The AIC-BIC dilemma: An in-depth look

by

Yuqian Song

to obtain the degree of Bachelor of Science at the Delft University of Technology,
to be defended publicly on Friday July 17, 2020 at 12:00 PM.

Student Name:	Yuqian Song	
Student Number:	4630521	
Project Duration:	April 2020 - July 2020	
Thesis committee:	Dr. J. Söehl	TU Delft, supervisor
	Dr. ir. F.H. van der Meulen	TU Delft
	Drs. E.M. van Elderen	TU Delft

An electronic version of this thesis is available at:
`http://repository.tudelft.nl`.

Abstract

In research there is often a need to choose between multiple competing models. Two popular criteria for model selection are the AIC and BIC. The AIC excels in estimating the best model for the unknown data generating process. The BIC on the other hand is consistent in finding the true model. It is clear that for model selection these two information criterion give answers to different selection criteria. The question that arises is whether it is possible to construct a model selection criterion which combines the strengths of both AIC and BIC. In this study we will show that it is impossible to construct a model selection criterion which shares the above mentioned two strenghts by revisiting the proof of (Yang, 2005) : That is, any consistent model selection criterion must be sub-optimal in the minimax convergence rate for regression estimation compared to the AIC.

Contents

1 Introduction

Statistics is used in all branches of science and industries. An important task in statistics is modelling data, it allows us to study, explain and predict data. A question that frequently comes up during modelling is: "Which model represents the process which generated the observed data best?". Statistics provides a tool for model selection in the form of information criterion. Information criteria are used to measure the "quality" of a model by taking the goodness of fit and the complexity of the model into consideration.

There are many information criteria to choose from, with two of the most well-known information criteria are the AIC (Aikaike Information Criterion) (Akaike, 1973) and the BIC (Bayesian Information Criterion), (Schwarz, 1978). While we will not go into detail about the process of selecting an information criterion we will stress that it is important to ask ourselves what the goals of the study are, which kind of model we are building, and what the considerations, circumstances, sample size and assumptions of the study are.

In this thesis we will have an in-depth look at the AIC and the BIC. With the main focus being on the properties of these criteria, in particular the minimax-rate optimality property of the AIC and the consistency property of the BIC. And the proof that for any consistent model selection criterion, it must be sub-optimal in the minimax convergence rate for regression estimation compared to the AIC.

2 Background material

The goal of this section is to inform and/or refresh the reader about a few concepts in statistics as well as to fix the notation. This will make it easier for the reader to follow the content in this thesis.

2.1 Kullback-Leibler information

As stated in the introduction, there are many information criteria. One of these is the Kullback-Leibler information (KL information) (Kullback & Leibler, 1951). The KL information measures the information that is lost when we use function g to approximate function f and is defined as:

$$I(f,g) = \int f(x) log\left(\frac{f(x)}{g(x)}\right) dx \tag{1}$$

Where log denotes the natural logarithm.

We use function f to represent the reality or true model, while function g is used to represent the approximating model. We aim to find the approximating model g which loses the least amount of information and thus minimize $I(f,g)$ over g.

The KL information is of importance for this study because it is one of the main components behind the theory of the AIC.

2.2 Regression analysis

Regression analysis is a method which estimates the relationship between a response variable and one or more independent variables. The aim of regression analysis is to construct a model which describes or explains the relationship between variables (we refer to (Seber & Lee, 2012) for an in-depth introduction to this subject).

Let us denote the response variable as Y and the set of independent variables as $X_1, X_2, ..., X_p$, where p is the number of independent variables. Then a regression model which represents the true relationship between the response variable Y and the independent variables X_i may be formulated as:

$$Y = f(X_1, X_2, ..., X_p) + \epsilon \tag{2}$$

Where $f(X_1, X_2, ..., X_p)$ is the regression function which describes the relationship between the response variable Y and the independent variables X_i. And ϵ denotes the random error of the approximation. These random errors are assumed to be normally distributed with mean 0 and variance σ^2, $\epsilon \sim N(0, \sigma^2)$.

For the purpose of the proof in the later parts of this study we shall use the same regression model as in (Yang, 2005):

$$Y = f_k(x, \theta_k) + \epsilon \tag{3}$$

Where $\mathcal{F}_k = \{f_k(x, \theta_k), \theta_k \in \Theta_k\}$ is a linear family of regression functions for all k. And θ_k is the parameter of a finite dimension m_k.

When we use an approximating model to represent the reality or true model we have to decide on how many parameters to use. This choice can cause two problems. If too few parameters are used then it will cause a large approximation error since a function cannot be approximated well by a projection onto a small dimensional subspace. This also causes underfitting the observed data. Which implies that the resulting approximation model does not reflect the observed data. However, the usage of too many parameters causes a large stochastic error since there is a stochastic error in each of the many estimated parameters. It also causes overfitting of the observed data. Which results into the approximation model being too dependent on the observed data. This is undesirable as we are not trying to find a model that fits the observed data exactly, we are trying to find a model that represents the process or distribution which generated the observed data. Another undesirable effect of overfitting is the inability to fit or predict new data. The issue of overfitting and underfitting will also appear in Section 5.1.

2.3 Likelihood

Definition 2.1. *Let the random variables $Y_1, Y_2, \ldots, Y_n$ of sample size n have a joint distribution function $f(Y_1, Y_2, \ldots, Y_n|\theta)$ which depends on a set of parameters θ. Let $y_1, y_2, \ldots, y_n$ be the observed values corresponding to the afore mentioned random variables. Then the likelihood function is defined as $\mathcal{L}(\theta) = \mathcal{L}(\theta|y_1, y_2, \ldots, y_n) = f(y_1, y_2, \ldots, y_n|\theta)$.*

The likelihood function represents the probability that a specific set of parameters would yield the observed data. In maximuml likelihood estimation we want to choose the set of parameters which results in the highest probability of the observed data. The AIC and BIC both use maximum likelihood estimation as a means to calculate their values. The variant the AIC and BIC use is the log-likelihood, which simply $log(\mathcal{L}(\theta|y_1, y_2, \ldots, y_n))$. The usage of the natural logarithm is justified by its monotonically increasing property which ensures that the maximum value of the log of the probability occurs at the same point as the original probability function. While also being easier to work with mathematically with regards to differentiation and expressions.

2.4 Loss function and risk function

In the context of statistics the loss function is used to quantify the difference between the estimated parameter $\hat{\theta}$ and the true parameter θ_0, or that of a

predictor $\hat{y_i}$ and the true outcome y_i. If we define the estimation error as $\hat{\theta} - \theta_0$, then the loss function is a function which maps the estimation error to the set of real numbers. The choice of the loss function depends on the considerations of the study. It is important to note that the choice of the loss function impacts how well an information criterion such as the AIC and BIC performs as argued by (Vrieze, 2012).

An example of the loss function would be the sum of squared error, $SSE = (\hat{\theta} - \theta_0)^2$. The expected value of the loss function is called the risk function or simply risk. If we take the expected value of the squared estimation error then we would get the mean squared error, $R = MSE = E_{\theta_0}((\hat{\theta} - \theta_0)^2)$. Similarly the risk function of the predictor is, $R = MSE = \frac{1}{n}\sum_{i=i}^{n}(\hat{y_i} - y_i)^2$ for n samples. One of the aims of the AIC is to find a model which minimizes these functions or in other words, the estimation or prediction error.

2.5 Bayes factor

In Bayesian statistics the prior probability of a random event is the unconditional probability assigned to the random event before taking any evidence into account. In contrast to the prior probability, the posterior probability of a random event is the conditional probability assigned to the random event after the relevant evidence has been taken into account. For example, assume that a professor wants to find out how much percent of his students will attend to every lecture during this semester. Then he could look at the historical data from surveys which estimates it to be 70%. This is the prior probability. Assume that he hands out surveys in the middle of the semester which only a part of the students answer. He then takes the new data into consideration with the given old data and comes in to an estimate of 75%. This is the posterior probability. Let $Pr(M_i)$ be the prior probability of model i, $Pr(y)$ be the prior probability of observed data y. Then we can define $Pr(y|M_i)$ to be the probability of data y being produced under model M_i. We can define the posterior probability $Pr(M_i|y)$ using the Bayes theorem as

$$Pr(M_i|y) = \frac{P(y|M_i) \cdot Pr(M_i)}{Pr(y)}$$

One way to choose between two models M_i and M_j under the observed data y in a model selection problem is to look at the ratio of the the posterior probabilities of one model over another. If we assume that prior probabilities of both models are equal, $Pr(M_i) = Pr(M_j)$ then this ratio is called the Bayes factor B_{ij}. The Bayes factor is defined as

$$
\begin{aligned}
B_{ij} &= \frac{Pr(y|M_i)}{Pr(y|M_j)} \\
&= \frac{\frac{Pr(M_i|y)Pr(y)}{Pr(M_i)}}{\frac{Pr(M_j|y)Pr(y)}{Pr(M_j)}} \\
&= \frac{Pr(M_i|y)}{Pr(M_j|y)}
\end{aligned}
$$

The Bayes factor can be interpreted as the ratio quantifying the relative probability of some observed data for one model over another. If value of $B_{ij} > 1$ then M_i is more strongly supported by the observed data y than M_j. Therefore, we should choose model M_i. If $B_{ij} < 1$, then we should choose model M_j as M_j is more supported by the data y than M_i. Table 1 below will show the precise interpretation for the Bayes factor for different values for a model M_1 and M_2. It is a modified version of the interpretation of (Jeffreys, 1961) made by (Lee & Wagenmakers, 2014).

Bayes factor B_{12}	interpretation
> 100	Extreme evidence for M_1
$30 - 100$	Very strong evidence for M_1
$10 - 30$	Strong evidence for M_1
$3 - 10$	Moderate evidence for M_1
$1 - 3$	Anecdotal evidence for M_1
1	No evidence for M_1
$\frac{1}{3} - 1$	Anecdotal evidence for M_2
$\frac{1}{10} - \frac{1}{3}$	Moderate evidence for M_2
$\frac{1}{30} - \frac{1}{10}$	Strong evidence for M_2
$\frac{1}{100} - \frac{1}{30}$	Very strong evidence for M_2
$< \frac{1}{100}$	Extreme evidence for M_2

Table 1: Interpretation of the Bayes factor values for two models M_1 and M_2.

The Bayes factor is treated in this section because to compare two models with equal prior probability, one can use the Bayes factor to measure which model is more supported by the data. And the BIC has its roots in posterior probabilities.

3 AIC

The Akaike Information Criterion, or commonly abbreviated as the AIC was introduced by Akaike in 1973 when he found a relation between the relative expected Kullback-Leibler information (KL information) and the maximized log-likelihood. The AIC is used in model selection as a measure for how good an estimating model out of a set of candidate estimating models is for representing the process which generated the observed data.
The formula of the AIC given by (Burnham & Anderson, 2002) is:

$$AIC = -2log(\mathcal{L}(\hat{\theta}|y)) + 2K \tag{4}$$

Where $log(\mathcal{L}(\hat{\theta}|y)$ is the numerical value of the log-likelihood at its maximum point. And K as the the number of estimated parameters.

We shall not give an derivation of the AIC, however, a derivation from the aspect of the KL information can be found in Chapter 2.2 of (Burnham & Anderson, 2002). The AIC does not assume that the "true model" is in the set of the candidate estimating models. This stems from the belief that a model cannot reflect reality, which is also why the KL information cannot be used directly as an information criterion, because that would imply that we know the truth (see (Burnham & Anderson, 2004)). The AIC was derived as an estimate of the KL information, it gives an estimate of the expected, relative distance between a candidate estimating model g and the unknown true model f. Let g_1 and g_2 be estimating models for the true model f. We do not know the absolute distance between an estimating model g_1 and the true model f, however, we can compare its expected distance to the expected distance of g_2 and f. The smaller the expected relative distance, the less information is lost in the KL information and the more this estimating model approaches the true model. Therefore, the estimating model which has the lowest AIC value is the preferred model for representing the unknown true model.

A point of consideration is that the estimate of the expected, relative distance mentioned before only holds asymptotically. For sample sizes which are too small the AIC tends to overfit and a stronger penalty term is recommended. As the sample size grows the risk of overfitting and underfitting diminishes for the AIC. In general it is recommended to use the AIC if there is an emphasis on avoiding underfitting (we refer to (Bozdogan, 1987) for an in-depth analysis of how sample size affects overfitting and underfitting for the AIC). We also refer to (Hurvich & Tsai, 1989) as it goes into on the proposed AIC_c criterion which addresses this problem. The lower case "c" in the AIC_c indicates that this is the version of the AIC which counterbalances the overfitting for small sample sizes with additional penalty.

3.1 Minimax-rate optimality

The AIC has a minimax property, it minimizes the maximum possible risk. This means that the AIC will perform the best in the worst possible situation that is allowed under the set of conditions. (Yang, 2005) gives an example on AIC yielding the minimax-rate optimal estimators of the regression function under a squared-error-type loss. We shall not go into the details of the proof but attempt to summarize the most important steps. For clarity we will take over his notation and definitions.

First, find a loss function to use for estimating the regression function f, as well as the according risk function. In this example we assume that it is the mean squared error (MSE). Let δ be the model selection criterion that selects model $\hat{k}$, $\hat{\theta}_{\hat{k}}$ be the least squares estimator of the parameter in the selected model, then $MSE(f_{\hat{k}}) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f_{\hat{k}}(x_i, \hat{\theta}_{\hat{k}}))^2$. The according risk function of the MSE is: $R(f, \delta, n) = \frac{1}{n}\sum_{i=1}^{n} E(f(x_i) - f_{\hat{k}}(x_i, \hat{\theta}_{\hat{k}}))^2$.

Second, define minimax-rate optimality:

Definition 3.1. *A model selection criterion δ is said to be minimax-rate optimal over a class of regression functions $\mathcal{F}$ if $\sup_{f\in\mathcal{F}} R(f, \delta, n)$ converges at the same rate as $\inf_{\hat{f}} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} E(f(x_i) - f_{\hat{k}}(x_i, \hat{\theta}_{\hat{k}}))^2$, where $\inf_{\hat{f}}$ is over all estimators based on the observations of $Y_1, \ldots, Y_n$.*

Third, proving that the rate of convergence is equal. (Yang, 2005) uses the result of his past study, (Yang, 1999), to formulate a proposition as well as a corollary.
We first need introduce the notation used in (Yang, 2005). Let Γ be the set of candidate models, which size can be either finite or countably infinite. We denote N_m as the number of models that have the same dimension m in the set of candidate models Γ. And m_k as the dimension of model k. We assume that there exists a constant $c > 0$ such that $N_m \leq e^{cm}$. We denote the estimator of f based on the outcome of AIC as δ_{AIC}. We denote M_k as the projection matrix of model k and let r_k denote the rank of projection matrix M_k. We note that $r_k \leq m_k$ as the rank of the projection matrix of model k cannot be larger than the dimension of model k itself in this case. For simplicity we assume that $\sigma^2 = 1$.

Proposition 3.1. *(Proposition 1 in (Yang, 2005)) There exists a constant $C > 0$ depending on c such that for every regression function f, we have*

$$R(f, \delta_{AIC}, n) \leq C \inf_{k\in\Gamma}\left(\frac{||f - M_k f||_n^2}{n} + \frac{r_k}{n}\right).$$

Where $||\cdot||$ denotes the Euclidean norm of an n-dimensional vector.
This proposition states that the risk function has to be smaller or equal than a constant C multiplied by the infimum over all possible models of $\frac{||f-M_k f||_n^2}{n}$,

which stands for the approximation error of model k, plus $\frac{r_k}{n}$, which stands for the stochastic error of model k.

Corollary 3.1. *(Corollary 1 in (Yang, 2005)) Suppose that model $k^* \in \Gamma$ is the true model. Then*

$$sup_{f \in \mathcal{F}_{k^*}} R(f, \delta_{AIC}, n) \leq \frac{C m_{k^*}}{n}.$$

Corollary 3.1 implies that the worst case risk of δ_{AIC} under the true model k^* decays at the rate of $\frac{1}{n}$. This means that the estimator δ_{AIC} is minimax-rate optimal whenever the true model is in the set of candidate models Γ. When the true regression function is infinite-dimensional relative to Γ, then $\frac{||f - M_k f||_n^2}{n}$ is nonzero for all models k. We can also look at smoothness classes such as Sobolev balls. For these classes we can choose an appropriate set of candidate models such as polynomial splines. The choice was made in such a way that it results into $\inf_{k \in \Gamma} \left(\frac{||f - M_k f||_n^2}{n} + \frac{r_k}{n} \right)$ being of the same order as the minimax-rate of convergence. Hence the δ_{AIC} is minimax-rate optimal over smoothness classes without the need to know the true smoothness order. Therefore, the δ_{AIC} is minimax-rate optimal with a convergence rate of $\frac{1}{n}$ when one of the candidate model holds. It is also minimax-rate optimal when the true regression function is of infinite dimension.

4 BIC

The Bayesian Information Criterion, or abbreviated as the BIC was derived by Schwartz in 1978 when he proposed a Bayesian argument for adopting Akaike's work. The formula given by (Burnham & Anderson, 2002) is:

$$BIC = -2log(\mathcal{L}(\hat{\theta}|y)) + K \cdot log(n). \tag{5}$$

Where the n denotes the sample size.

The BIC unlike the AIC is not an estimate for the KL distance, it is an estimate of the Bayes factor. To compare two models M_i and M_j, one can use the BIC to estimate the Bayes factor B_{ij} by

$$B_{ij} \approx \exp(-\frac{1}{2}BIC_i + \frac{1}{2}BIC_j). \tag{6}$$

The candidate model with the smallest BIC value, is the candidate model with the highest Bayesian posterior probability. And therefore the "best" performing candidate model is the model with the lowest BIC value.

4.1 Consistency

The BIC is part of a class of criteria that uses "consistency" or "dimension-consistency" as an approach to model selection. The main assumption the BIC makes is that a true model exists which represents the reality fully. And when the true model is in the set of candidate models, then the probability of choosing the true model goes to one as sample size increases. The formal definition is:

Definition 4.1. *Assume that the "true model" exists and is in the set of candidate models. Furthermore, assume that the dimension of the true model remains fixed as the sample size grows, and the number of parameters in the true model is finite. Then an information criterion is consistent if the probability to choose "the true model" approaches 1 as the sample size increases to infinity.*

There is much debate and confusion about the existence of a true model. Information theorists do not believe that a model exists which can fully represent reality. It even creates a paradox, because if the true model existed, and is assumed to be in the set of candidate models, then would you not already know the true model?

It is worth mentioning that while the AIC does not have the consistency property it might select a better performing model compared to the BIC. This is due to its minimax-rate optimality which minimzes the loss function as discussed in (Vrieze, 2012).

5 AIC versus BIC

It should be clear by now that the inherent goals of the AIC and BIC are different. The AIC attempts to find the model which has the best predictive power to predict future observations. The method is preferred in cases where the true model is too complex as it contains an inifinite amount of parameters. While the BIC attempts to find the true model which generated the observed data. This causes the BIC to be preferred in cases where the true model exists and is in the candidate set of models. The BIC is not minimax-rate optimal, because no consistent model selection criteria can be minimax-rate optimal (Yang, 2005). The AIC is not consistent as it has a nonzero probability of failing to choose the true model as the smaple size goes to infinity. This is true even under the assumption that the true model is in the candidate set of models.

5.1 Underfitting and overfitting

The formula of the AIC can be separated into two parts. The first part is $-2log(\mathcal{L}(\hat{\theta}|y))$. This part contains the numerical value of the log-likelihood at its maximum point. Due to the term being multiplied with a "-2" it rewards goodness of fit as it lowers the AIC value. However, this causes the risk of overfitting. The second part of the formula is $2K$, which penalizes overfitting as it discourages the usage of more estimated parameters. As a result, the AIC manages to avoid both overfitting as well as underfitting.

We observe that the formula for the BIC is very similar to that of the AIC, The only difference being the penalty term against overfitting, $log(n)$. This fuction becomes bigger than 2 when the sample size reaches 8. Therefore, the BIC penalizes overfitting more for big sample sizes compared to the AIC. This implies that the BIC will prefer more "simple" or "smaller" models (models with less parameters) compared to the AIC for sample sizes bigger than 8. As an additional result the models chosen by the AIC and BIC will be remarkably different for large sample sizes.

5.2 Numerical example

This section will feature a simple example to illustrate the differences between the AIC and BIC in a simulated model selection problem. We will compare the AIC and BIC values for different sample sizes as well as looking at the results of the cross validation. The cross validation method is a model selection method used to measure the predictive ability of models. It is a method used as a measure against overfitting in a predictive model. Cross validation partitions the data in a predetermined amount of partitions, runs the analysis on each partition, and then takes the average of the overall error estimate, which is in our case the squared error. In cross validation the amount of partitions is denoted by the letter k. We also say that we apply k-fold cross validation for the amount of partitions we divide our data into. We refer to (Shao, 1993) for

further reading about the cross validation method.

5.2.1 Methodology

Consider the following model:

$$y = \alpha x + \beta x^2 + \gamma x^3 + \epsilon$$

where y is the response variable, x the predictor and ϵ the random error. This will be our true model which generates the data.

- The values for $[\alpha, \beta, \gamma]$ were chosen arbitrarily to be $[-2, 5, 7]$.
- The predictor "x" is assumed to be uniformly distributed between -2 and 2.
- The random error "ϵ" is assumed to be normally distributed with mean 0 and variance 1.
- The candidate models will consist out of five polynomials from degree one to five.
- The candidate models will be generated using the R function "lm" with argument "poly" to fit the generated data.
- The cross validation will be five fold.
- Usage of the free statistical software R.
- The code is used in the simulations is a modified version from (Petrkeil, 2013).
- I will use a fixed seed for the simulations, "4630521" which is my student number. A fixed seed is chosen so we can observe how the models perform each time new data gets added to the old data set. This is also such that the simulations are replicable. I will do an additional Monte Carlo simulation without a seed.

I will denote the sample size by the letter "n" and the amount of simulations with the letter "m". The goal of the simulations is to let the AIC and BIC choose between the fitted set of candidate models which consists out of polynomials of degrees one to five and choose the "best" model out of them.

5.2.2 Simulation of model selection scores

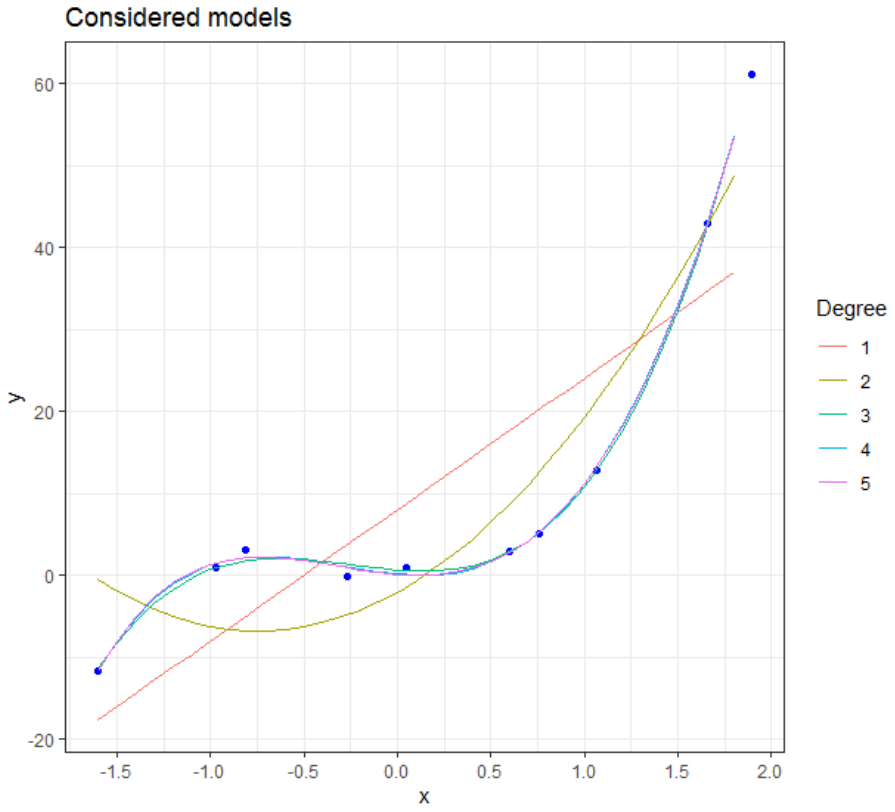

Figure 1: The candidate models are represented by the lines in different colours and the observations generated by the true model are represented by the blue dots. The sample size is n = 10.

From Figure 1 one can observe the generated data points of the true model and the fits of the different degree polynomials. It becomes immediately clear that the polynomials of degrees one and two are underfitting. The higher degree polynomials will cause overfitting, as they are be able to fit the data more precisely due to their higher degree.
From the above Table 2 we observe that both the AIC and BIC choose the polynomial of degree four as the "best" model. This coincides with the cross validation method designating the polynomial of degree four having the best predictive ability. A possible explanation for why the BIC has failed to choose the true model here could be that the sample size is too small. If we estimate the Bayes factor for M_3 and M_4 which are the candidate models of degree three and four , respectively, the value is $B_{34} \approx 0.47$. According to Table 1 this corresponds to an anecdotal evidence for M_4. Therefore, it seems that the BIC

degree	AIC	BIC	cross validation
1	84.00	84.90	2456.59
2	76.03	77.25	2310.24
3	31.15	32.66	33.53
4	29.34	31.16	12.76
5	31.15	33.27	28.31

Table 2: The values of the AIC, BIC and cross validation for different polynomial degrees rounded to two decimals for sample size n=10.

is not too inaccurate. We will look at the results again for a sample size of n = 50.

degree	AIC	BIC	cross validation
1	378.49	384.22	5620.80
2	364.05	371.70	4381.31
3	144.98	154.54	54.09
4	137.79	149.25	45.75
5	138.82	152.21	46.40

Table 3: The values of the AIC, BIC and cross validation for different polynomial degrees rounded to two decimals for sample size n=50.

We observe a similar result as before. Both AIC and BIC agree that the "best" model is the candidate model with polynomial degree four. And the cross validation method chose the candidate model with polynomial degree four as the best predictive model again. For the next simulation we will increase the sample size to n = 250.

degree	AIC	BIC	cross validation
1	1844.52	1855.08	23374.19
2	1778.00	1792.09	18039.14
3	713.78	731.39	252.01
4	713.94	735.07	251.90
5	715.72	740.37	253.39

Table 4: The values of the AIC, BIC and cross validation for different polynomial degrees rounded to two decimals for sample size n=250.

The BIC has chosen the candidate model with degree three which is the polynomial degree of the true model. It is clear that the BIC prefers the candidate model with degree three over degree two. The estimated Bayes factor for candidate models of degree three and four is $B_{34} = 6.30$. This result implies according to Table 1 that there is moderate evidence for the candidate model with degree three. This observation demonstrates the consistency property of

the BIC partially, because the BIC has chosen the correct model as the sample size increased. The AIC has also chosen the polynomial of degree three, however, the difference between the AIC values of the model with polynomial degrees three and four differ very little. The cross validation method once again has chosen the model with polynomial degree four as the model with the best predictive capability. This partially illustrates the AIC's property to prefer models for prediction throughout all three simulations.

A possible explanation for why all three methods preferred polynomials of degree three or higher is because the true model itself is a polynomial of degree three. The polynomials of degree one and two could not fit the data generated by a polynomial of degree three adequately and have been heavily penalized for underfitting. This issue appears consistently for all three sample sizes as we can observe that the values of the criteria for those two candidate models are much higher. In contrast, if a third, fourth or fifth degree polynomial are used to approximate a third degree polynomial then they would all be very precise. We know that higher degree polynomials can approximate lower degree polynomials better compared to the other way around. This resulted into the overfitting to be of a much lower degree compared to the underfitting of the lower degree candidate models. Hence it would not be penalized as much which explains why the criterion scores for these three higher degree candidate models are so close to each other.

5.2.3 Monte Carlo simulations

I will now perform a second simulation where I illustrate the consistency property of the BIC more clearly. The simulation will calculate the probability of the AIC and BIC selecting the candidate model with degree three which is the same degree as the true model for different sample sizes (n) and different numbers of simulations (m). I've removed the code for calculating the cross validation in the Monte Carlo simulations as the computation time is too long when I increase the sample size or the amount of simulations.

AIC	n = 10	n = 50	n = 250	n = 1000
m = 25	36%	80%	84%	76%
m = 100	47%	76%	80%	73%
m = 250	47.2%	75.6%	80.4%	78.5%
m = 500	46.2%	75.6%	79.6%	78.8%
m = 1000	48.6%	74.9%	79.7%	77.5%

Table 5: Probability of AIC selecting the candidate model with degree three.

BIC	n = 10	n = 50	n = 250	n = 1000
m = 25	44%	88%	100	100%
m = 100	51%	93%	98%	98%
m = 250	51.6%	93.6%	98.8%	98%
m = 500	51.2%	92.6%	98.6%	98.2%
m = 1000	52.9%	92.1%	98.4%	98.5%

Table 6: Probability of BIC selecting the candidate model with degree three.

From Tables 5 and 6 we observe multiple things. The first is that in general the sample size affects whether the candidate model with the true degree was chosen for both AIC and BIC. The second is that the BIC is more consistent in choosing the candidate model with the degree compared to the AIC. As seen before in Section 5.2.2 this is caused by the AIC preferring models with stronger explanatory ability compared to the BIC. Lastly, we observe that the probability to select the candidate model with the true degree approaches one as the sample size grows for the BIC. This also shows of the consistency property of the BIC to a degree.

We will perform Monte Carlo simulations without using a fixed seed. This will make it so that each simulation is randomized and simulations with bigger sample sizes and amount of simulations no longer have any data related with the simulations with smaller sample sizes and amount of simulations.

AIC	n = 10	n = 50	n = 250	n = 1000
m = 25	60%	80%	68%	84%
m = 100	49%	77%	77%	80%
m = 250	56%	74.4%	77.6%	79.2%
m = 500	45.2%	76.4%	79.4%	77.4%
m = 1000	47.4%	76%	76.5%	78.2%

Table 7: Probability of AIC selecting the candidate model with degree three. (No seed used).

BIC	n = 10	n = 50	n = 250	n = 1000
m = 25	60%	100%	92%	100%
m = 100	55%	93%	97%	99%
m = 250	60%	93.2%	98%	100%
m = 500	50%	94%	98%	99.6%
m = 1000	52.1%	93.1%	97.1%	98.8%

Table 8: Probability of BIC selecting the candidate model with degree three. (No seed used.)

From Tables 7 and 8 we can observe roughly the same results as before. The probability to choose the candidate model with degree three appears to go to one as the sample size increases for the BIC. The AIC also has an increased chance to select the candidate model with the true degree. However, for sample sizes 50, 250 and 1000 the probability of selecting the true model is hovering a round 80%. We cannot conclude that the AIC has reached a bottleneck for consistency at around 80% because we haven't done enough simulations to confirm this numerically. However, we can attribute this result to AIC's property to prefer candidate models with better predictive power. It still holds that the BIC is more consistent compared to the AIC for these simulations. And that sample size affects the probability to select the true model for both model selection criteria. Compared to the simulations before which used a seed the probabilities for sample size $n = 10$ for both AIC and BIC are on average higher. The explanation for this is because the random number generator is no longer restricted by a seed and therefore it would result into different percentages. We do not observe other significant differences between the two Monte Carlo simulations.

5.2.4 Discussion

It should be noted that the simulations performed in this section were not in-depth. A simulation for the expected loss function could be added for each candidate model for an inspection of the minimax-rate properties for both AIC and BIC. It would also be worthwhile to find out another way to simulate the cross validation score for the Monte Carlo simulations in a way that it is feasible time wise. Furthermore the simulations were performed using a single response variable. This could be expanded with multiple response variables. There are many more simulations that can be done to get a better understanding of how the AIC and BIC perform under application for model selection. However, that is outside the scope of this thesis.

6 Combining the strength of AIC and BIC

The topic of interest in this study is whether we can create a method which has both the minimax-rate optimality of the AIC and the consistency of the BIC. Hypothetically this would lead to a method of model selection which would result in better chosen models.

6.1 Proof that it is not possible to combine the strengths

Unfortunately under the standard definition of the minimax-rate optimality it is not possible for a method to be both consistent and minimax-rate optimality. I will go through the proof given by (Yang, 2005) for this claim.

6.1.1 Assumptions

The structure of the proof begins with making some assumptions.

Assumption 1 (Assumption 1 in (Yang, 2005))**.** *There exists two models* $k_1, k_2 \in \Gamma$ *such that*

1. $\mathcal{F}_{k_1} = \{f_{k_1}(x, \theta_{k_1}) : \theta_{k_1} \in \Theta_{k_1}\}$ *is a sub-linear space of* $\mathcal{F}_{k_2} = \{f_{k_2}(x, \theta_{k_2}) : \theta_{k_2} \in \Theta_{k_2}\}$.
2. *There exists a function* $\phi(x)$ *in* $\mathcal{F}_{k_2}$ *orthogonal to* $\mathcal{F}_{k_1}$ *(at the design points) with* $\frac{1}{n}\sum_{i=1}^{n} \phi^2(x_i)$ *being bounded between two positive constants for a large enough* n.
3. *There exists a function* $f_0 \in \mathcal{F}_{k_1}$ *such that* f_0 *is not in any family* $\mathcal{F}_k (k \in \Gamma)$ *that does not contain* $\mathcal{F}_{k_1}$.

The term design point in part two of Assumption 1 means the points on which you observe something. An example of a design point is when you split a function $f(x)$ on the interval $x \in [0, 1]$ into 4 equidistant intervals. And only look at the values of $f(x)$ on the design points $x \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. The second part of the assumption is satisfied for a reasonable design. The third part of the assumption holds when there are a finite number of models or a countable list of (nested) models. It is said that two models are nested if the parameters in one of the two models are a subset of the other model.

6.1.2 Theorem

Theorem 6.1 (Theorem 1 in (Yang, 2005))**.** *Under Assumption 1, if any model selection method* δ *is consistent in model selection, then we must have*

$$n \sup_{f \in \mathcal{F}_{k_2}} R(f, \delta, n) \to \infty. \tag{7}$$

Theorem 6.1 basically says that in a parametric case, a case where we estimate the function f to be finite dimensional, consistency in model selection comes at a

high price for estimation of the regression function. The risk function multiplied by a constant n goes to infinity, this is the high cost. In the parametric case, we can typically achieve a rate of $\frac{1}{n}$ for minimax-rate optimal model selection methods. Since this term goes to infinity, it cannot be minimax-rate optimal.

6.1.3 Proof in a special case

The proof by (Yang, 2005) reduces the problem to a hypothesis testing problem. If we do this then we can apply hypothesis testing theory. We will first prove Theorem 6.1 for a simple case and afterwards for a more general case.
We first define the null model M_0 to be:

$$Y_i = \alpha + \epsilon_i, \qquad i = 1, 2, ..., n \tag{8}$$

Afterwards we define the simple linear model M_1 to be:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, 2, ..., n \tag{9}$$

where x is an one-dimensional design variable and ϵ is the error. We can assume that the design has the property $\bar{x}_n = 0$ without loss of generality. Furthermore, similarly to the second part of Assumption 1 we have assume that $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ is bounded between two positive constants for all n.
Consider a consistent model selection criterion δ and let A_n be the event that M_1 is chosen with n observations. Then the corresponding estimator of $f(x_{M_0})$ is

$$\hat{f}(x_{M_0}) = \hat{\alpha} + \hat{\beta} x_{M_0} I_{A_n}$$

where I_{A_n} is the indicator function of event A_n which outputs 1 when event A_n happens and 0 otherwise.
The risk at x_{M_0} under squared error loss is:

$$\frac{\sigma^2}{n} + x_{M_0}^2 E\big(\hat{\beta} I_{A_n} - \beta\big)^2 + 2x_{M_0} E(\hat{\alpha} - \alpha)(\hat{\beta} I_{A_n} - \beta).$$

Where σ is the standard deviation and σ^2 is the variance. These terms appear in the expression because the calculation involves taking the expected value of the function $\hat{f}(x_{M_0})$ substituted into the formula for the squared error loss. We want to take the expected value of this equation, the first part $\frac{\sigma^2}{n}$ is just a constant so we can use linearity of the expectation to take it out of the expectation. In the middle section we have the expected value $E\big(\hat{\beta} I_{A_n} - \beta\big)^2$, we know that the expected value of the expected value is itself. The last part of the equation $2x_{M_0} E(\hat{\alpha} - \alpha)(\hat{\beta} I_{A_n} - \beta)$ disappears because of our assumption that $\bar{x}_n = 0$.
The mean average squared error is therefore:

$$R(f, \delta, n) = \frac{\sigma^2}{n} + \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) E\big(\hat{\beta} I_{A_n} - \beta\big)^2.$$

Now that we have an expression for $R(f, \delta, n)$ we can adapt it to the expression in Theorem 6.1 and prove the theorem.
In particular, we show that for any consistent model selection method, for each $c > 0$, we must have

$$\begin{aligned}
& n \sup_{|\beta| \le c} E_\beta\big(\hat{\beta} I_{A_n} - \beta\big)^2 \to \infty \\
& \iff \sup_{|\beta| \le c} E_\beta\big(\sqrt{n}\hat{\beta} I_{A_n} - \sqrt{n}\beta\big)^2 \to \infty \\
& \iff \sup_{|\beta| \le c} E_\beta\big(\sqrt{n}(\hat{\beta} - \beta) I_{A_n} - \sqrt{n}\beta I_{A_n^c}\big)^2 \to \infty \\
& \iff \sup_{|\beta| \le c} \big(E_\beta n(\hat{\beta} - \beta)^2 I_{A_n} + n\beta^2 P_\beta(A_n^c)\big) \to \infty.
\end{aligned}$$

Where A_n^c is the complement of A_n. We notice that the expression in the last equivalency goes to infinity if the second term, $n\beta^2 P_\beta(A_n^c)$, goes to infinity. Therefore, it suffices to prove that for each $c > 0$

$$\sup_{|\beta| \le c} n\beta^2 P_\beta(A_n^c) \to \infty. \tag{10}$$

Now we set up a testing problem as follows. The true model which generated the observations is:

$$Y_i = \beta x_i + \epsilon_i, \qquad i = 1, 2, ..., n \tag{11}$$

where ϵ_i are independent standard normally distributed errors. This model is a sub family of model (9) with $\alpha = 0$ and $\sigma = 1$. We will now construct a hypothesis testing problem. Let the null hypothesis be: $H_0 : \beta = 0$ and the alternative hypothesis be $H_1 : \beta > 0$. For this hypothesis test we take A_n as the rejection region and δ as the testing rule with probability of type I error approaching zero. Using the Neyman-Pearson Lemma (Neyman & Pearson, 1933) it can be shown that for any test with probability of type I error going to zero, it has to hold that $\sup_{|\beta| \le c} n\beta^2 P_\beta(\tilde{A}_n^c) \to \infty$. With $\tilde{A}$ being the rejection region of the test. Let us denote the joint probability density function of $Y_1, ..., Y_n$ of model (11) to be $f(y_1, ..., y_n; \beta)$. Then for $\beta_1 > \beta_0 \ge 0$,

$$\begin{aligned}
\frac{f(y_1, ..., y_n; \beta_1)}{f(y_1, ..., y_n; \beta_0)} &= \exp\left(\frac{1}{2}\sum_{n=1}^{n}\Big((y_i - \beta_0 x_i)^2 - (y_i - \beta_1 x_i)^2\Big)\right) \\
&= \exp\left((\beta_1 - \beta_0)\sum_{n=1}^{n} x_i y_i + \frac{1}{2}(\beta_0^2 - \beta_1^2)\sum_{i=1}^{n} x_i^2\right).
\end{aligned}$$

We observe that this ratio is non decreasing for the statistic $\sum_{i=1}^{n} x_i Y_i$. This family of models therefore has the property that is also known as a monotone likelihood ratio. Using the Karlin-Rubin Theorem that an uniformly most powerful (UMP) test exists. Which rejects H_0 when $\sum_{i=1}^{n} x_i Y_i > C$ where C is a

constant which we choose to be d_n. This d_n has been chosen in such a way that $P_{\beta=0}(\sum_{i=1}^{n} x_i Y_i \geq d_n) = P_{\beta=0}(A_n)$. Let $A_{n,*}$ be the event $\{\sum_{i=1}^{n} x_i Y_i \geq d_n\}$. By the UMP property of $A_{n,*}$ we have that for all $\beta > 0$

$$P_\beta(A_{n,*}) \geq P_\beta(A_n).$$

As a result we also have that:

$$\sup_{|\beta|\leq c} n\beta^2 P_\beta(A^c_{n,*}) \geq \sup_{|\beta|\leq c} n\beta^2 P_\beta(A^c_n).$$

We know that $\sum_{i=1}^{n} x_i Y_i$ has a normal distribution, therefore for $\beta = 0$:

$$\begin{aligned} P_{\beta=0}(A_{n,*}) &= P_{\beta=0}(\sum_{i=1}^{n} x_i Y_i \geq d_n) \\ &= P\bigg(N(0,1) \geq \frac{d_n}{\sqrt{\sum x_i^2}}\bigg) \end{aligned}$$

and for $\beta > 0$

$$\begin{aligned} P_\beta(A_{n,*}) &= P_\beta(\sum_{i=1}^{n} x_i Y_i < d_n) \\ &= P\bigg(N(0,1) < \frac{d_n - \beta \sum x_i^2}{\sqrt{\sum x_i^2}}\bigg). \end{aligned}$$

With our earlier choice of d_n such that $P_{\beta=0}(\sum_{i=1}^{n} x_i Y_i \geq d_n) = P_{\beta=0}(A_n)$ and our model selection criterion δ consistent which implies that $P_{\beta=0}(A_n) \to 0$. We have that $P_{\beta=0}(\sum_{i=1}^{n} x_i Y_i \geq d_n) = P_{\beta=0}(A_n) \to 0$. This implies that $\frac{d_n}{\sqrt{n}} \to \infty$. If we choose $\beta_n = \min\big(\frac{d_n}{2\sum x_i^2}, c\big)$, then:

$$\sup_{|\beta|\leq c} n\beta^2 P_\beta(A^c_{n,*}) \geq n\beta_n^2 P_{\beta_n}(A^c_{n,*}).$$

Because $\frac{d_n}{\sqrt{n}} \to \infty$, it means that d_n grows faster than $\sqrt{n}$. Therefore, it becomes clear that $n\beta_n^2 \to \infty$. For the choice of β_n we have that $P_{\beta_n}(A^c_{n,*}) \geq P\big(N(0,1) < \frac{d_n}{2\sqrt{\sum x_i^2}}\big)$, and therefore $P_{\beta_n}(A^c_{n,*}) \to 1$. With these results we can now conclude that $\sup_{|\beta|\leq c} n\beta^2 P(A^c_n) \geq \sup_{|\beta|\leq c} n\beta^2 P(A^c_{n,*}) \to \infty$. We have now proven Equation (10) and our proof for the special case is complete.

6.1.4 Proof in a general case

The second part of the proof will prove Theorem 6.1 for the general case (Yang, 2005). Let all three parts of Assumption 1 hold. Let k_1 and k_2 be two nested models according to Assumption 1. And B_n be the event such that the model selection method δ does not choose model k_1. If we assume that the model

selection method δ is consistent, then $P_{f_0}(B_n) \to 0$ as $n \to \infty$. The f_0 here is the f_0 in part three of Assumption 1. This says that the probability that the probability of not choosing model k_1 given function f_0 for which the associated true model is k_1 is zero as the sample size grows to infinity. This is simply the consistency property of δ.
Consider the simplified model:

$$Y_i = f_0(x_i) + \beta\phi(x_i) + \epsilon_i, \qquad i = 1, 2, ..., n. \tag{12}$$

Where $\phi(x_i)$ is the function which satisfies part two of Assumption 1. Now let us construct a hypothesis test similar to the proof in the special case. Let null hypothesis be $H_0 : \beta = 0$ and the alternative hypothesis be $H_1 : \beta > 0$. Furthermore, for the null hypothesis the observations come from model k_1 while the observations for the alternative hypothesis H_1 comes from the regression functions in $\mathcal{F}_{k_2}$. The model selection criterion δ is used to construct the following test: H_0 is accepted when model k_1, the true model, has been chosen by δ. And H_0 is rejected in any other cases. Because we have chosen δ to be consistent, the probability of a type I error goes to zero as n goes to infinity.
In the context of Equation (12), let $\vec{f} = [f(x_1), ..., f(x_n)]^T$ where $f(x_i) = f_0(x_i) + \beta\phi(x_1)$, $\vec{Y} = [Y_1, ..., Y_n]^T$, $\vec{\epsilon} = [\epsilon_1, ..., \epsilon_n]^T$, $\vec{\phi} = [\phi(x_1), ..., \phi(x_n)]^T$. These are all vectors and the superscript T denotes the transpose. Let M_{k_1} be the projection matrix of model k_1. The loss function for model k_1 is

$$||\vec{f} - M_{k_1}\vec{Y}||_n^2 \tag{13}$$
$$= ||\vec{f} - M_{k_1}\vec{f}||_n^2 + \vec{\epsilon}^T M_{k_1}\vec{\epsilon} \tag{14}$$
$$= ||\beta\vec{\phi} - \beta M_{k_1}\vec{\phi}||_n^2 + \vec{\epsilon}^T M_{k_1}\vec{\epsilon} \tag{15}$$
$$= \beta^2||\vec{\phi}||_n^2 + \vec{\epsilon}^T M_{k_1}\vec{\epsilon}. \tag{16}$$

The first equality follows from taking out the error. The second equality follows from that $[f_0(x_1), ..., f_0(x_n)]^T$ is the column space of M_k and $\vec{\phi}$ is orthogonal to the column space of M_k therefore the term $\vec{f_0}$ vanishes. The third equality follows again from that $\vec{\phi}$ is orthogonal to the column space of M_k. The risk function is therefore:

$$
\begin{aligned}
R(f,\delta,n) &= \frac{1}{n}\sum_{k\in\Gamma} E_\beta ||\vec{f} - M_k\vec{Y}||_n^2 I_{\{\hat{k}=k\}} \\
&\geq \frac{1}{n} E_\beta ||\vec{f} - M_{k_1}\vec{Y}||_n^2 I_{\{\hat{k}=k_1\}} \\
&\geq \frac{\beta^2}{n} E_\beta ||\vec{\phi}||_n^2 I_{\{\hat{k}=k_1\}} \\
&= \frac{\sum_{i=1}^n \phi^2(x_i)}{n} \beta^2 P_\beta(\hat{k}=k_1),
\end{aligned}
$$

where k is a candidate model in the set of candidate models Γ and k_1 is the aforementioned true model. The first equality is simply the expected value of the loss functions. The first inequality follows from that the sum being larger or equal than a single element of the sum. The second inequality follows from Equation (16). And the last equality follows from expanding the expression.
To prove Theorem 6.1 and thus the expression $n\sup_{f\in\mathcal{F}_{k_2}} R(f,\delta,n)\to\infty$ holds we can simply modify Expression (10) from the special case and construct a similar proof. That is, it suffices to prove that for each $c>0$

$$
\sup_{|\beta|\leq c} n\beta^2 P_\beta(B_n^c)\to\infty.
$$

This statement only holds for our hypothesis testing problem if we can show that for any test with rejection A_n satisfying $P_{\beta=0}(A_n)\to 0$ we must have $\sup_{|\beta|\leq c} n\beta^2 P(A_n^c)\to\infty$. We can construct a random variable $Z_i = Y_i - f_0(x_i)$ such that $Z_1,...,Z_n$ are independent Gaussian random variables with distribution $N(\beta\phi(x_i),\sigma^2)$. We can apply the arguments which we used in the simple two-model case again as they hold similarly. Thus we can conclude that $\sup_{|\beta|\leq c} n\beta^2 P_\beta(B_n^c)\to\infty$ and the proof of Theorem 6.1 is complete for the general case.

6.2 A positive outlook

We have just proven that it is impossible to combine the minimax-rate optimality property of the AIC and the consistency property of the BIC. But this hasn't discouraged researchers to stop searching for a way to combine the strengths. Multiple studies have successfully combined the strengths of the AIC and BIC in other ways, for example; (Erven, Grünwald, & de Rooij, 2012), used a "switch distribution" to achieve both model consistency and the minimax-rate optimal convergence rate to a degree under fairly weak conditions. The study of (Zheng & Loh, 1995) uses a set of penalty functions and a procedure of sorting covariates based on t-statistics to generalize Mallows' C_p, AIC, BIC and ϕ information criterion (we recommend (Hannan & Quinn, 1979) page 191 for more reading on the proposed ϕ criterion) . The methods obtained by these two studies give

a result which is better than using either AIC or BIC alone. A recent preprint on the AIC-BIC dilemma (Kirichenko & Grünwald, 2020) foregoes the assumption that the sample size is fixed in advance. By redefining the definition of minimax-rate optimality as "weakly adversarial time-robust minimax optimal" (Section 2.3 of (Kirichenko & Grünwald, 2020)) they have achieved a way to circumvent the AIC-BIC dilemma.

7 Summary and conclusion

The goal of this thesis was to take an in-depth look at the two of the most well known information criteria, the AIC and BIC, and revisit the proof of (Yang, 2005) that it is not possible to combine two of their most important properties.

To introduce these two model selection criterion we revisited some concepts in statistics such as; the KL information which is used to measure the loss of information between two models, regression analysis which is is used to examine the relationships between a response and (multiple) independent variables, likelihoods, loss and risk functions and the Bayes factor.

Afterwards I have introduced the AIC and BIC and how it relates to the concepts before. The AIC is a model selection criterion which is used for its capability to select the model with the most predictive ability. Furthermore, the AIC has an important property which is its ability to be minimax-rate optimal for both parametric and nonparametric cases. The BIC is known for its ability to choose the true model when the true model exists and is in the set of candidate models. It also has a higher penalty term compared to the AIC for large sample sizes which causes it to prefer more simple models. In the limited simulations I have performed we have seen the AIC's tendency to pick models with high predictive power as well as the consistency property for the BIC.

For applications in model selection problems both consistency and minimax-rate optimality are important. This has sparked debate around whether it is possible to combine both properties into a superior model selection criterion. Some studies have successfully shown that a superior criterion can be made which combines the strenghts of the AIC and BIC to a degree. However, with the current definition of the minimax-rate optimality it is not possible to combine both consistency as well as minimax-rate optimality. The proof of (Yang, 2005) I revisited shows this.

The research into optimizing the model selection process by combining desired properties of model selection criteria is a very complex one. And while it is not possible to have the very best of both worlds in this case, it is certain that there will be more discoveries which will advance the model selection process.

8 R code

This code is a modified version of the code listed on:
`https://www.r-bloggers.com/aic-bic-vs-crossvalidation/` written by a person who goes by the username of "Petrkeil". Most parts of the code remains the same. I've edited the code under the comment "# plotting the data and the fitted models" which plotted graph 1 because the original code written by Petrkeil didn't work properly. I modified the code such that it became easier to run multiple simulations with it. Such as adding a variable for the value of the sample size, and adding variables α, β and γ to make it easier to change the parameters of the true model. I've removed the parts in the original code which printed the graph for the AIC and BIC scores as well as the graph for the cross validation. Furthermore, I modified the code to have a loop for the Monte Carlo simulations. And allow it to identify the candidate model with the lowest AIC and BIC scores.

8.1 Code used in Section 5.2.2

```
# the figures require ggplot2 library and
# all packages it depends on
library(ggplot2)

#Setting seed
set.seed(4630521)

#Sample size
n <- 1000
# generate the x predictor
x <- runif(n,-2,2)

# set the parameters
alpha <- -2
beta <- 5
gamma <- 7
epsilon <- rnorm(n)

# generate the y response
y <- alpha*x + beta*x^2 + gamma*x^3 + epsilon
xy <- data.frame(x=x, y=y)
# specify the maximum polynomial degree that will be explored
max.poly <- 5

# creating data.frame which will store model predictions
# that will be used for the smooth curves in Fig. 1
```

```
x.new <- seq(min(x), max(x), by=0.1)
degree <- rep(1:max.poly, each=length(x.new))
predicted <- numeric(length(x.new)*max.poly)
new.dat <- data.frame(x=rep(x.new, times=max.poly),
                      degree,
                      predicted)

# fitting lm() polynomials of increasing complexity
# (up to max.degree) and storing their predictions
# in the new.dat data.frame
for(i in 1:max.poly)
{
  sub.dat <- new.dat[new.dat$degree==i,]
  new.dat[new.dat$degree==i,3] <- predict(lm(y~poly(x, i)),
                                          newdata=data.frame(x=x.new))
}

# plotting the data and the fitted models
p <- ggplot() +
     geom_point(aes(x, y), xy, colour="blue") +
     geom_line(aes(x, predicted, colour=as.character(degree)), new.dat) +
     scale_x_continuous(breaks = round(seq(-2, 2, by = 0.5),1)) +
     scale_colour_discrete(name = "Degree") +
     theme_bw() +
     labs(title="Considered models")
print(p, comment = FALSE)

# creating empty data.frame that will store
# AIC and BIC values of all of the models
AIC.BIC <- data.frame(criterion=c(rep("AIC",max.poly),
                                  rep("BIC",max.poly)),
                      value=numeric(max.poly*2),
                      degree=rep(1:max.poly, times=2))

# calculating AIC and BIC values of each model
for(i in 1:max.poly)
{
  AIC.BIC[i,2] <- AIC(lm(y~poly(x,i)))
  AIC.BIC[i+max.poly,2] <- BIC(lm(y~poly(x,i)))
}

# function that will perform the "leave one out"
# crossvalidation for a y~poly(x, degree) polynomial
crossvalidate <- function(x, y, degree)
{
  preds <- numeric(length(x))
```

```
  for(i in 1:length(x))
  {
    x.in <- x[-i]
    x.out <- x[i]
    y.in <- y[-i]
    y.out <- x[i]
    m <- lm(y.in ~ poly(x.in, degree=degree) )
    new <- data.frame(x.in = seq(-3, 3, by=0.1))
    preds[i]<- predict(m, newdata=data.frame(x.in=x.out))
  }
  # the squared error:
  return(sum((y-preds)^2))
}

# crossvalidating all of the polynomial models
# and storing their squared errors in
# the "a" object
a <- data.frame(cross=numeric(max.poly))
for(i in 1:max.poly)
{
  a[i,1] <- crossvalidate(x, y, degree=i)
}
```

8.2 Code used in Section 5.2.3

```
#Setting seed
set.seed(4630521)

#initialize variables
AIC_cor <- 0
BIC_cor <- 0

#Sample size
n <- 250

#amount of simulations
m <- 500

for(j in 1:m)
{
x <- runif(n,-2,2)
alpha <- -2
beta <- 5
gamma <- 7
epsilon <- rnorm(n)
y <- alpha*x + beta*x^2 + gamma*x^3 + epsilon
```

```
xy <- data.frame(x=x, y=y)
max.poly <- 5
x.new <- seq(min(x), max(x), by=0.1)
degree <- rep(1:max.poly, each=length(x.new))
predicted <- numeric(length(x.new)*max.poly)
new.dat <- data.frame(x=rep(x.new, times=max.poly),
                      degree,
                      predicted)

for(i in 1:max.poly)
{
  sub.dat <- new.dat[new.dat$degree==i,]
  new.dat[new.dat$degree==i,3] <- predict(lm(y~poly(x, i)),
                                          newdata=data.frame(x=x.new))
}

AIC.BIC <- data.frame(criterion=c(rep("AIC",max.poly),
                                  rep("BIC",max.poly)),
                      value=numeric(max.poly*2),
                      degree=rep(1:max.poly, times=2))

for(i in 1:max.poly)
{
  AIC.BIC[i,2] <- AIC(lm(y~poly(x,i)))
  AIC.BIC[i+max.poly,2] <- BIC(lm(y~poly(x,i)))
}

Data_AIC <- subset(AIC.BIC, criterion=="AIC")
Data_BIC <- subset(AIC.BIC, criterion=="BIC")

# finding minimum
Data_AICmin <- Data_AIC[which.min(Data_AIC$value),]
Data_BICmin <- Data_BIC[which.min(Data_BIC$value),]

# increase count if correct degree has been selected
if (Data_AICmin[,3] == 3){
  AIC_cor = AIC_cor + 1
}
if (Data_BICmin[,3] == 3){
  BIC_cor = BIC_cor + 1
}
}
(AIC_cor/m)*100
(BIC_cor/m)*100
```

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Burnham, K. P., & Anderson, D. R. (2002). A practical information-theoretic approach. *Model selection and multimodel inference, 2nd ed. Springer, New York*.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261–304.

Erven, T. v., Grünwald, P., & de Rooij, S. (2012). Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 361–417.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 190–195.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

Jeffreys, H. (1961). Theory of probability. *International Series of Monographs on Physics*.

Kirichenko, A., & Grünwald, P. (2020). Minimax rates without the fixed sample size assumption. *arXiv preprint arXiv:2006.11170*.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Neyman, J., & Pearson, E. S. (1933). Ix. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289–337.

Petrkeil. (2013). *Petrikeil: AIC & BIC vs. crossvalidation.* Retrieved from `https://www.r-bloggers.com/aic-bic-vs-crossvalidation/`

Schwarz, G. (1978). Estimating the dimension of a model, 1978. *The Annals of Statistics*, *6*, 461–464.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329). John Wiley & Sons.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the bayesian information criterion (BIC). *Psychological methods*, *17*(2), 228.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica*, 475–499.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. *Biometrika*, *92*(4), 937–950.

Zheng, X., & Loh, W.-Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association*, *90*(429), 151–156.