.87	0.85	0.44	0.96	0.99	0.99	0.97
0.9	TRUE	0.54	0.78	0.74	0.77	0.74	0.43	0.83	0.84	0.84	0.84
1.0	FALSE	0.54	0.78	0.81	0.81	0.81	0.48	0.92	0.97	0.97	0.94
1.0	TRUE	0.58	0.76	0.70	0.71	0.70	0.41	0.78	0.80	0.79	0.79

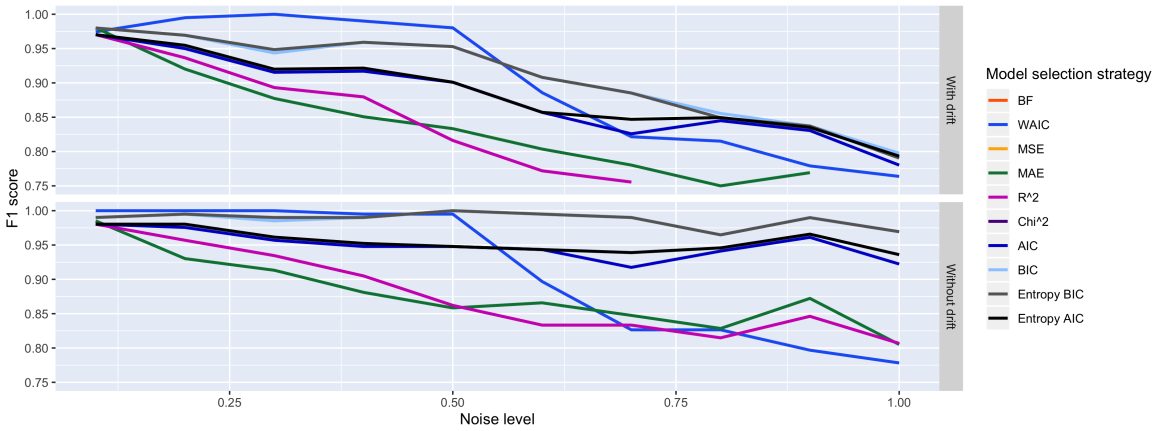


Figure 11: Best model selction models in terns of F1 score

Table 3: F1 score by model selection

Noise level	Drift	BF	WAIC	MSE	MAE	R2	Chi2	AIC	BIC	Shannon BIC	Shannon AIC
0.1	FALSE	0.56	1.00	0.98	0.98	0.98	0.50	0.98	0.99	0.99	0.98
0.1	TRUE	0.53	0.98	0.97	0.98	0.97	0.52	0.97	0.98	0.98	0.97
0.2	FALSE	0.47	1.00	0.96	0.92	0.96	0.52	0.98	1.00	1.00	0.98
0.2	TRUE	0.50	1.00	0.94	0.92	0.94	0.45	0.95	0.97	0.97	0.96
0.3	FALSE	0.56	1.00	0.93	0.90	0.93	0.47	0.96	0.98	0.99	0.96
0.3	TRUE	0.50	1.00	0.89	0.87	0.89	0.48	0.92	0.94	0.95	0.92
0.4	FALSE	0.48	1.00	0.90	0.86	0.90	0.54	0.94	0.99	0.99	0.95
0.4	TRUE	0.52	0.99	0.87	0.84	0.87	0.48	0.92	0.96	0.96	0.92
0.5	FALSE	0.50	1.00	0.84	0.84	0.84	0.48	0.94	1.00	1.00	0.94
0.5	TRUE	0.52	0.98	0.80	0.81	0.80	0.43	0.90	0.96	0.96	0.90
0.6	FALSE	0.48	0.88	0.80	0.84	0.80	0.52	0.94	1.00	1.00	0.94
0.6	TRUE	0.46	0.88	0.74	0.78	0.74	0.51	0.86	0.92	0.92	0.86
0.7	FALSE	0.52	0.79	0.80	0.82	0.80	0.46	0.91	0.99	0.99	0.94
0.7	TRUE	0.50	0.78	0.72	0.76	0.72	0.46	0.82	0.90	0.90	0.85
0.8	FALSE	0.52	0.79	0.78	0.80	0.78	0.48	0.94	0.96	0.96	0.94
0.8	TRUE	0.49	0.78	0.70	0.72	0.70	0.50	0.86	0.87	0.86	0.86
0.9	FALSE	0.52	0.74	0.82	0.86	0.82	0.47	0.96	0.99	0.99	0.96
0.9	TRUE	0.52	0.72	0.72	0.76	0.72	0.46	0.84	0.86	0.86	0.85
1.0	FALSE	0.50	0.72	0.77	0.77	0.77	0.50	0.92	0.97	0.97	0.94
1.0	TRUE	0.54	0.70	0.67	0.68	0.67	0.46	0.80	0.83	0.82	0.82

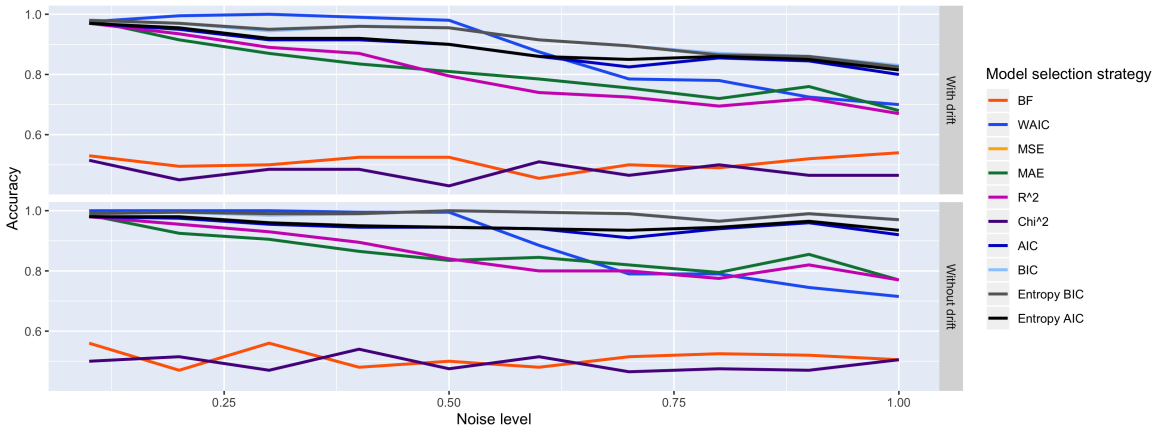


Figure 12: Accuracy of model selection methods

References

Aho, Ken, Dewayne Derryberry, and Teri Peterson. 2014. "Model selection for ecologists: The worldviews of AIC and BIC." https://doi.org/10.1890/13-1452.1.

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*. https://doi.org/10.1109/TAC.1974.1100705.

Amaran, Satyajith, Nikolaos V. Sahinidis, Bikram Sharda, and Scott J. Bury. 2016. "Simulation optimization: a review of algorithms and applications." *Annals of Operations Research*. https://doi.org/10.1007/s10479-015-2019-x.

Arlot, Sylvain, and Alain Celisse. 2010. "A survey of cross-validation procedures for model selection." *Statistics Surveys*. https://doi.org/10.1214/09-SS054.

Bem, Daryl J., and Andrea Allen. 1974. "On predicting some of the people some of the time: The search for cross-situational consistencies in behavior." *Psychological Review*. https://doi.org/10.1037/h0037130.

Bishop, Christopher M. 2006. *Machine Learning and Pattern Recogniton*.

Blankenshipa, Erin E., Micah W Perkinsb, and Ron J. Johnsonc. 2002. "THE INFORMATION-THEORETIC APPROACH TO MODEL SELECTION: DESCRIPTION AND CASE STUDY." *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1200.

Breiman, Leo. 1996. "Bagging predictors." *Machine Learning*. https://doi.org/10.1007/

bf00058655.

Brooks, Steve, Andrew Gelman, Galin L. Jones, and Xiao Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*.

Buehler, David A., Timothy J. Mersmann, James D. Fraser, and Janis K. D. Seegar. 1991. "Effects of Human Activity on Bald Eagle Distribution on the Northern Chesapeake Bay." *The Journal of Wildlife Management*. https://doi.org/10.2307/3809151.

Cavanaugh, Joseph E. 1997. "Unifying the derivations for the Akaike and corrected Akaike information criteria." *Statistics and Probability Letters*. https://doi.org/10.1016/s0167-7152(96)00128-9.

Chipman, Hugh, Edward I. George, and Robert E. McCulloch. 2001. "The Practical Implementation of Bayesian Model Selection." In. https://doi.org/10.1214/lnms/1215540964.

Cho, Jaeseol, and Kevin D. Dorfman. 2010. "Brownian dynamics simulations of electrophoretic DNA separations in a sparse ordered post array." *Journal of Chromatography A*. https://doi.org/10.1016/j.chroma.2010.06.057.

Cickovski, Trevor M., Chengbang Huang, Rajiv Chaturvedi, Tilmann Glimm, H. George E. Hentschel, Mark S. Alber, James A. Glazier, Stuart A. Newman, and Jesús A. Izaguirre. 2005. "A framework for three-dimensional simulation of morphogenesis." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/TCBB.2005.46.

Claeskens, Gerda, and Nils Lid Hjort. 2008. *Model selection and model averaging*. https://doi.org/10.1017/CBO9780511790485.

Cosma, Shalizi. 2015. "Lecture 21 : Model selection."

Ding, Jie, Vahid Tarokh, and Yuhong Yang. 2018. "Model Selection Techniques: An Overview." *IEEE Signal Processing Magazine*. https://doi.org/10.1109/MSP.2018.2867638.

Dormann, Carsten F., Justin M. Calabrese, Gurutzeta Guillera-Arroita, Eleni Matechou, Volker Bahn, Kamil Bartoń, Colin M. Beale, et al. 2018. "Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference." https://doi.org/10.1002/ecm.1309.

Downey, Allen B. 2012. *Think Bayes: Bayesian Statistics in Python.*

Evans, Nathan J. 2019. "Assessing the practical differences between model selection methods in inferences about choice response time tasks." https://doi.org/10.3758/s13423-018-01563-9.

Fernandez, Miguel Angel Luque. 2015. "Cross-validation." In *Faculty of Epidemiology and Population Health Department of Non-Communicable Disease.*

Firat, Mehmet, Emre Karaman, Ebru Başer, and Dogan Narinc. 2016. "Bayesian Analysis for the Comparison of Nonlinear Regression Model Parameters: An Application to the Growth of Japanese Quail." *Revista Brasileira de Ciência Avícola* 18 (September): 19–26. https://doi.org/10.1590/1806-9061-2015-0066.

Foreman-Mackey, Daniel, David W. Hogg, Dustin Lang, and Jonathan Goodman. 2013. "emcee : The MCMC Hammer." *Publications of the Astronomical Society of the Pacific.* https://doi.org/10.1086/670067.

Fu, Michael C., Sigrún Andradóttir, John S. Carson, Fred Glover, Charles R. Harrell, Yu Chi Ho, James P. Kelly, and Stephen M. Robinson. 2000. "Integrating optimization and simulation: Research and practice." *Winter Simulation Conference Proceedings.* https://doi.org/10.1109/WSC.2000.899770.

Gelman, A, J B Carlin, H S Stern, and D B Rubin. 2004. "Bayesian Data Analysis Second Edition.PDF." https://doi.org/10.1002/wcs.72.

Gelman, Andrew, Jessica Hwang, and Aki Vehtari. 2014. "Understanding predictive information criteria for Bayesian models." *Statistics and Computing.* https://doi.org/10.1007/s11222-013-9416-2.

Gelman, Andrew, and Donald B. Rubin. 1995. "Avoiding Model Selection in Bayesian Social Research." *Sociological Methodology.* https://doi.org/10.2307/271064.

Gerrodette, T. 1987. "A power analysis for detecting trends." *Ecology.* https://doi.org/10.2307/1939220.

Gottschalk, Paul G., and John R. Dunn. 2005. "The five-parameter logistic: A characterization and comparison with the four-parameter logistic." *Analytical Biochemistry.* https://doi.org/10.1016/j.ab.2005.04.035.

Harris, Edouard A., Eun Jee Koh, Jason Moffat, and David R. McMillen. 2016. "Auto-

mated inference procedure for the determination of cell growth parameters." *Physical Review E*. https://doi.org/10.1103/PhysRevE.93.012402.

Johnson, Jerald B., and Kristian S. Omland. 2004. "Model selection in ecology and evolution." https://doi.org/10.1016/j.tree.2003.10.013.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes factors." *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1995.10476572.

Kéry, Marc, and J. Andrew Royle. 2016. "Introduction to Data Simulation." In *Applied Hierarchical Modeling in Ecology*. https://doi.org/10.1016/b978-0-12-801378-6.00004-7.

Kim, Ji Hyun. 2009. "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap." *Computational Statistics and Data Analysis*. https://doi.org/10.1016/j.csda.2009.04.009.

Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177729694.

Lambert, Ben. 2018. *A Students Guide to Bayesian Statistics*.

Liddle, Andrew R. 2007. "Information criteria for astrophysical model selection." https://doi.org/10.1111/j.1745-3933.2007.00306.x.

Liu, Jia, Daniel J. Nordman, and William Q. Meeker. 2016. "The Number of MCMC Draws Needed to Compute Bayesian Credible Bounds." *American Statistician*. https://doi.org/10.1080/00031305.2016.1158738.

Liu, Wei, and Yuhong Yang. 2011. "Parametric or nonparametric? A parametricness index for model selection." *The Annals of Statistics*. https://doi.org/10.1214/11-aos899.

McQuarrie, Allan, Robert Shumway, and Chih Ling Tsai. 1997. "The model selection criterion AICu." *Statistics and Probability Letters*. https://doi.org/10.1016/s0167-7152(96)00192-7.

Montagna, Sara, and Andrea Omicini. 2017. "Agent-based modeling for the self-management of chronic diseases: An exploratory study." *Simulation*. https://doi.org/10.1177/0037549717712605.

Murari, Andrea, Emmanuele Peluso, Francesco Cianfrani, Pasquale Gaudio, and Michele Lungaroni. 2019. "On the use of entropy to improve model selection criteria." *Entropy*. https://doi.org/10.3390/e21040394.

Nguimkeu, Pierre. 2014. "A simple selection test between the Gompertz and Logistic growth models." *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2014.06.017.

Nocedal, J, and S Wright. 2006. *Numerical optimization, series in operations research and financial engineering*.

Park, Barum. 2018. "No Title."

Patil, Anand, David Huard, and Christopher J. Fonnesbeck. 2010. "PyMC: Bayesian Stochastic modelling in Python." *Journal of Statistical Software*. https://doi.org/10.18637/jss.v035.i04.

Paul-Christian Bürkner, Jonah Gabry, Aki Vehtari. 2019. "Approximate leave-future-out cross-validation for Bayesian time series models."

Porciani, C. 2012. "Posterior Probability." In *Obervational Cosmology Lecture 3 - 2012*. https://astro.uni-bonn.de/{~}kbasu/ObsCosmo/Slides2012/Lecture3{_}2012.pdf.

Raftery, ADRIAN E. 1999. "Bayes Factors and BIC." *Sociological Methods & Research*. https://doi.org/10.1177/0049124199027003005.

Ravenzwaaij, Don van, Pete Cassey, and Scott D. Brown. 2018. "A simple introduction to Markov Chain Monte–Carlo sampling." *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-016-1015-8.

Sala-i-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller. 2004. "Determinants of long-term growth: A bayesian averaging of classical estimates (BACE) approach." https://doi.org/10.1257/0002828042002570.

Schulze-hartung, Tim, and Peter Melchior. 2014. "NO Dos and don ' ts of reduced chi-squared." *Astro-Ph*.

ScienceDirect. 2019. "Akaike Information Criterion." https://www.sciencedirect.com/topics/medicine-and-dentistry/akaike-information-criterion.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal*. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shao, Jun. 1997. "An asymptotic theory for linear model selection." *Statistica Sinica*.

Snider and Brimlow, J. N. 2013. "An Introduction to Population Growth." *Nature*

Education Knowledge 4 (4): 3.

Stone, M. 1974. "Cross-validation and multinomial prediction." *Biometrika*. https://doi.org/10.1093/biomet/61.3.509.

Strauss, Sharon Y. 1991. "Indirect effects in community ecology: Their definition, study and importance." https://doi.org/10.1016/0169-5347(91)90023-Q.

Vajpeyi Avi, Smith Rory, Kanner Jonah. 2016. "Use of the Bayes Factor to Improve the Detection of Binary Black Hole Systems."

VanderPlas, Jake. 2014. "Frequentism and Bayesianism: A Python-driven Primer." In *Proceedings of the 13th Python in Science Conference*. https://doi.org/10.25080/majora-14bd3278-00e.

Van Der Ploeg, Tjeerd, Peter C. Austin, and Ewout W. Steyerberg. 2014. "Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints." *BMC Medical Research Methodology*. https://doi.org/10.1186/1471-2288-14-137.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*. https://doi.org/10.1007/s11222-016-9696-4.

Watanabe, Sumio. 2013. "A widely applicable bayesian information criterion." *Journal of Machine Learning Research*.

Wooldridge, Jeffrey M. 2003. "Introductory Econometrics: A Modern Approach." *Economic Analysis*. https://doi.org/10.1198/jasa.2006.s154.

Yang, Yuhong. 2007. "Consistency of cross validation for comparing regression procedures." *Annals of Statistics*. https://doi.org/10.1214/009053607000000514.

Zhang, Yongli, and Yuhong Yang. 2015. "Cross-validation for selecting a model selection procedure." *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2015.02.006.