COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

INFORMATION CRITERIA

MASTER'S THESIS

2021 Thu Nguyen Quynh

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

Information criteria

MASTER'S THESIS

Study programme:	Probability and Statistics
Branch of study:	6211 Statistics
Department:	Department of Applied Mathematics and Statistics
Supervisor:	Mgr. Samuel Rosa, PhD.

Bratislava 2021 **Thu Nguyen Quynh**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Thu Nguyen Quynh
Študijný program: pravdepodobnosť a matematická štatistika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Information criteria
Informačné kritérá

Anotácia: Informačné kritérá (napríklad Akaikeho a bayesovské) sú bežne používané na výber spomedzi sady modelov. Z praktického hľadiska sú veľmi jednoducho aplikovateľné, obzvlášť v lineárnej regresii, a teda aj značne populárne. Napriek tomu sa študent štatistiky zväčša s hlbšou teoretickou analýzou týchto kritérií nestretne. V práci dôkladne naštudujeme a spíšeme teoretické podloženie informačných kritérií,preskúmame ich správanie na simulovaných dátach, a kritériá aplikujeme pri riešení problému na reálnych dátach.

Vedúci: Mgr. Samuel Rosa, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Dátum zadania: 22.11.2019

Dátum schválenia: 22.11.2019 prof. RNDr. Marek Fila, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Abstrakt

Výber modelu je základným problémom štatistiky, ako aj mnohých iných vied, ako je strojové učenie a ekonometria. Existuje mnoho informačných kritérií na výber skutočného modelu, ako sú Akaikeho informačné kritérium (AIC), Bayesovské informačné kritérium (BIC).

Naša diplomová práca predstavuje dve dôležité informačné kritériá, ktorými sú Akaikeho informačné kritérium a Bayesovské informačné kritérium. Táto práca sa zaoberá tvorbou, matematickými odvodeniami, motiváciou a aplikáciou týchto dvoch kritérií. Taktiež diplomová práca demonštruje vzťah a rozdiely medzi týmito dvoma kritériami pri aplikácii na lineárnu regresiu. Vykonávame niekoľko simulácií, aby sme demonštrovali výsledky, ktoré sme študovali z teórie. Simulácie vedú k zaujímavým výsledkom. Pre daný súbor údajov máme dostatok argumentov na výber vhodných kritérií.

Kľúčové slová: Výber modelu, informačné kritériá, Akaikeho informačné kritériá, Bayesovské informačné kritériá, lineárna regresia.

Abstract

Model selection is an essential problem of statistics as well as many other sciences such as machine learning, and econometrics. There are many information criteria to choose the true model such as Akaike's information criteria (AIC), Bayesian Information Criteria (BIC).

Our thesis introduces two important information criteria that are Akaike's information criterion and Bayesian's information criterion. This thesis studies the formation, mathematical derivation, motivation, and application of the two criteria. Also, the thesis demonstrates the relationship and differences between the two criteria when applied to linear regression. We perform some simulations to demonstrate the results we studied from the theory. The simulations lead to some interesting results. Given a set of data, we have enough arguments to choose appropriate criteria.

Keywords: Model selection, information criteria, Akaike's information criteria, Bayesian Information Criteria, linear regression.

Acknowledgements I would like to express my sincere thanks to my supervisor, Mgr. Samuel Rosa, PhD. for his support, patience, and encouragement. This work would not have been produced without his devoted engagement at every stage of the process.

Contents

Introduction

A given data set can possibly be put into a variety of models. The idea behind model selection is how to know if we have a good model? How do we make a fair comparison among all of the varieties of models that might fit our data? Would a regression model that has smaller residual parts or a model having less complexity be better? Because entire reality cannot be included in a model, a good model certainly is a good fit the data set under investigation and when adding more variabes to the model, the apparent fit is better. On the other hand, selecting the variables for use is an important assignment.

Model selection is the key to data analysis for statistical inference, prediction of reliable and reproducible ingredients. It is often performed in various fields such as biology, engineering, and ecology. It has gained widespread attention due to its importance. Before the 1970s, it was mainly focused on estimating the model's precision and parametric estimation. The selection of a model is crucial in the statistical analysis of data. After the appearance of Akaike's research, model selection started to attract the attention of the statisticians.

There has been a long history of information criteria that arise from research in statistics, information theory, and signal processing. The adjusted R-squared R^2_{adj} known as the first model selection criterion was first published in 1921 by the geneticist Sewall Wright still be used today in many regression materials. As is well known, the R-squared R^2 is not an appropriate method for choosing among alternative specifications because when adding a variable to the model it always increases, and thus it leads to a choice of the model with the highest possible dimensionality. For this always-increasing property, R^2_{adj} is an alternative criterion to the basic R^2.

In the 1970s, Akaike proposed in the book [2] the seminal Akaike Information Criterion (AIC) which is based on the Kullback-Leibler distance. He pointed out that statistical hypothesis tests are not suitable for model selection. Furthermore, he found out another procedure known as Minimum Akaike Information Criterion Estimate (MAICE) which is more suitable and does not create many inaccuracies for testing model assumptions. The author also claimed that, although MAICE can be used for

most statistical models, this metric is most useful for time series models. In 1975, Tong developed in [25] a procedure to determine the order of AR (autoregressive) models using the AIC. This procedure has been proven to work very well on both simulated and real data sets. Interestingly, Tong's problem is similar to classical factor analysis problems when the parameters are estimated by the ML method and the order is determined by MAICE. Therefore, Tong's procedure is considered as an extension of the order determination procedure in factor analysis (see more in [25]).

Shibata analyzed the statistical properties of AIC and proposed a statistical model fitting method in his study [23]. In particular, the author checked the consistency of the AIC and showed that the AIC is inconsistent in the selection of orders in the AR (autoregressive) model. In the late 1970s, there was an eruption in the information theory field: the Bayesian Information Criterion (BIC, first introduced in [3]), the Hannan and Quinn criterion (HQ, first proposed in [11]), and The Final Prediction Error criterion (FPE–[4]). Hannan and Quinn in [11] pointed out that BIC is a strongly consistent criterion for model selection of AR models. In [20], Nishii using cross-validation have shown that using AIC and FPE are equivalent. In the work [23] of Shibata, he also proved that AIC and FPE are equivalent. Similarly, AIC and BIC are equivalent in many studies, such as in [11]. In general, many later studies using both AIC and BIC indexes showed similar results. In [14], Hurvich and Tsai developed an enhanced unbiased small-sample estimator of the Kullback-Leibler divergence, AICc, based on Sugiura's results in the book [24]. Their simulation results showed that AICc has proven itself to be one of the strongest model selection criteria.

We will discuss the two popular information criteria AIC and BIC. In the first chapter, we will introduce the Kullback-Leibler information which is the original idea of AIC, and the relationship between the maximum likelihood estimation and the Kullback-Leibler information. We will formulate the definition of AIC, the relationship between AIC and the Kullback-Leibler information, and also some related criteria such as the Takeuchi criterion (TIC), Corrected AIC (AICc), and some commonly used quantities such as AIC differences, Akaike weights.

We will discuss the Bayesian information and its derivation from Bayesian theory

in the second chapter. Then we will spend a part discussing the differences between AIC and BIC in terms of derivation, formulas, statistical consistency, and asymptotic efficiency.

In the next chapter, using the theory of linear regression, we will introduce and formulate in detail the formulas of AIC and BIC for regression models and other criteria for evaluating the regression model.

In the last chapter, we will begin with a simple case study of the Bodyfat data so as to illustrate how BIC and AIC work in R and also their application in the linear regression model. In the next sections, we will focus on simulating data to study the differences between AIC and BIC and compare the results with the theory shown in previous chapters.

1 Akaike's Information Criterion

Akaike's information criterion is known as the AIC which means "an information criterion", was introduced in 1973 by the Japanese statistician Hirotogu Akaike in the book [2]. The Akaike Information Criterion (AIC) was developed with information theory. Information theory is a branch of applied mathematics that deals with the quantification (the process of counting and measuring) of information. When it comes to comparing multiple models, the AIC is well known as a method that takes into account both descriptive accuracy and parsimony ([2], [3], and emphasized in book [6]). AIC works in a range of applications. For instance, AIC is applied in factor analysis introduced in [2], regression (see [6], [16]), and latent class analysis (in [10]). It measures the relative Kullback-Leibler information between the likelihood function specified by a fitted candidate model and the unknown true likelihood model that generated the data. Therefore, AIC is related to behaviour of the Kullback-Leibler information and the log-likelihood function, as we will discuss in the following sections.

1.1 Maximum Likelihood Estimation

One of the most important ideas in statistics is random sampling. We can understand that random sampling is similar to voting. People go to the polls to represent them. In statistics, we aim to collect data in such a way that the observations in the sample represent the entire population. However, we do not know the values of all observations in the population, we pick one at random and expect it to represent the entire population.

Maximum Likelihood Estimation (MLE) is based on the idea that what can be seen is the most likely thing to happen. As a result, the density functions at the observations are multiplied, and the parameters are found such that the likelihood function reaches its maximum value. The Akaike Information Criterion can be interpreted as an extension of the MLE method, allowing us to not only approximate parameters of the model, but also select the model. In this section, we will mostly follow [16], [13], and [29] in order to approach the concept of maximum likelihood estimation method.

1.1.1 The Likelihood Function and The Maximum Likelihood Estimator

We start with a vector $X = (x_1, x_2, ..., x_n)$ as notation for n observations given by a population. This population can be characterized by the probability density function (pdf) denoted by $\{f(x|\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$. Similarly, discrete probability models with a probability mass function (pmf) will be denoted by $\{p(k|\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$, where $K = (k_1, k_2, ..., k_n)$ is a random sample of size n. Let's refer $\theta = (\theta_1, \theta_2, ..., \theta_p)^T$ as p-dimensional unknown parameter to estimate. Following Konishi and Kitagawa in book [16], we define the *likelihood function*, $L_n(.|\theta)$ which is written as

$$L_n(K|\theta) = \prod_{i=1}^{n} p(k_i|\theta), \quad \text{for discrete observations,} \tag{1.1}$$

and

$$L_n(X|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \quad \text{for continuous observations.} \tag{1.2}$$

However, we often work with the natural logarithm of the likelihood function for several reasons:

- Arithmetic stability: The values of PDF or PMF are smaller than 1 so when they are multiplied, we will obtain a very "small" value, and it is difficult to handle errors.
- Ability to convert products to sums: simplification of calculations because finding the maximum of a sum will be easier than of a product.
- The relationship to information theory.

The natural logarithm of the likelihood function is call the log-likelihood, and is represented as

$$l_n(K|\theta) = \log(L_n(K|\theta)) = \sum_{i=1}^{n} \log p(k_i|\theta), \quad \text{for discrete observations,} \tag{1.3}$$

and

$$l_n(X|\theta) = \log(L_n(X|\theta)) = \sum_{i=1}^{n} \log f(x_i|\theta), \quad \text{for continuous observations,} \tag{1.4}$$

where log denotes the natural logarithm.

The principle of MLE results in a set of parameters that maximizes the likelihood function (resp. log-likelihood) over the parameter space. This value is a *maximum likelihood estimator* for parameter θ.

Definition 1.1. *(see [17]-p.45) Let* $L_n(K|\theta) = \prod_{i=1}^n p(k_i|\theta)$, *and* $L_n(X|\theta) = \prod_{i=1}^n f(x_i|\theta)$ *be the likelihood functions corresponding to random samples* $k_1, k_2, ..., k_n$ *from the discrete pdf* $p(k|\theta)$ *and* $x_1, x_2, ..., x_n$ *from the continuous pdf* $f(x|\theta)$, *respectively, where* θ *is an unknown p-dimensional parameter. In each case, let* $\hat{\theta}$ *be a value of the parameter such that* $L(.|\hat{\theta}) \geq L(.|\theta)$ *for all possible of* θ. *Then* $\hat{\theta}$ *is called a maximum likelihood estimator for* θ

$$\hat{\theta} = \arg\max_{\theta\in\Theta}(L_n(.|\ \theta)) = \arg\max_{\theta\in\Theta}(l_n(.|\ \theta)). \tag{1.5}$$

1.1.2 Score Vector, Hessian Matrix, and Fisher Information

In the case that log-likelihood function $l_n(.|\theta)$ is continuously differentiable, maximum likelihood estimators is found by solving the likelihood equation

$$\frac{\partial}{\partial\theta} l_n(.|\theta)|_{\theta=\hat{\theta}} = \mathbf{0}, \tag{1.6}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$ and $\mathbf{0}$ is a zero vector of dimension p, $\mathbf{0} = (0, 0, \ldots, 0)^T$. In this section, we will use $l_n(\theta)$ instead of $l_n(.|\theta)$.

We define

$$s_n(\theta) = \frac{\partial l_n(\theta)}{\partial\theta}, \tag{1.7}$$

and

$$H(\theta) = \frac{\partial^2 l_n(\theta)}{\partial\theta\partial\theta^T}. \tag{1.8}$$

In the expression (1.7), $s_n(\theta)$ is a p-dimensional vector function, often called the *score vector* of the model with components $\frac{\partial l_n(\theta)}{\partial\theta_i}$ for $i = 1, \ldots, p$:

$$s_n(\hat{\theta}) = \begin{pmatrix} \frac{\partial l_n(\theta)}{\partial\theta_1}|_{\theta=\hat{\theta}} \\ \vdots \\ \frac{\partial l_n(\theta)}{\partial\theta_p}|_{\theta=\hat{\theta}} \end{pmatrix}. \tag{1.9}$$

The equation (1.6) yields

$$s_n(\hat{\theta}) = \begin{pmatrix} \frac{\partial l_n(\theta)}{\partial\theta_1}|_{\theta=\hat{\theta}} \\ \vdots \\ \frac{\partial l_n(\theta)}{\partial\theta_p}|_{\theta=\hat{\theta}} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{1.10}$$

While the score vector tells us whether $\hat{\theta}$ is a stationary point of function $l_n(\theta)$, then a second-order partial derivative of $l_n(\theta)$ allows us to test whether $\hat{\theta}$ is a local maximum,

local minimum, or a saddle point. This matrix is called the Hessian matrix. Its components are the mixed second-order derivatives

$$\frac{\partial^2 l_n(\theta)}{\partial\theta_j\partial\theta_k}\ \forall j,k=1,\ldots,p. \tag{1.11}$$

Hessian matrix is used for numerically finding the maximum likelihood estimators and characterising their behaviour.

Remark 1. The Hessian matrix is a $p \times p$ matrix:

$$H(\theta)=\frac{\partial^2 l_n(\theta)}{\partial\theta\partial\theta^T}=\begin{pmatrix} \frac{\partial^2 l_n(\theta)}{\partial\theta_1^2} & \frac{\partial^2 l_n(\theta)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 l_n(\theta)}{\partial\theta_1\partial\theta_p} \\ \frac{\partial^2 l_n(\theta)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 l_n(\theta)}{\partial\theta_2^2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 l_n(\theta)}{\partial\theta_p\partial\theta_1} & \cdots & \cdots & \frac{\partial^2 l_n(\theta)}{\partial\theta_p^2} \end{pmatrix}. \tag{1.12}$$

In addition to the score vector and Hessian matrix, Fisher information associates with maximum likelihood estimation. The Fisher information matrix can be defined as the variance-covariance matrix of the score vector:

$$J_n(\theta)=Var_\theta(s_n(\theta)). \tag{1.13}$$

The Fisher information matrix is also given by:

$$J_n(\theta)=E_\theta(s_n(\theta)\times s_n(\theta)^T). \tag{1.14}$$

Under some regularity conditions, one property of the Fisher information matrix is that it equals the negative of the expected value of the Hessian matrix of the log-likelihood evaluated at the true parameter value θ_0 as the following alternative definition:

$$J_n(\theta)=E_\theta(-H(\theta)). \tag{1.15}$$

1.1.3 Properties of Maximum Likelihood Estimators

In this section we will try to understand why maximum likelihood estimators are "good". We will prove that maximum likelihood estimators are consistent and asymptotically normal as long as these regularity conditions are satisfied.

Conditions (Regularity conditions for maximum likelihood estimators - see [17]).
R1. The number of parameters p is constant.

R2. The true but unknown parameter value θ_0 of the "true" model f is defined as:

$$\theta_0 = \arg\max_{\theta\in\Theta} E_f(l_n(.|\theta)). \tag{1.16}$$

A common name for θ_0 is known as the *least false* or *best approximating* parameter value ([8]).

R4. The parameter space Θ is a compact set and the true value of θ lies in it.

R5. The first three log-likelihood derivatives $l_n(.|\theta)$ are continuously differentiable and finite in a neighborhood of θ_0.

Intuitively, we can see the that as n increases, $\hat{\theta}$ becomes more and more concentrated to θ_0. In this part, we denote the maximum likelihood estimator by $\hat{\theta}_n$ to emphasize the dependence on n. Let $g(x|\theta) : \theta \in \Theta \ \subset \mathbb{R}^p$ be a continuous parametric model, with $\theta \in \Theta \subset \mathbb{R}^p$ is p-dimensional parameter vector. Let $x_1, \ldots, x_n$ be the data generated from the "true" distribution $f(x)$. Assume that θ_0 is the least false parameter defined in (1.16). Under the above regularity conditions, θ_0 is a solution of

$$\int f(x)\frac{\partial \log(g|\theta)}{\partial\theta} = 0.$$

It can be shown that $\hat{\theta}_n$ converges in probability to θ_0 when n goes to infinity. The following lemma is taken from Newey and McFadden (1994)–[19].

Lemma 1.2. ***(Consistency)*** *Under regularity conditions and the assumptions described above, we have the maximum likelihood estimator $\hat{\theta}_n$ converges almost surely (i.e. converges in probability) to θ_0 when $n \to \infty$*

$$\hat{\theta} \xrightarrow[n\to\infty]{p} \theta_0. \tag{1.17}$$

Following some studies of [6], [16], [19], and [8], the lemma can be proved as follows.

Proof. $g(x|\hat{\theta}_n)$ is the best model approximating the true $f(x)$ by using the maximum likelihood method. The maximum likelihood estimator $\hat{\theta}_n$ is defined as $\arg\max_{\theta\in\Theta} l_n(x|\theta)$. The log-likelihood function $l_n(x|\theta)$ is

$$l_n(x|\theta) = \sum_{i=1}^{n} \log g(x_i|\theta). \tag{1.18}$$

On the other hand, applying the strong law of large numbers when n (the number of observations) tends to infinity

$$\frac{1}{n}\sum_{i=1}^{n}\log g(x_i|\theta) \overset{n\to\infty}{\to} E_f \log(g(x|\theta)). \tag{1.19}$$

Thus, by denoting $p\lim$ as a limit in probability then putting things above together, we obtain

$$\begin{aligned} p\lim\hat{\theta}_n &= p\lim(\arg\max_{\theta\in\Theta}\frac{1}{n}l_n(x|\theta)) \\ &= \arg\max_{\theta\in\Theta}(p\lim\frac{1}{n}l_n(x|\theta)) \\ &= \arg\max_{\theta\in\Theta}E_f(l_n(x|\theta)) \\ &= \theta_0. \end{aligned} \tag{1.20}$$

The last equality is true, because it is the assumption of θ_0 in the regularity condition **R2**. Take a note that in this proof we do not deal with the technical details (e.g., we do not prove the type of convergence) and thus it is technically not complete. □

In the next theorem, we will show that the maximum likelihood estimator $\hat{\theta}_n$ is approximately distributed according to a normal distribution for a large sample size n.

Theorem 1.3. *Under regularity conditions and the assumptions described above, let $\hat{\theta}_n$ be the maximum likelihood estimator which is a sequence $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \ldots, x_n)$ and is consistent to the true parameter $\theta_0 \in \Theta$. If $J(\theta_0)$ is the Fisher information matrix for one observation at $\theta = \theta_0$ then we get:*

$$\hat{\theta}_n \overset{asy}{\approx} N(\theta_0, n^{-1}J^{-1}(\theta_0)). \tag{1.21}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N(0, J^{-1}(\theta_0)). \tag{1.22}$$

Proof. Applying the first order Taylor expansion to the score, with initial point θ_0:

$$s_n(\theta) = \frac{\partial l_n(\theta)}{\partial\theta} = \frac{\partial l_n(\theta)}{\partial\theta}|_{\theta=\theta_0} + \frac{\partial^2 l_n(\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta_0}(\theta-\theta_0) + R(\theta,\theta_0). \tag{1.23}$$

Dividing by n and rewriting as a function of sum of $x_1, x_2, \ldots, x_n$ then we get

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log g(x_i|\theta)}{\partial\theta} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log g(x_i|\theta)}{\partial\theta}|_{\theta=\theta_0} \\ &+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log g(x_i|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta_0}(\theta-\theta_0) \\ &+\frac{1}{n}\sum_{i=1}^{n}R(x_i|\theta,\theta_0). \end{aligned} \tag{1.24}$$

If (1.24) is computed at $\hat{\theta}_n$ then with the assumption of $\hat{\theta}_n$, the left hand side is zero. Thus we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(x_i|\theta)}{\partial\theta}|_{\theta=\theta_0}+\frac{1}{n}\sum_{i=1}^{n}R(x_i|\hat{\theta}_n,\theta_0)=-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log g(x_i|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta_0}(\hat{\theta}_n-\theta_0).$$

We multiply both sides by $\sqrt{n}$

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log g(x_i|\theta)}{\partial\theta}|_{\theta=\theta_0}+\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}R(x_i|\hat{\theta}_n,\theta_0)=-\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log g(x_i|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta_0}(\hat{\theta}_n-\theta_0). \tag{1.25}$$

For this proof, we have made a few corrections to the proof from [27] since in their proof, the fraction $\frac{1}{n}$ in (1.25) is disappeared. Equation (1.25) is equivalent to

$$\begin{aligned}&-\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log g(x_i|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta_0}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log g(x_i|\theta)}{\partial\theta}|_{\theta=\theta_0}\\&\quad+\left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log g(x_i|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta_0}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}R(x_i|\hat{\theta}_n,\theta_0)=\sqrt{n}(\hat{\theta}_n-\theta_0).\end{aligned} \tag{1.26}$$

For a conveniently large n, the residual term becomes negligible. Thus, we have

$$\sqrt{n}(\hat{\theta}_n-\theta_0)\approx\left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log g(x_i|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta_0}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log g(x_i|\theta)}{\partial\theta}|_{\theta=\theta_0}. \tag{1.27}$$

By applying a law of large number, we can see the average of n terms $\frac{\partial^2 \log g(x_i|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\theta_0}$ converges in probability to their expected value which equals the Fisher information matrix $J(\theta_0)$. In the second sum each of the terms $\frac{\partial \log g(x_i|\theta)}{\partial\theta}|_{\theta=\theta_0}$ has expected value zero and variance matrix $J(\theta_0)$, thus according to the Central Limit Theorem (CLT) we obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log g(x_i|\theta)}{\partial\theta}|_{\theta=\theta_0}\xrightarrow{d}N(0,J(\theta_0)).$$

Therefore, we can conclude that

$$\sqrt{n}(\hat{\theta}_n-\theta_0)\xrightarrow{d}N(0,J^{-1}(\theta_0)). \tag{1.28}$$

Thus $(\hat{\theta}_n-\theta_0)$ is approximately distributed according a normal distribution with zero mean and $n^{-1}J^{-1}(\theta_0)$ variance-covariance matrix

$$\begin{aligned}(\hat{\theta}_n-\theta_0)&\approx N(0,n^{-1}J^{-1}(\theta_0)),\\&\text{or equivalently}\\\hat{\theta}_n&\approx N(\theta_0,n^{-1}J^{-1}(\theta_0)).\end{aligned} \tag{1.29}$$

□

If we denote the Fisher information matrix associated to the sample size n as $J_n(\theta)$, then [27] has demonstrated that

$$J_n(\theta) = nJ(\theta). \tag{1.30}$$

Therefore another way, to write (1.29), is that for large sample size n, the maximum likelihood estimator is approximately distributed according a normal distribution with the mean θ_0 and variance $J_n^{-1}(\theta_0)$

$$\hat{\theta}_n \approx N(\theta_0, J_n^{-1}(\theta_0)). \tag{1.31}$$

1.2 Kullback-Leibler Information

The marriage of mathematical statistics and information theory started with the Kullback-Leibler divergence in Kullback's book (1959)-[13]. Kullback-Leibler divergence is a measure of how much information we lose when we choose an approximation. The Kullback-Leibler divergence was first introduced in Kullback and Leibler (1951) as the "directed divergence" between two models and the term "discrimination information" is preferred by Kullback (1987) -[3]. According to some materials (e.g., [6], [8], also [16]), we tend to use *Kullback-Leibler information* (K-L information) or *Kullback-Leibler divergence* (K-L divergence). We do not use the term *Kullback-Leibler distance* as some materials used. It may be enticing to think of divergence as a distance metric, but in fact the distance between two distributions or the distance between two models in general cannot be calculated by using K-L divergence. This is because K-L divergence is not symmetrical. Specifically, the information lost when using model A to appoximate model B is not equal to the information lost when B is used to approximate A.

Considering n independent observations $x_1, x_2, \ldots, x_n$ from an undefined probability distribution function $F(x)$, we use $F(x)$ to denote the true distribution (the cumulative distribution function, which generates data as the true model). Although there are no models that specifically reflect full reality, it is possible to denote the full truth as $F(x)$. Our aim is to evaluate "how well" a given candidate model $G(x)$ approximates the true model. Both $F(x)$ and $G(x)$ are simple probability distributions. If $f(x)$ and $g(x)$ respectively, are called density functions, then they are called

continuous models or continuous distribution models. In contrast, they are considered as discrete models if they are represented as probabilities of events

$$f_i = f(x_i) \equiv Pr(\omega;\ Y(\omega) = y_i), \quad i = 1, 2, \ldots$$

$$g_i = g(x_i) \equiv Pr(\omega;\ Y(\omega) = y_i), \quad i = 1, 2, \ldots$$

while $x_1, x_2, \ldots, x_k, \ldots$ is a set of discrete, finite or countably infinite points.
To explain "how well" in the language of mathematics, we use the notation $KL(f,\ g)$ to calculate the "information lost when g is used to approximate f" or "the divergence from g to f". The K-L divergence quantifies how much information is lost by looking at the expectation of the natural logarithm difference between f and g. If we think in terms of natural logarithm, we can seek how many bits of information we expect to lose when we encode our amount of information lost. We could write our formula in term of expectation with respect to the true model f as

$$KL(f,\ g) = E_f \log(\frac{f(.)}{g(.)}), \tag{1.32}$$

where log denotes the natural logarithm (see [6]).

The more common way to see K-L divergence written is as the two following definitions.

Definition 1.4. *If $f(x)$ and $g(x)$ are the density functions of continuous models, then the K-L divergence can be expressed as the integral*

$$KL(f,\ g) = \int f(x) \log(\frac{f(x)}{g(x)})dx. \tag{1.33}$$

Definition 1.5. *If $f(x_i)$ and $g(x_i)$, $\quad i = 1, 2, \ldots$, are the probabilities of discrete models, then the K-L divergence can be expressed as the sum*

$$KL(f,\ g) = \sum_{i=1}^{\infty} f(x_i) \log(\frac{f(x_i)}{g(x_i)}). \tag{1.34}$$

(see [6]).

The better our approximation, the less lost information. This implies to minimizing $KL(f, g)$.

The analytic properties of K-L divergence are extensively discussed in [13] and also repeated in [16]. Here we list some of the important ones, which will play a crucial role in the development of AIC.

Lemma 1.6. *Let $g(x)$ and $f(x)$ be respectively density functions of two probability distribution functions $G(x)$ and $F(x)$. The K-L information has the following properties:*

1. $KL(f,\ g) \geq 0$, *whenever* $f(x) \neq g(x)$.
2. $KL(f,\ g) = 0 \iff f(x) = g(x)$, *almost everywhere in the possible range of* x.
3. *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables. Then the K-L information of the whole sample is $KL_n(f,\ g) = nKL(f,\ g)$. On the other hand, if the random variables are independent, the K-L information is additive.*

Proof. See [16] -p. 30, 31. □

1.3 Expected Log-Likelihood and Corresponding estimator

We will start by a specific example given in the book [16] of how expected log-likelihood operates and used.

Example 1.1. *Assume that 2 following probability distributions are the probabilities for rolling two dice, with the numbers ranging from one to six*

$$g_a = \{0.2;\ 0.12;\ 0.18;\ 0.12;\ 0.20;\ 0.18\}$$

$$g_b = \{0.18;\ 0.12;\ 0.14;\ 0.19;\ 0.22;\ 0.15\}$$

***Question**: Which is the fairer die between g_a and g_b, considering the true model f of a fair die is $f = \{\frac{1}{6}; \ldots ; \frac{1}{6}\}$.*

The idea is that we calculate the K-L information $KL(f,\ g_a)$ and $KL(f,\ g_b)$. The smaller value gives us the closer to the ideal fair die. By using the formula for discrete distribution

$$KL(f,\ g) = \sum_{i=1}^{6} f_i \times \log \frac{f_i}{g_i}$$

After calculation, we have $KL(f,\ g_a) = 0.023$ and $KL(f,\ g_b) = 0.02$ Thus, die with probability distribution g_b is "fairer" than the other.

Take a note that if E_f is an expectation with respect to the true model f, then K-L divergence can be expressed as a difference between two statistical expectations

$$KL(f,\ g_a) = E_f[\log(f(x))] - E_f[\log(g_a(x))]$$

$$KL(f,\ g_b) = E_f[\log(f(x))] - E_f[\log(g_b(x))]$$

The first expectation $E_f[\log(f(x))]$ is known as a constant C. Let g_b be a better approximation than g_a. Thus,

$$KL(f,\ g_b) < KL(f,\ g_a).$$

Then

$$KL(f,\ g_b) - C < KL(f,\ g_a) - C.$$

Hence,

$$E_f[\log(g_b(x))] > E_f[\log(g_a(x))].$$

Moreover,

$$KL(f,\ g_b) - KL(f,\ g_a) = E_f[\log(g_b(x))] - E_f[\log(g_a(x))].$$

Then, we can identify to which extent model g_b is better than g_a. By calculating the K-L divergence, we can evaluate the appropriateness of a given model.

In actual data analysis, we do not know the term $E_f[\log(f(x))]$, so its value cannot be calculated directly. Thus, we do not want to focus more on this term. We are more interested in using $E_f[\log(g(x))]$ for selecting the best model. This term is referred to as the *expected log-likelihood* ([16]-p.47) or the *relative directed distance* between f and g ([6] -p.85). The better approximated model is the larger this value and the smaller its K-L divergence.

Definition 1.7. *The expected log-likelihood of an approximating model $g(x)$ with respect to the true model f can be expressed as*

$$E_f[\log(g(x))] = \sum_{i=1}^{\infty} f(x_i)\log(g(x_i)) \quad \textit{(for discrete models)} \tag{1.35}$$

$$E_f[\log(g(x))] = \int_{-\infty}^{\infty} f(x)\log(g(x))dx \quad \textit{(for continuous models)} \tag{1.36}$$

1.4 Akaike's Information Criterion

We now look at the foundation criterion of the K-L divergence based on model selection criteria, the Akaike information criterion (AIC). AIC was first defined by

Akaike (Figure 1) as "an information criterion". To select a model from a list of candidates, AIC is one of the most well known and versatile strategies.

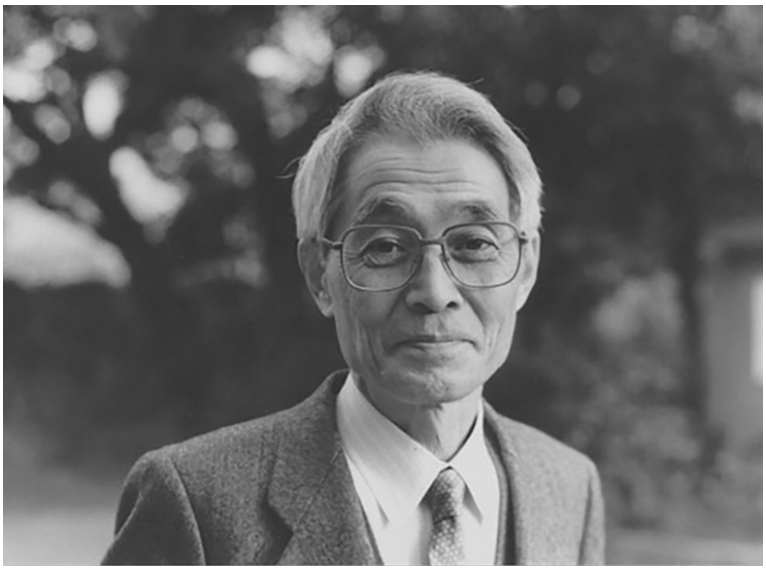

Figure 1: Hirotugu Akaike (November 5, 1927 – August 4, 2009) was a Japanese statistician. In the early 1970s, he formulated the Akaike information criterion (source: [28]).

Definition 1.8. *Akaike's information criterion or AIC has general form*

$$AIC = -2[\log((L(.|\hat{\theta}))) - p]$$

or

$$AIC = -2\log(L(.|\hat{\theta})) + 2p, \tag{1.37}$$

where $\log(L(.|\hat{\theta}))$ *represents the log-likelihood function at maximum likelihood estimator point* $\hat{\theta}$ *and* p *is the dimension of the parameter* θ*,* p *can be counted as the number of estimable parameters in the model.*

The first term $-\log(L(.|\hat{\theta}))$ is a measure of lack of model fit. Meanwhile, the second term can be regarded as a penalty for a model with increasing dimension which is more sophisticated than a simple one. As indicated by equation (1.37), AIC rewards descriptive accuracy by the maximum log-likelihood, and penalizes parsimony shortage by the number of free parameters. In practice, we determine the AIC for each candidate model and find the model with the smallest AIC value. Models that yield smaller AIC values can be considered to have a smaller difference from the true model. AIC provides

a simple and efficient way of selecting the most appropriate model for the true model. Regarding general linear models, AIC is rather good for small samples, but in large samples the criterion does not attempt to choose the true model.

Here are some properties of the AIC ([16]):

1. AIC differences between models based on different data sets are not comparable.
2. The order of computation of AIC values is irrelevant.
3. Models which are not included in the selection is not considered.

1.5 Mathematical reasons behind the AIC version

As discussed at the initial chapter, there is a relationship between AIC and the K-L information. Models that produce smaller AIC values can be considered to have a smaller difference from the true model. This section will elaborate on the mathematical derivation of AIC.

The AIC formula is simple but its derivation is often cryptic and requires many assumptions. In this section, the K-L information is taken as a criterion for investigating advantages or disadvantages of a selected model that approximately illustrates the true one. With that being said, the AIC is derived by correcting the asymptotic bias of the log-likelihood when estimating the K-L information.

1.5.1 Expected log-likelihood and maximum log-likelihood

We have the true model $f(.)$ with full reality. In actuality, we do not know $f(.)$ and we want to approximate $f(.)$ by its class candidate models $g(.|\theta_1)$, $g(.|\theta_2)$, ... Logically, we can follow these steps

1. First, using K-L information to find the closeness of two models $g(.|\theta_i)$ and $f(.)$. The model returning smallest value is selected and denoted by

$$g(.|\theta_0) = \arg\min_{\theta_i \in \Theta} KL(g(.|\theta_i), f(.)). \tag{1.38}$$

The K-L information for the best approximating model $g(.|\theta_0)$ is

$$\begin{aligned} KL[f(.), g(.|\theta_0)] &= \int f(x) \log[\frac{f(x)}{g(x|\theta_0)}]dx \\ &= E_f[\log \frac{f(.)}{g(.|\theta_0)}] \\ &= E_f[\log f(.)] - E_f[\log g(.|\theta_0)] \\ &= const - E_f[\log g(.|\theta_0)] \\ &= const - Q_0. \end{aligned} \tag{1.39}$$

2. We do not know θ_0 because the concept of true model f is abstract. Therefore in the second step, we find the maximum likelihood estimator that maximizes the log-likelihood of distribution $g(.|\theta)$ and denote by $\hat{\theta}$. Note that $\hat{\theta}$ is a random variable that depends on the data generated from the true model $f(.)$. We measure the information loss when using $g(.|\hat{\theta})$ to approximate $f(.)$

$$\begin{aligned} KL[f(.), g(.|\hat{\theta})] &= \int f(x) \log[\frac{f(x)}{g(x|\hat{\theta})}]dx \\ &= const - E_f[\log g(.|\hat{\theta})] \\ &= const - R_n, \end{aligned} \tag{1.40}$$

   where R_n is a random variable based on the data and the observations n. Thus, we study the expectation of R_n, which we denote by Q_n

$$Q_n = E_f(R_n) = E_f(\int f(x) \log g(x|\hat{\theta})dx). \tag{1.41}$$

We consider θ_0 as the true parameter. Our task is to evaluate the goodness of model $g(.|\hat{\theta})$ in terms of comparison to $g(.|\theta_0)$. It is worth to note that estimator $\hat{\theta}$ would not equal θ_0 almost surely and the probability of equality $\hat{\theta} = \theta_0$ is much less than 1 ([6]). Thus, the relationship

$$E_f[KL(f(.), g(.|\hat{\theta}))] > KL[f(.), g(.|\theta_0)] \tag{1.42}$$

always holds.

This equivalent to

$$E_f(const - R_n) > const - Q_0. \tag{1.43}$$

Thus,

$$Q_n = E_f(R_n) = E_f(\int f(x) \log g(x|\hat{\theta}))dx < Q_0. \tag{1.44}$$

1.5.2 How is AIC calculated?

The task of statistical inference is to build a model from a population whose distribution is $f(x)$ or $F(x)$. We take out a data set to observe, usually, the data is divided into two parts called training set and test set. $X = \{x_1, x_2, \ldots, x_n\}$ is denoted as the set of the training data set and $Y = \{y_1, y_2, \ldots, y_m\}$ represents the test set. The structure of X is captured by a parametric model $\{g(X|\theta); \theta \in \Theta \subset R^p\}$. We estimate θ by using the maximum likelihood method and found $\hat{\theta}$ as the result.

Our goal is to evaluate the goodness or badness of model $g(x|\hat{\theta})$. The test set Y or more specifically the model $g(y|\hat{\theta})$ is used to evaluate the model $g(x|\hat{\theta})$. From the point of view of K-L divergence, the goodness of the model $g(y|\hat{\theta})$ should be evaluated in terms of the expected log-likelihood $E_f[\log g(Y|\hat{\theta})]$. Akaike based on this argument stated

$$-2nE_f[\log g(Y|\hat{\theta})],$$

in order to criticise $E_f[\log g(Y|\hat{\theta})]$.

From the law of large number, $E_f[\log g(Y|\hat{\theta})]$ can be computed based on $\{x_1, x_2, \ldots, x_n\}$

$$E_f[\log g(Y|\hat{\theta})] \approx \frac{1}{n}\sum_{i=1}^{n} \log g(x_i|\hat{\theta}). \tag{1.45}$$

The right term $\sum_{i=1}^{n} \log g(x_i|\hat{\theta})$ is the log-likelihood function of data $x_1, x_2, \ldots, x_n$ at the point $\hat{\theta}$. Hence, the general form of AIC information criterion can be obtained by evaluating the bias $D = l_n(\hat{\theta}) - nE_f[\log g(Y|\hat{\theta})]$ and correcting the asymptotic bias of the log-likelihood as

$$AIC = -2[\text{log-likelihood of statistical model } l_n(\hat{\theta}) \text{ - bias estimator } D]. \tag{1.46}$$

There are many methods to find D, Konishi and Kitagawa introduced a method that is decomposition D into 3 parts[1] and found $D =$ the number of parameters p (see [16]–p. 66-72).

Therefore, the AIC is given by

$$AIC = -2l_n(\hat{\theta}) + 2p. \tag{1.47}$$

[1] as in Figure 2

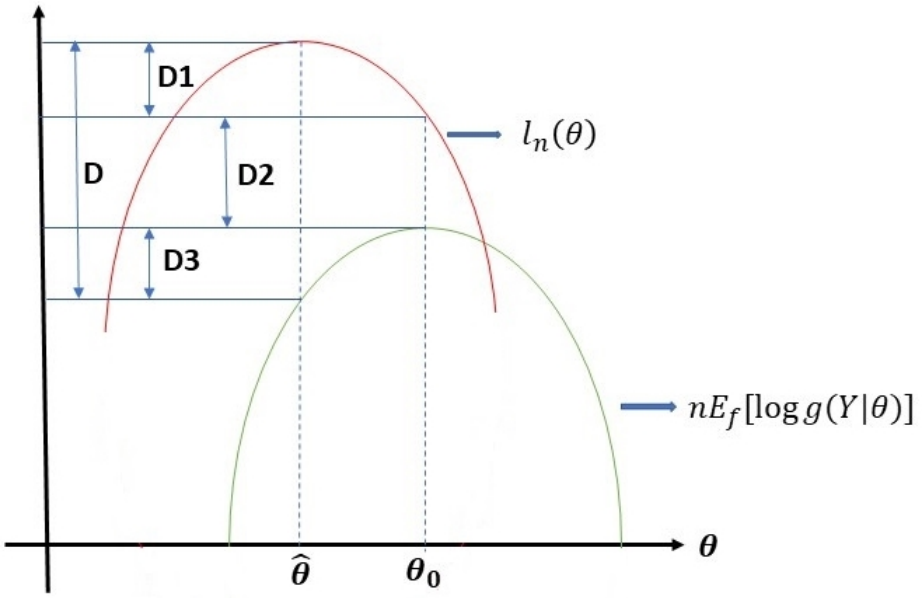

Figure 2: A relationship between log-likelihood and expected log-likelihood.

The approximate asymptotic bias equals p is too optimistic and the resulting penalty for model complexity is too weak. In case n is small or not large relative to p, AIC does not perform well. Several researchers have suggested using a corrected version, AIC_c which we will discuss in the next section.

1.6 Takeuchi

Takeuchi information criterion (TIC) is an alternative for AIC while TIC is applied when the true model is not a candidate. In 1976, Takeuchi derived a model-robust version that applies in general in case one does not want to make the assumption about the true model–[26].

Takeuchi proposed such an estimator and the corresponding criterion,

$$TIC = 2\log(L(\hat{\theta})|X) - 2\hat{K^*},$$

with $\hat{K^*} = tr(J^{-1}(\hat{\theta})I(\hat{\theta}))$, where $J(\hat{\theta})$ defined as the Fisher information matrix at $\hat{\theta}$

$$J(\hat{\theta}) = \frac{-1}{n}\left[\sum_{i=1}^{n}\frac{\partial^2 \log g(x_i|\theta)}{\partial\theta\partial\theta^T}\right]_{\theta=\hat{\theta}}, \tag{1.48}$$

and $I(\hat{\theta})$ is obtained as

$$I(\hat{\theta}) = \frac{1}{n}\left[\sum_{i=1}^{n}\frac{\partial \log g(x_i|\theta)}{\partial\theta}\sum_{i=1}^{n}\frac{\partial \log g(x_i|\theta)}{\partial\theta^T}\right]_{\theta=\hat{\theta}}. \tag{1.49}$$

TIC requires large sample sizes to be able to estimate the elements of matrix $J(\hat{\theta})$ and $I(\hat{\theta})$. Burham(2002) showed that AIC is an approximation to TIC where $tr(J^{-1}(\hat{\theta})I(\hat{\theta}))$ $\approx p$. Therefore, in many practice cases, we rather use AIC and AICc because of their convenience in computing and effectiveness in many applications.

1.7 Corrected AIC, AICc

1.7.1 Overfitting and Underfitting

Overfitting and underfitting are two common phenomena that result in bad performance of selecting models. The causes of these phenomena might be complicated. If the selected model is either too large or contains more variables than the model closest to the true model, overfitting happens. The opposite scenario is called underfitting. When a model is underfitted, it has too few variables and does not perform well compared to the true model. We want to avoid both overfitting and underfitting because both of them can lead to problems when we predict a model. Figure 3 illustrates overfitting and underfitting.

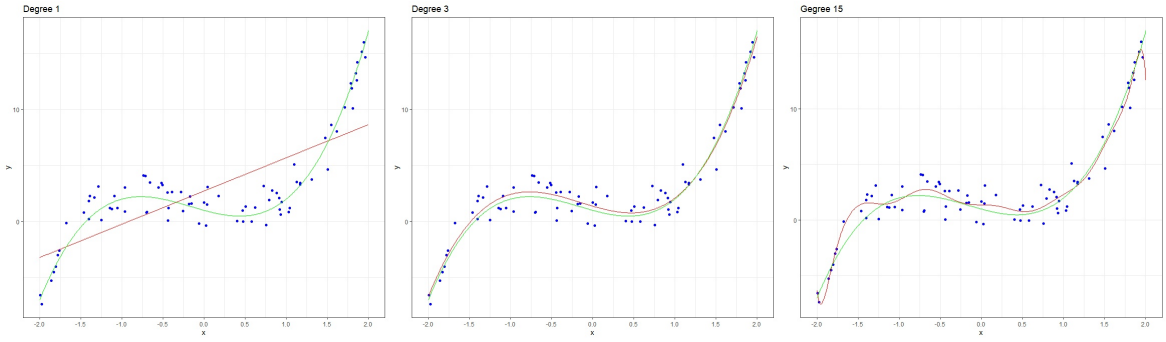

Figure 3: Illustration of overfitting and underfitting on a simple regression case–polynomial. Blue dots are data points that are generated from polynomial of degree 3 (green curve), and model fits are shown as red curve.

Green model is a 3-degree function that fits the data well and is considered as the true model. The red model in the first image is an example of underfitted, while the

model found is too simple with the degree of 2. On the contrary, the red model in the third image fits the data closely and this result in a high explanatory power. Howerver, it is so complicated that it leads to a higher error on prediction with new data samples. An overfitted model may be unreliable if our data contains extraneous variables or too much noise. In contrast, an underfitting model may lead to a poor predictive ability of a model due to a lack of detail in it. A good approach or a criterion is preferable if it can balance the tendencies to overfit and underfit.

1.7.2 The derivation of AICc

If there are too many parameters with respect to the sample size, AIC may perform poorly and the penalty will be negligible. It is worth noting that the more the sample size increases, the more complex the model AIC selects. This happens because the maximum log-likelihood will increase linearly when sample size n increases while the penalty term depends on the number of parameters p. When the dimension p of the model is large relative to n, or when n is small, for any p, there is a small-sample (second-order bias) correction which is termed as AIC_c. AIC_c was introduced in 1989 by Hurvich and Tsai ([14]).

They defined AIC_c as

$$AIC_c = -2\log[L(\hat{\theta}|X)] + 2p(\frac{n}{n-p-1}). \tag{1.50}$$

We can rewrite it as

$$AIC_c = -2\log[L(\hat{\theta}|X)] + 2p + \frac{2p(p+1)}{n-p-1}. \tag{1.51}$$

This result is equivalent to

$$AIC_c = AIC + \frac{2p(p+1)}{n-p-1}, \tag{1.52}$$

where the number of parameters is p in the model and n–the sample size. AIC_c has the bias-correction term more complex than with the standard version of AIC, with less chance of overfitting the model. However, if the sample size n is large with respect to p this additional bias-correction is negligible and AIC is sufficient. Burnham and Anderson (2002) advocated AIC_c should be used in particular when the ratio $\frac{n}{p} < 40$ for the model with the largest value of p. Additionally, that because AIC_c converges to AIC, as n gets large with p fixed, so AIC_c should be used in practice.

1.8 AIC differences, $\Delta_i(AIC)$ and Akaike weights, $\omega_i(AIC)$

1.8.1 AIC differences, $\Delta_i(AIC)$

The appearance of a number of models and their respective AIC scores is one of the key goals of measuring AIC. We may calculate how much better the best approximation model is relative to the next best model by comparing the multiple models. The easiest way to do this is to measure the difference. For the i-th model, the differences in AIC with respect to the AIC of the best candidate model can be expressed as

$$\Delta_i(AIC) = AIC_i - AIC_{min}, \tag{1.53}$$

where AIC_i is the AIC of the i-th model and AIC_{min} is the lowest one obtained among the set of models examined.

There are some rules in practise outlined by [6]

1. Models with $\Delta_i(AIC) > 10$ have essentially no support;
2. if $4 < \Delta_i(AIC) < 7$, then there is considerably less support for the i-th model;
3. if $\Delta_i(AIC) < 2$, then there is substantial support for the i-th model, and the suggestion that it is a proper representation is highly probable.

The best estimated model has $\Delta_i(AIC) = \Delta_{min}(AIC) = 0$. There always exists at least one best AIC estimated model in the set of all candidate models. The $\Delta_i(AIC)$ values also allow for ranking of models within the choice set. The simple methodological rule is we would like to choose the model with $\Delta_i(AIC) = \Delta_{min}(AIC) = 0$. In the next section, we will discuss a refinement of this rule.

1.8.2 Akaike weights, $\omega_i(AIC)$

From the differences in AIC $\Delta_i(AIC)$, we can then obtain an estimate of the relative likelihood given both the data and model. The likelihood of model M_i, given the data X can be simply computed by the transform

$$L(M_i|data) \propto \exp(-\frac{1}{2}\Delta_i(AIC)), \tag{1.54}$$

where $\propto$ stands for "is proportional to".

To obtain Akaike weights $\omega_i(AIC)$, following [6] we normalize the relative likelihoods.

In other words, we calculate the relative likelihood for each model of the set K models then devide it by the sum of these values across all K models

$$\omega_i(AIC) = \frac{\exp -\frac{1}{2}\Delta_i(AIC)}{\sum_{k=1}^{K} \exp -\frac{1}{2}\Delta_k(AIC)}, \tag{1.55}$$

so that the Akaike weight is a value between 0 and 1, with $\sum_{i=1}^{K} \omega_i(AIC) = 1$. Note that the Akaike weight depends on the sampling variability; therefore, given a different sample will most likely generate a different set of weights for the model in the candidate set. Burnham and Anderson (2002) refers to these weights as "model probabilities" or "the weight of evidence in favour of model i". Akaike weights ratios are equal to relative likelihood ratios (i.e., for a pair of models i and j, $L(g_i|x)/L(g_j|x) = \frac{\omega_i}{\omega_j}$), which are in the AIC literature taken to "... represent the evidence about fitted models as to which is better in a K-L information sense" ([6]-pp 78).

2 The Bayesian information (BIC)

2.1 BIC and its Bayesian Motivation

The Bayesian information criterion (BIC), proposed by Schwarz in the 1978 paper–[22] and hence also referred to as the Schwarz's information criterion (SIC), is another popular method for model selection but set within a Bayesian context. The BIC has a theoretical motivation in Bayesian statistical analysis which we will introduce in this section. The general idea of this section is mostly based on [16].

Let's suppose

1. $M_1, M_2, \ldots, M_k$ are k candidate models.
2. Vector $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ as notation for n observations is given.
3. $\theta_i \in \Theta_i \subset \mathbb{R}^{k_i}$ is the k_i-dimensional parameter of model M_i and θ_i belongs to the parameter space Θ_i.

Each candidate model M_i can be characterized by

1. A parametric distribution $g(\mathbf{x}|M_i, \theta_i)$, also known as the likelihood of the data (sometimes denoted by $L(\theta_i|M_i)$), given the i-th model and its parameter θ_i.
2. The prior distribution $\pi(\theta_i|M_i)$ of the parameter θ_i given model M_i. The prior distribution indicates the density of parameter θ_i before any data is observed.

Specifically, the marginal probability of the model M_i can be computed as

$$p(\mathbf{x}|M_i) = \int_{\Theta_i} g(\mathbf{x}|M_i, \theta_i).\pi(\theta_i|M_i)d\theta_i. \tag{2.1}$$

We can consider $p(\mathbf{x}|M_i)$ as the likelihood of the model M_i or the marginal likelihood of the data $\mathbf{x}$. The BIC is characterized by posterior probability of the model which is evaluated (see also [16]). A Bayesian procedure chooses the model, which is a posteriori most likely, if different models are feasible. The principle is we calculate the posterior of each model then choose the model with the biggest posterior probability. The posterior probability of a model is the conditional probability that calculates the probability of an unknown data being generated by the model when previous data are observed.

According to Bayes' theorem, the posterior probability of model M_i is provided as

$$P(M_i|\mathbf{x}) = \frac{p(\mathbf{x}|M_i).P(M_i)}{p(\mathbf{x})}, \quad i = 1, \ldots, k, \tag{2.2}$$

where

1. $P(M_i)$ is the prior probability of model M_i;

2. $p(\mathbf{x})$ is the unconditional likelihood of the data and obtained as

$$p(\mathbf{x}) = \sum_{j=1}^{k} p(\mathbf{x}|M_j)P(M_j), \tag{2.3}$$

and $p(\mathbf{x})$ is constant across k models.

As previously discussed, under the Bayes paradigm, the model with the highest posterior probability is selected. In the comparison of posterior probabilities $P(M_i|\mathbf{x})$ across different models, the model that maximizes the number of $p(\mathbf{x}|M_i).P(M_i)$ must be selected, since $p(\mathbf{x})$ is constant. If we further assume that in all candidate models, the prior probabilities $P(M_i)$ are equal then the crucial aspect is to maximizes the marginal likelihood $p(\mathbf{x}|M_i)$.

2.2 Laplace Approximation of High Dimensional Integrals

To approximate the marginal likelihood of data, we use Laplace's method for integrals. In this section, we introduce the basic theory of Laplace approximation method.

Let's assume we want to approximate this integral

$$\int e^{nf(\mathbf{x})} d\mathbf{x}. \tag{2.4}$$

The function is just any function possible, so in general to solve this integral is impossible. The question is what would we do in that case if we still want to get some information about this integral's behavior. To do it we perform a Taylor expansion of $f(\mathbf{x})$ around $\mathbf{x}_0$ - a unique global maximum of function $f(\mathbf{x})$ ($f(\mathbf{x}_0) > f(\mathbf{x})$):

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0)^T - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T J_f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + ..., \tag{2.5}$$

where

$$J_f(\mathbf{x}_0) = -\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T}\Big|_{\mathbf{x}=\mathbf{x}_0}, \tag{2.6}$$

is the Jacobian of the transformation.

The first derivative from (2.5) equals 0 because $\mathbf{x}_0$ is the unique global maximum of function $f(\mathbf{x})$. Substituting the (2.5) into (2.4), the integral yields an asymptotic expansion in the form

$$\int e^{\{n.[f(\mathbf{x}_0)-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T J_f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)+\ldots)]\}} d(\mathbf{x}_0) \approx e^{nf(\mathbf{x}_0)} \int e^{\{-\frac{n}{2}(\mathbf{x}-\mathbf{x}_0)^T J_f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)\}} d\mathbf{x}_0. \tag{2.7}$$

We get a multivariate Gaussian integral

$$\int e^{\{-\frac{n}{2}(\mathbf{x}-\mathbf{x}_0)^T J_f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)\}} d\mathbf{x}_0 = \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}} |J_f(\mathbf{x}_0)|^{\frac{1}{2}}}, \tag{2.8}$$

where p is the dimension of $\mathbf{x}$ (see [16]).

Therefore, we obtain

$$\int e^{nf(\mathbf{x})} dx \approx e^{nf(\mathbf{x}_0)} \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}} |J_f(\mathbf{x}_0)|^{\frac{1}{2}}}. \tag{2.9}$$

This expansion is accurate to order $O(\frac{1}{n})$, because we only consider the second order terms of the Laplace approximation.

Besides, we can state an approximation of $\int e^{nf(\mathbf{x})} g(\mathbf{x}) d\mathbf{x}$

$$\int e^{nf(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = \int e^{nf(\mathbf{x})+n\frac{1}{n}\log g(\mathbf{x})} d\mathbf{x} = \int e^{nh(\mathbf{x})} d\mathbf{x}, \tag{2.10}$$

where $h(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{n}\log g(\mathbf{x})$.

Using the Laplace approximation method for $\int e^{nh(\mathbf{x})} d\mathbf{x}$, then we can obtain

$$\int e^{nf(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} \approx e^{nf(\mathbf{x}_0)} \frac{(2\pi)^{\frac{p}{2}} g(\mathbf{x}_0)}{n^{\frac{p}{2}} |J_f(\mathbf{x}_0)|^{\frac{1}{2}}}. \tag{2.11}$$

2.3 Derivation of the BIC and BIC_{exact}

We would like to get an approximation of the marginal likelihood $p(\mathbf{x}_n|M_i)$. In this section, we will drop all the notations related to the model M_i. The marginal likelihood can be written as

$$p(\mathbf{x}) = \int_\Theta L(\theta)\pi(\theta) d\theta \tag{2.12}$$

Rewrite equation (2.12) as

$$p(\mathbf{x}) = \int_{\Theta} e^{\log L(\theta)} \pi(\theta) d\theta = \int_{\Theta} e^{l(\theta)} \pi(\theta) d\theta, \tag{2.13}$$

where $l(\theta)$ is the log-likelihood function $l(\theta) = \log L(\theta) = \log f(\mathbf{x}|\theta)$.

To apply the Laplace approximation, we wish to have $p(\mathbf{x})$ under the form

$$p(\mathbf{x}) = \int e^{nh(\theta)} K d\theta,$$

where K is a constant and $h(\theta) = \frac{l(\theta)}{n}$.

Let's assume $l(\theta)$ has the maximum likelihood estimator $\hat{\theta}$. The Taylor expansion of function $l(\theta)$ around $\hat{\theta}$ provides

$$l(\theta) \approx l(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^T J_l(\hat{\theta})(\theta - \hat{\theta}), \tag{2.14}$$

where $J_l(\hat{\theta})$ is the Hessian matrix also known as the Jacobian matrix of the gradient of the function $l(\theta)$.

Rewrite (2.14) as

$$l(\theta) \approx l(\hat{\theta}) - \frac{n}{2}(\theta - \hat{\theta})^T J(\hat{\theta})(\theta - \hat{\theta}), \tag{2.15}$$

where $J(\hat{\theta}) = \frac{J_l(\hat{\theta})}{n}$ is the Fisher information matrix.

Going back to equation (2.13) then substituting (2.15) into, we obtain

$$\begin{aligned} p(\mathbf{x}_n) &= \int_{\Theta} e^{l(\theta)} \pi(\theta) d\theta \\ &\approx \int_{\Theta} [e^{l(\hat{\theta}) - \frac{n}{2}(\theta - \hat{\theta})^T J(\hat{\theta})(\theta - \hat{\theta})}] \pi(\theta) d\theta. \end{aligned}$$

An approximation of $p(\mathbf{x}_n)$ is found as

$$\begin{aligned} p(\mathbf{x}_n) &\approx \int_{\Theta} e^{l(\hat{\theta}) - \frac{n}{2}(\theta - \hat{\theta})^T J(\hat{\theta})(\theta - \hat{\theta})} \pi(\theta) d\theta \\ &= e^{l(\hat{\theta})} \int_{\Theta} e^{-\frac{n}{2}(\theta - \hat{\theta})^T J(\hat{\theta})(\theta - \hat{\theta})} \pi(\theta) d\theta. \end{aligned}$$

After applying the consequence of Laplace approximation method in equation (2.11), when the observation n is large, the marginal likelihood can be approximated as

$$p(\mathbf{x}_n) \approx e^{l(\hat{\theta})} \pi(\hat{\theta}) \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}} |J(\hat{\theta})|^{\frac{1}{2}}}. \tag{2.16}$$

Now we take the logarithm and multiply by -2 and denote by BIC_{exact}. In other words, denote $-2 \log p(\mathbf{x}_n) = BIC_{exact}$, the BIC_{exact} then yields

$$BIC_{exact} \approx -2l(\hat{\theta}) - 2 \log \pi(\hat{\theta}) - p \log(2\pi) + p \log n + \log |J(\hat{\theta})|$$

$$= -2l(\hat{\theta}) + p\log n - 2\log\pi(\hat{\theta}) - p\log(2\pi) + \log|J(\hat{\theta})|.$$

The two dominant terms of BIC_{exact} is $-2l(\hat{\theta})$ and $p\log n$ which are accurate respectively to orders $O(n)$ and $O(\log n)$. The others are accurate to $O(1)$. BIC is found by ignoring these lower-order terms of BIC_{exact}.

Definition 2.1. *The BIC is formally defined as*

$$BIC = -2\log(L(X|\hat{\theta})) + p\log n, \tag{2.17}$$

where $\log(L(X|\hat{\theta}))$ *represents the log-likelihood function at maximum likelihood estimator point* $\hat{\theta}$ *given the observed data* X *and* p *is the dimension of the parameter* θ*. The parameter* p *can be counted as the number of estimable parameters in the model ([8], [16]).*

2.4 A discussion on distinctions between AIC and BIC

A Web of Science search showed that the two most common indicators of parsimony were the AIC and BIC for ecological publications using databases from 1993 to 2013. Specifically, 84% used AIC, 14% used BIC, and just 2% used some other methodologies, for publications that used structured multi-model inference approaches ([1]). In this section, we compare some aspects that distinguish AIC and BIC. We concentrate on some mathematical contrast characteristics, such as mathematical motivations, derivation, efficiency, consistency, and parsimony.

2.4.1 About motivation and formulas

Both criteria are constructed through a comparison with likelihood ratio tests while models are nested. The forms of AIC and BIC are respectively introduced as (1.37) and (2.1):

$$AIC = -2\log(L(X|\hat{\theta})) + 2\text{length}(\theta).$$

$$BIC = -2\log(L(X|\hat{\theta})) + 2\text{length}(\theta)\log n.$$

While the formulas are simply analyzed, it is possible to misinterpret AIC and BIC as competing criteria intended to achieve the same goal. However, the initial question, which candidate models "best" suit the results, can be paraphrased in a variety of ways,

and until the question is more clearly stated, AIC and BIC are both answering different questions. By using the K-L information, Akaike (1973)–[2] tried to approximate the probability of a model when correcting for the bias introduced by maximum likelihood probability. On the other hand, BIC has a theoretical incentive in Bayesian statistical analysis, in particular the Bayes Factor. The approach of BIC is the Bayesian posterior probabilities (see the previous Section 3.1, 3.3).

2.4.2 Statistical Consistency and Asymptotic efficiency

There are two main problems in model selection when it comes to asymptotics: consistency and efficiency. If there is one estimator model that converges to the true model when the observations rise, it's linked to consistency. The criterion is weakly consistent if, with a likelihood that tends to be one as the sample size tends to be infinite, the criterion is capable of choosing the true model from the candidate models. Strong consistency is reached when the selection of the true model is almost certain. Often, models are considered without necessarily requiring the true model to be amongst the candidate models, and only current data be used. If instead, we are able to believe that there is a candidate model that is closest in K-L information to the true model ([8]); numerous sources (e.g., [29]) have stated that AIC is not a statistically consistent estimation. However, the BIC is statistically consistent.

As far as BIC is concerned, when the sample size n increases, the criterion selects the true model with a probability approaching 1. A powerful outcome of this nature requires a group of assumptions. In this case, some important assumptions are: (a) the true model is under consideration; (b) the true model dimension (denoted p_0) remains constant when n increases, and (c) the number of parameters in the true model is finite ([20]). The BIC is consistent as penalty $2\text{length}(\theta)\log n$ satisfies the assumption of consistency. Unlike the BIC, the AIC with a fixed term $2\text{length}(\theta)$ penalty can not pick a true model with a non-vanishing probability when n grows large even when the true model is under consideration. A reasonable explanation is that an information criterion must have a penalty approaching infinity when n approaches infinity to be consistent ([11]). The penalty of the AIC is a fixed term, which is not able to approach

infinity as n increases. Otherwise, AIC can be concerned as weak at consistency [2].

The consistency characteristic of BIC makes it robust in theory. Researchers may collect data, fit models, and compare models with BIC, knowing that when n is very big, BIC is dependent on in identifying the appropriate model with probability equals 1. However, this finding is unhelpful for real data, where the truth is unknown or as n increases, the form of the truth is not fixed, but changes accordingly. We will attempt to demonstrate the consistency characteristic of BIC through various simulations in the last chapter compared to AIC and AICc.

Significantly, while AIC does not attain the consistency, it is nevertheless capable of selecting a more efficient model compared to the BIC. We will consider another circumstance while one of the assumptions mentioned above is broken. In particular, when the true model is not included in the set of candidate models or does not exist, we tend to minimize the value of the loss function (e.g. root mean squared error of estimation). AIC is considered as a asymptotically efficient while BIC is not. In 1938, Shibata came up that in large samples, an information criterion is said to be asymptotically efficient if it chooses the model with minimum mean squared error deviation in [23]. The efficiency concerns finding a model that might yield a good inference, compared to the consistency which concerns finding a true model as correctly as possible. Shibata (1983) in [23] also pointed out that in other cases such as when the number of parameters in the true model is infinite or when the sample size n increases, the dimension of the true model also increases, AIC is considered as efficient and BIC is not. It is worth noting here that because AIC focuses on choosing the most efficient model and minimizing errors, AIC asymptotically overfits the true dimension, also known as overfitting, was mentioned in Section 1.7.1. We will attempt some simulations in Chapter 4 to observe these assertions.

[2]Theorem about weak consistency and strong consistency are formally written in Claeskens and Hjort (2008)–[8]

3 Information Criteria for selecting Linear Regression Models

3.1 Maximum Likelihood for Linear Regression

Suppose we have a response vector $Y = (y_1, ..., y_n)^T$ and vectors $X_1, X_2, ..., X_n$ are n explanatory observations (or n covariate vectors). Each $X_i^T = (x_{i1}, ..., x_{ip})$, whereby index i goes from 1 to n. The general linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{3.1}$$

where the residual terms ϵ_i are independently drawn from $N(0, \sigma^2)$.
Denote vector $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$, and $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$. Then ϵ is multivariate-normal distribution $N_n(\mathbf{0}_n, \sigma^2 I_n)$.
We also need to denote covariate matrix $\bar{X}$ as a $n \times (p+1)$ matrix

$$\bar{X} = \begin{pmatrix} 1 & X_1^T \\ 1 & X_2^T \\ \vdots & \vdots \\ 1 & X_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}. \tag{3.2}$$

Equation (3.1) is written in matrix form as

$$Y = \bar{X}\beta + \epsilon. \tag{3.3}$$

Therefore the conditional distribution of the response variables y_i given the explanatory varables $X_i = (x_{i1}, \cdots, x_{ip})^T$ is expressed as

$$p(y_i | x_{i1}, \cdots, x_{ip}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}}{\sigma})^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2}. \tag{3.4}$$

Note that the response variables y_i are independent because the epsilons ϵ_i were assumed to be independent. Thus, the likelihood function of the regression model is

$$L(\beta, \sigma^2) = \prod_{i=1}^n p(y_i | x_{i1}, \cdots, x_{ip}) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2}. \tag{3.5}$$

Then, the log-likelihood function is

$$l(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2. \tag{3.6}$$

Maximization the log-likelihood function with respect to β is equivalent to minimization of

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 = ||Y - \bar{X}\beta||^2 = SSE(\beta). \tag{3.7}$$

Then the maximum likelihood estimator of regression coefficients $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p)^T$ of the vector regression coeficient $\beta = (\beta_0, ..., \beta_p)^T$ is obtained as the solution to the system of linear equations

$$\bar{X}^T\bar{X}\beta = \bar{X}^T Y. \tag{3.8}$$

Thus, the solution can be written as

$$\hat{\beta} = (\bar{X}^T\bar{X})^{-1}\bar{X}^T Y. \tag{3.9}$$

We assume that $\bar{X}$ is matrix $n \times (p+1)$ with full rank $(p+1)$ to make $\bar{X}^T\bar{X}$ be an invertible $(p+1) \times (p+1)$ matrix.

The maximum likelihood estimator $\hat{\sigma}^2$ of σ^2 is the maximizer of $l(\hat{\beta}, \sigma^2)$. Therefore, we have

$$\hat{\sigma}^2 = \frac{1}{n}SSE(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} res_i^2 = \frac{1}{n}||res||^2, \tag{3.10}$$

where the residuals are $res_i = y_i - \bar{X}_i^T\hat{\beta}$ and $\bar{X}_i^T$ is the i-th row of matrix $\bar{X}$.

The maximum log-likelihood is given by plugging in $\hat{\sigma}^2$ and $\hat{\beta}$

$$l_{\max} = l(\hat{\beta}, \hat{\sigma}^2) = l(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p, \hat{\sigma}^2) = -\frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2} - \frac{n}{2}\log(2\pi). \tag{3.11}$$

3.2 Derivation of the Information Criteria Formulas for Regression Models

1. The form of AIC.

   As we have known AIC is found as an estimate of the Kullback-Leibler divergence and moreover, it is asymptotically unbiased for K-L. In general, AIC can be computed as

   $$AIC = -2 \times \text{log likelihood} + 2 \times \text{number of parameters},$$

   (see equation (1.37)).

Using the maximum log likelihood l_{max} found in (3.11), we have

$$-2l_{max} = -2l(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2) = n(\log(2\pi) + 1 + \log\hat{\sigma}^2). \tag{3.12}$$

There are $(p+2)$ free parameters contained in the model regression, then the AIC for this model is

$$\begin{aligned} AIC_{reg} &= n(\log(2\pi) + 1 + \log\hat{\sigma}^2) + 2(p+2) \\ &= n\log(2\pi) + n + n\log\hat{\sigma}^2 + 2p + 2. \end{aligned} \tag{3.13}$$

Again, across all candidate models, a model is selected which has the AIC minimum for different possible combinations of the explanatory variables. Then the constants $n\log(2\pi)$, n, and 2 are not essential. We can ignore them and then scale AIC by $\frac{1}{n}$

$$AIC_{reg} = \log\hat{\sigma}^2 + \frac{2p}{n}. \tag{3.14}$$

2. The form of AICc.

   We next consider the AICc which is intended to correct the overfitting tendencies of AIC when the sample size is small. AICc adjusts the second term in the form of AIC to be $2(p+2)(\frac{n}{n-p+1})$, yielding

   $$AICc_{reg} = n\log\hat{\sigma}^2 + (\frac{2n(p+2)}{n-p+1}).$$

   Similarly, AICc is scaled by $\frac{1}{n}$ to express as a rate

   $$AICc_{reg} = \log\hat{\sigma}^2 + (\frac{2(p+2)}{n-p+1}). \tag{3.15}$$

3. The form of BIC.

   In BIC the second term $2p$ is replaced by $p\log n$, the penalty for overfitting is much stronger than AIC

   $$BIC_{reg} = \log\hat{\sigma}^2 + \frac{p\log n}{n}. \tag{3.16}$$

3.3 Some other criteria for evaluating the regression model

In this section, we will list some criteria that can be used to test the quality of the regression model. All of these criteria aim to measure the difference between actual n observations y_i and n corresponding predicted values $\hat{y}_i$, where $i = 1, \ldots, n$.

3.3.1 Coefficient of determination, R-squared and Adjusted R-squared

R-squared is a metric that indicates how well data fits into a regression model. It calculates how much of the variation in the dependent variable Y can be explained by the independent variables X. The value of R-squared is between 0 and 1 and independent of scale of Y. R-squared is calculated using the formula:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \tag{3.17}$$

where TSS is total sum of squares, $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ ($\bar{y}$ denotes the observation mean) and RSS is the residual sum of squares, $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

R-squared has a tendency to increase when we add predictors to the model, regardless of how the model is effective. Because of that, the more variables we add to the model, the better fit it appears to be. This may not be accurate. The proportion of variance explained by just the independent factors that impact the dependent variable is calculated using adjusted R-squared (R^2_{adj}). Adding independent variables that do not match the model will be penalized by the adjusted R-Squared. Adjusted R-squared will always be smaller than or equal to R-squared. The adjusted R-Squared formula is as follows:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}, \tag{3.18}$$

where p is the number of independent variables.

3.3.2 Criteria based on absolute difference

In this group we have:

1. Mean absolute error - MAE

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}. \tag{3.19}$$

2. Sum Absolute Error - SAE

$$\text{SAE} = \sum_{i=1}^{n}(|\hat{y}_i - y_i|). \tag{3.20}$$

3. Mean absolute percentage error - MAPE

$$\text{MAPE} = \frac{\sum_{i=1}^{n}(|\frac{\hat{y}_i - y_i}{\hat{y}_i}|)}{n}. \tag{3.21}$$

All three criteria are easy to understand. The first two criteria measure the difference between the actual y_i and the predicted $\hat{y}_i$. The absolute value is used to avoid errors in cases where the model leads to errors over than 0 and less than 0 and cancels each other out. The third criterion MAPE measures the difference as a percentage, used for cases where the outcome variable has too low or too high units.

3.3.3 Criteria based on square of error

The criteria in this group do not use the absolute values, but are based on the square of the error, we have in turn:

1. Mean Square Error - MSE

   In statistics, the mean squared error (MSE) of an estimate is the average of the squares deviation between the predicted values with the actual data observation. MSE is a function of lack of risk, corresponding to the expected value of squared error loss. Errors may occur due to randomness, or the model found is not really good. When discussing the mean error of a statistical model, it is difficult to determine how much error is due to the model and how much is due to randomness. MSE simply refers to the mean of the squared difference between the predicted parameter and the observed parameter to evaluate the quality of an estimator as the following formula:

   $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \tag{3.22}$$

   where y_i are actual observations and $\hat{y}_i$ are predicted values. The MSE is extremely sensitive to outliers since all values are squared. Strong outliers among the prediction errors are indicated by an MSE that is significantly higher than the MAE.

2. Root mean squared error - RMSE

   Although R-squared is said to be the standard unit to measure the lack of fit of

a linear model, it does not guarantee high reliability. Some researchers do not accept R-squared but accept the criteria with higher reliability RMSE. Similar to the standard deviation, when taking the square root of the MSE, we get the root-mean-square error (RMSE).

$$\text{RMSE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{3.23}$$

RMSE is a useful and standard measure of how well the model predicts the response and is resistant to outliers.

3. Sum of squared error - SSE

$$\text{SSE} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2. \tag{3.24}$$

Like MAE, MedAE, SAE, and MAPE, the lower value of SSE, RMSE, MSE indicates the better the model fit.

4 Simulations and Case study

4.1 Bodyfat data

4.1.1 About data and variables

The data include estimates of percentage of body fat calculated by underwater weighing and other body circumference measures for 252 males which is supplied by Behnke and McArdle–[15] p.113. We have known that it is expensive to accurately measure of body fat, therefore we need an easy method of estimating body fat that is not inconvenient. The data consist of 13 independent variables which are denoted from X_1 to X_{13} and one dependent variable denoted by Y. Here are more details about the variables

- Y - $1/D$, where D is the density determined from underwater weighing
- X_1 - Age (years)
- X_2 - Weight (lbs)
- X_3 - Height (inches)
- X_4 - Neck circumference (cm)
- X_5 - Chest circumference (cm)
- X_6 - Abdomen circumference (cm)
- X_7 - Hip circumference (cm)
- X_8 - Thigh circumference (cm)
- X_9 - Knee circumference (cm)
- X_{10} - Ankle circumference (cm)
- X_{11} - Biceps (extended) circumference (cm)
- X_{12} - Forearm circumference (cm)
- X_{13} - Wrist circumference (cm)

4.1.2 AIC and BIC for regression model in R

We consider the multiple regression of full model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{13} X_{13} + \epsilon.$$

We start by running a model that includes all predictors (often called full model) so we can see how they are interacting. In R, we can use the general formula `AIC(object,..., k=2)` to compute the classical AIC and $k = \log(n)$ for computing BIC.

```
> # 2. Finding AIC, BIC manual + function provided in R for full model
> bodyfat.lmfull <- lm(Y~., data = df)
> bodyfat.ssefull <- sum(resid(bodyfat.lmfull)^2)
> p <- 14
> n <- nrow(df)
> AIC.lmfull <- n + n*log(2*pi) + n*log(bodyfat.ssefull/n) + 2*(p+1)
> AIC.lmfull
[1] -1648.125
> AIC(bodyfat.lmfull, k=2)
[1] -1648.125
>
> BIC.lmfull <- n + n*log(2*pi) + n*log(bodyfat.ssefull/n) + log(n)*(p+1)
> BIC.lmfull
[1] -1595.184
> AIC(bodyfat.lmfull, k=log(n))
[1] -1595.184
```

Figure 4: Calculating AIC and BIC for full model linear regression for bodyfat data.

From Figure 4, we can see that the results when manual computing AIC and BIC using the two formulas 3.13, 3.16 and two functions provided in R are the same.

4.1.3 Fitting a Multiple Linear Regression Model using Backward Elimination Method

There are 13 potential predictors, then we have a set of $2^{13} - 1 = 8191$ possible regression models to be considered. It is quite complicated when we calculate AIC and BIC values for all 8191 models. Furthermore, it is impossible if the number of predictors is large. The stepwise procedure is typically used on much larger data sets for which it is not feasible to attempt to fit all of the possible regression models. The three main approaches for stepwise regression are forward selection, backward elimination, and bidirectional elimination:

- The forward selection, which involves starting with a null model that contains no variables and then adds the most significant variables. We repeat this process

until none of these significantly improves the model or a pre-specified stopping rule is satisfied or the model runs out of variables.

- The backward elimination starts with a full model that contains all variables under consideration and continues with removing the least significant variable to test its importance relative to overall results. The process is repeated until a pre-specified stopping rule is reached or until no further variable is left in the model.

- Bidirectional elimination, which is a combination of the two above approaches. At each stage, we test for variables to be included or eliminated.

In R the built-in *step*() function from the stats package performs stepwise selection, which uses the following syntax: `step(object, scope, scale = 0, direction = c("both", "backward", "forward"), trace = 1, keep = NULL, steps = 1000, k = 2, ...)`. The argument k of function `step()` defaults to 2, which gives the AIC and if we set it to $k = \log(n)$, the function considers the BIC. The function has an option called `direction`, which can have the "forward", "backward", and "both" values.

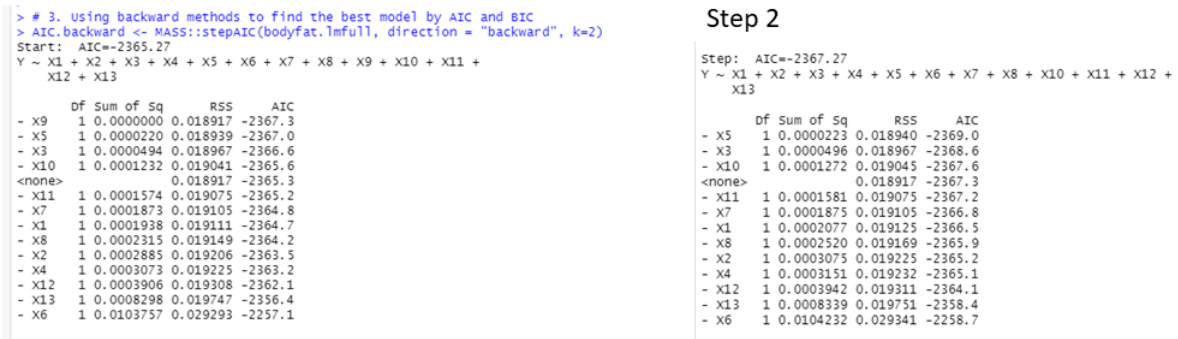

Figure 5: Function stepAIC() using "backward elimination". The first two steps for stepwise using AIC starting with full model.

In our thesis, we determine AIC and BIC for backward elimination to select the best subset. The function begins with the full model. As we can see in the output (Figure 5), X_9 has lowest AIC value which means the amount of information loss by

removing X_9 is minimum according to AIC criteria. Then for the next step, the variable X_9 is removed and the remaining set of variables is used to run the stepAIC function. We can see the AIC value at the first step is Start: $AIC = -2365.27$ and then it improves in the next step to step: $AIC = -2367.27$. The stopping rule is reached if the model with remaining variables has the lowest AIC. At this point, backward elimination will be terminate and return the model at current's state.

```
> summary(AIC.backward)

Call:
lm(formula = Y ~ X1 + X2 + X4 + X6 + X7 + X8 + X12 + X13, data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.032756 -0.005918 -0.000144  0.006199  0.020190

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.644e-01  2.436e-02  35.489  < 2e-16 ***
X1           1.107e-04  6.399e-05   1.731  0.08479 .
X2          -2.124e-04  8.298e-05  -2.560  0.01108 *
X4          -9.543e-04  4.670e-04  -2.043  0.04210 *
X6           1.989e-03  1.496e-04  13.298  < 2e-16 ***
X7          -4.351e-04  2.879e-04  -1.511  0.13203
X8           6.881e-04  2.683e-04   2.565  0.01093 *
X12          1.068e-03  3.874e-04   2.757  0.00628 **
X13         -3.268e-03  1.059e-03  -3.085  0.00227 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008903 on 243 degrees of freedom
Multiple R-squared:  0.7377,    Adjusted R-squared:  0.7291
F-statistic: 85.44 on 8 and 243 DF,  p-value: < 2.2e-16
```

Figure 6: Bodyfat data - The model's result of stepwise using AIC starting with the full model.

In the case using backward selection and AIC as a threshold (Figure 6), we find

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{12} X_{12} + \beta_{13} X_{13}$$

is the best selected model. Figure 7 gives the the best explanatory model in case we use BIC as selection criterion

$$E(Y) = \beta_0 + +\beta_2 X_2 + \beta_6 X_6 + \beta_{12} X_{12} + \beta_{13} X_{13}.$$

```
> summary(BIC.backward)

Call:
lm(formula = Y ~ X2 + X6 + X12 + X13, data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.033110 -0.006231 -0.000328  0.006394  0.018501

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.422e-01  1.505e-02  55.944  < 2e-16 ***
X2          -2.872e-04  5.143e-05  -5.585 6.14e-08 ***
X6           2.056e-03  1.165e-04  17.645  < 2e-16 ***
X12          1.019e-03  3.775e-04   2.699 0.007438 **
X13         -3.436e-03  9.198e-04  -3.735 0.000233 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009024 on 247 degrees of freedom
Multiple R-squared:  0.7261,    Adjusted R-squared:  0.7217
F-statistic: 163.7 on 4 and 247 DF,  p-value: < 2.2e-16
```

Figure 7: Bodyfat data - The model's result of stepwise using BIC starting with the full model.

One thing we can interpret is that AIC has a tendency to include more variables than BIC. At this point, we do not make any judgments about the accuracy of the two resulting models. The function does not compute the AIC and BIC for all potential models, but rather uses a search approach that compares models progressively, but with the advantage that no dubious p-values are used. One more disadvantage of the stepwise function is its unstable property, especially when the data has a small sample size in comparison to the number of variables. Besides, stepwise has its limits when it does not consider the causal relationship between variables, for instance, a variable should be included in the model by expert opinion to control errors.

4.1.4 Simulation based on variance-covariance matrix

It would be very difficult to understand the relationship between each variable by simply staring at the raw data. We visualize a correlation matrix among 13 predictors to quickly explore the relationships among them. The **corrplot()** function in the `corrplot` package is used for our visualization. In looking at our correlation matrix (Figure 8), the correlation among 13 variables are strong and most of them are positive which indicates that they are strongly positively correlated.

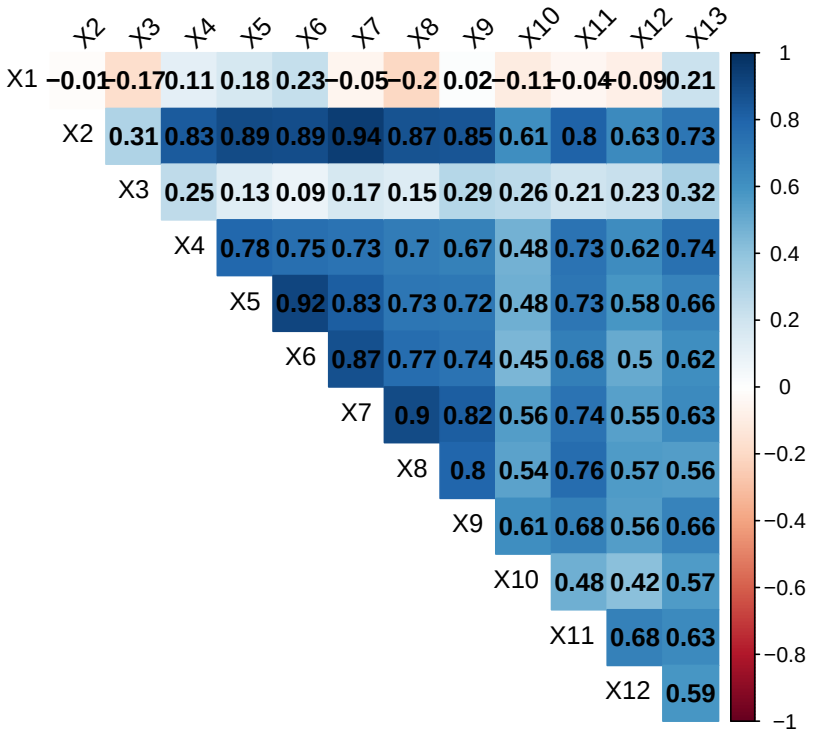

Figure 8: Correlation matrix among the 13 predictors of bodyfat dataset.

The data bodyfat has a full-rank design matrix. Hence for the next simulation, we carry out a small simulation study inspired by [7]. For that, we simulate data from a multivariate normal distribution in which the covariance matrix S of the variables is the covariance matrix of data bodyfat. We assume the covariates for the bodyfat data are stored in the matrix $X \in \mathbb{R}^{n\times p}$. Thus, the calculation for the covariance matrix can be expressed as

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X}^T). \tag{4.1}$$

For a given X, the dependent variable Y is generated following

$$Y = X\beta + \epsilon, \tag{4.2}$$

where β is the coefficients of full model using `lm(...)` funtion and ϵ is the random error, each value is independently generated from a standard normal distribution $N(0,1)$. For a given generated dataset, six datasets are randomly generated with number of samples increasing as the list $n = 2^i,\ i = 5,\ldots,10$.. For each dataset, we will divide the sample

into training and testing sets. The observations of the training set accounts for 70% of all the observations and is used to run a stepwise selection. We use the testing set to run a model with the selected variables to find the predicted $\hat{Y}$ by provided function `predict (object, ...)` in R, and then calculate the root-mean-square error (RMSE) between the "true value" Y with the predicted $\hat{Y}$.

The simulation is repeated 500 times for each dataset and displays the average RMSE in Table 1. It shows empirically some interesting results. When the sample size is

Sample size	AIC	BIC
32	0.0144	0.0141
64	0.0109	0.0103
128	0.0102	0.0101
256	0.00974	0.00972
512	0.01075	0.01091
1024	0.0102	0.0104

Table 1: The average of root mean square error for measuring the differences between model predicted by AIC or BIC selecting and the values observed. The results were estimated with 500 Monte Carlo runs.

small such as 32, 64, 128, and 256, BIC seems to be better performing than AIC. In case the sample size is larger such as 512, 1024, then AIC is more effective. This result is in agreement with the theory we gave in Section 1.7.2, where mentioned AIC works effectively when the sample size is large compared to the number of variables.

4.2 Overfitting and Underfitting

4.2.1 An intuitive approach

We perform a simple simulation to obtain intuitive observations of AIC and BIC about overfitting and underfitting features. Our conclusions on theoretical parts reveal that AIC has the lowest rate of underfitting but is frequently overfitting, whereas BIC

seldom overfits and often underfits. BIC chose the correct models in most situations, while AIC's accuracy was significantly lower due to its tendency to overfit.

Our data is generated from a polynomial of degree 3:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3 + \epsilon_i, \tag{4.3}$$

where y is the response variable, x the predictor and ϵ the random error.
More specific,

- List of coefficients $[\alpha_0, \alpha_1, \alpha_2, \alpha_3]$ is chosen as $[-2, -1, 3, 0.1]$.
- The sample size n is changed in range of list $[8, 32, 128, 512, 1024]$
- By each sample size n_i chosen, we generate n_i random predictors x that is distributed as a uniform distribution between -3 and 3.
- The random error "ϵ" is considered to be a standard normal distribution $N(0, 1)$ and independently generated.
- Six candidate models are used to fit the data. Starting from the polynomial of degree one and going up to 6th degree.
- $AIC()$ and $BIC()$ functions are used to find the "best" polynomial in terms of AIC and BIC.

From Figure 9, one can observe that both AIC and BIC choose the polynomial of degree six as the "best" model. RMSE values for predicted unseen data show that a polynomial with degree of six returns a very high value which means that model performs very badly. In this case, both metrics are inaccurate. A possible explanation for the bad performance of AIC and BIC could be that the sample size $n = 8$ is too small.

Degree	AIC	BIC	RMSE
1	61.55	61.79	8.95
2	29.01	29.33	**1.39**
3	26.71	27.11	1.44
4	28.10	28.57	1.93
5	29.15	29.71	13.91
6	**1.60**	**2.23**	241.60

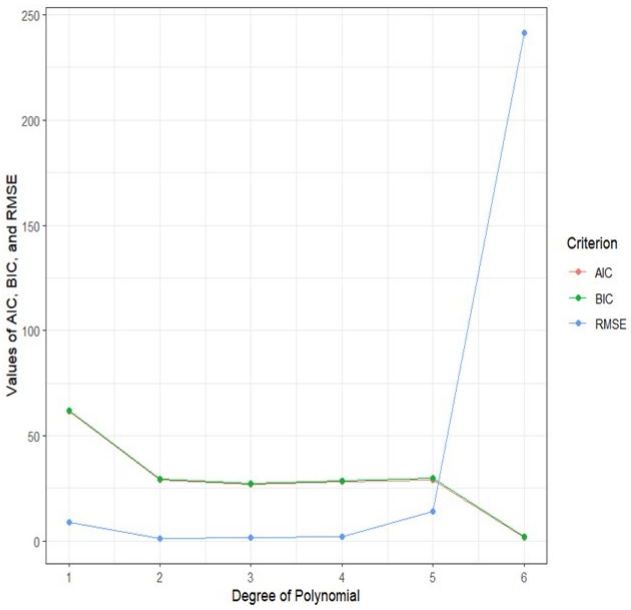

Figure 9: Values AIC and BIC on six candidate polynomials with sample size $n = 8$.

For a sample size of $n = 32$ (Figure 10), AIC and BIC select the "true" polynomial with a degree of four. However, when comparing the differences among polynomial of degree 5 and 6 to degree 4, these results are very similar, specially for AIC's case. We begin to catch symptoms of overfitting here.

Degree	AIC	BIC	RMSE
1	229.68	234.08	8.7
2	91.57	97.43	1.15
3	79.27	86.59	**0.96**
4	**72.74**	**81.54**	1.04
5	74.69	84.96	1.19
6	73.26	84.99	1.07

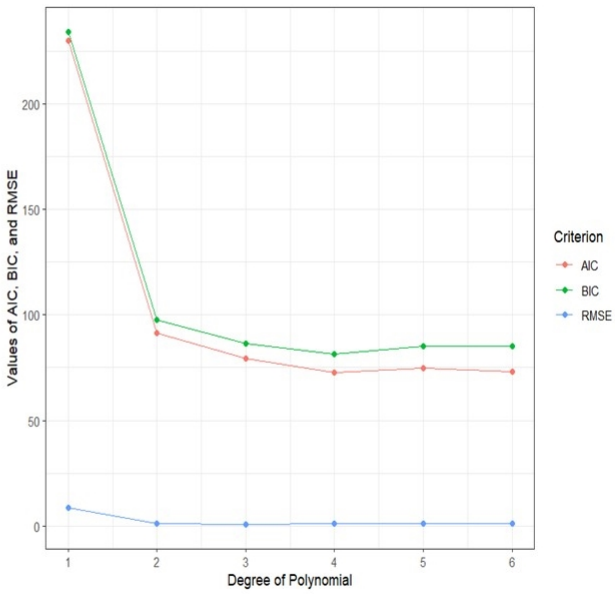

Figure 10: Values AIC and BIC on six candidate polynomials with sample size = 32.

Next simulation, the sample size increases to $n = 1024$ (Figure 11). We can see that both AIC and BIC select the "best" model is the candidate model with polynomial degree three. By observing the differences when using AIC, an insignificant difference between a model to the "best" model indicates that the model has a essential support in our data (see Section 1.8.1). Therefore, the polynomial of degree 3, 4, 5 and 6 should be under consideration.

Degree	AIC	BIC	RMSE
1	7167.56	7182.35	8.40
2	2992.58	3012.31	0.99
3	**2813.45**	**2838.12**	0.93
4	2814.96	2844.54	1.10
5	2816.67	2851.19	**0.81**
6	2816.68	2856.12	1.00

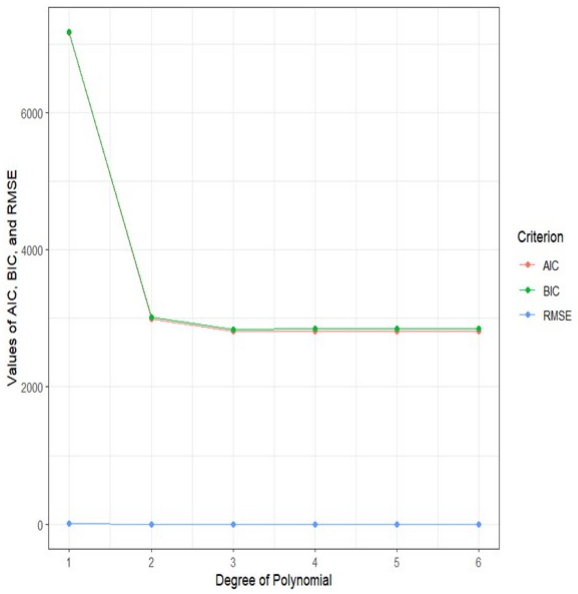

Figure 11: Values AIC and BIC on six candidate polynomials with sample size = 1024.

The results of these simulations are designed to provide an intuitive view and cannot return any conclusions because we utilize just one repetition for the simulation and the results produced are different after each simulation. From these above results, AIC "seems" to be overfitting as theoretical references. AIC sometimes chooses the wrong model, other times it "nearly" chooses the wrong one. Likewise, BIC gives a similar result as AIC. AIC and BIC return these similar values among polynomials of degree 3, 4, 5 and 6 could be explained as higer degree polynomials can approximate lower degree polynomials. Additionally, the "true" model is included in the set of all candidate models. Hence criteria would find the "true" model with a high precision.

4.2.2 A Comparison of AIC and BIC on linear regression model with multiple predictors

A caveat on the use of polynomial is that a polynomial has only one predictor x and the response variable y strongly depends on x, so AIC and BIC can effectively learn from the data and accurately deliver the true model. Therefore, we will perform a multiple linear regression in order to overcome weaknesses when using polynomials.

We consider a regression model:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{4.4}$$

where y is the response, X_1, X_2 are two predictors, and ϵ the random error.

A set of 200 observations is generated from model (4.4) as following:

- Coefficients $\beta_0, \beta_1, \beta_2$ are chosen as $[0, 2, 1]$.
- Random error $\epsilon \sim N(0, 1)$.
- Predictors $X_1, X_2 \sim N(0, 1)$.
- Each realization of X_1, X_2, ϵ is generated independently; and also that X_1, X_2 are independently generated.

We build a set of 200 regression models, in turn from the model with one predictor X_1 to model with 200 predictors $X_1, X_2, \ldots, X_{200}$; and similarly $X_3, \ldots, X_{200}$ are independently generated from $N(0, 1)$.

- $M_1 : y = \beta_0 + \beta_1 X_1 + \epsilon$.
- $M_2 : y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
  $\vdots$
- $M_{200} : y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{200} X_{200} + \epsilon$

Independently from each generating model, the data are generated 500 times by Monte Carlo simulations. Then we obtain 500 values of AIC and BIC. Figure 12 shows the average of 500 values for AIC and BIC.

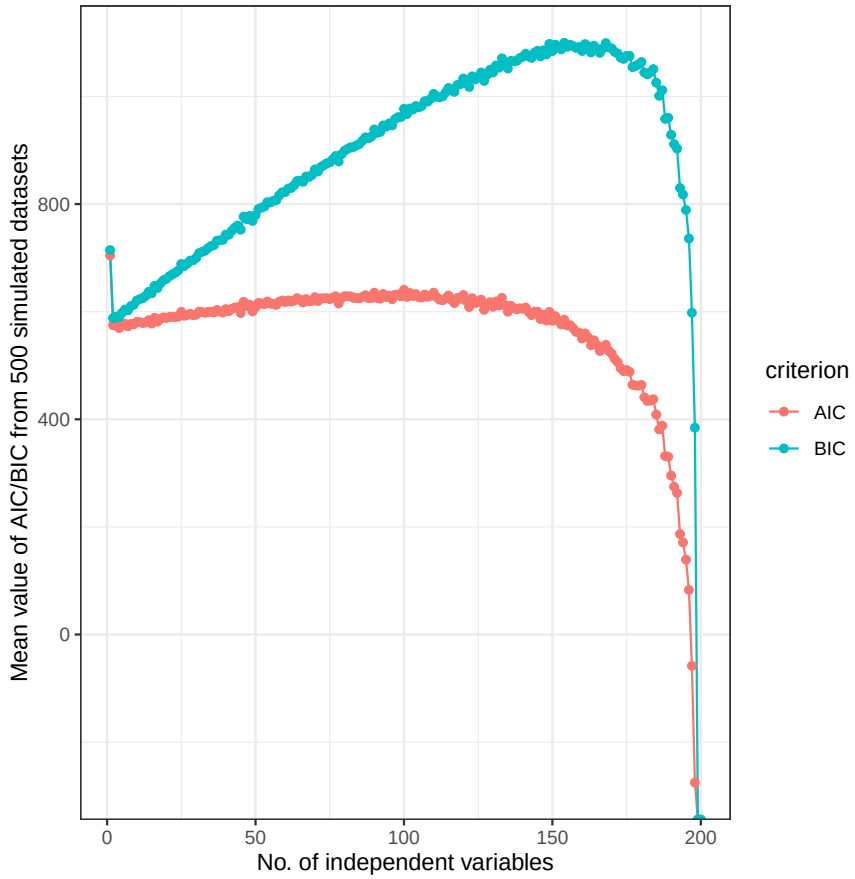

Figure 12: Average values of AIC and BIC on 200 linear regression models.

According to Figure 12, empirical evidence provides several interesting results. BIC agrees that the model with 2 predictors is the "best" model because it returns the smallest value. When we compare models with less than about 190 variables, BIC puts higher penalization than AIC when increasing variables. When the model is too large, for example with more than 190 predictors, BIC is inefficient when it comes to slopes and returns small values. This is because the sample size chosen was $n = 200$ and the number of predictors is almost 200. In such cases, BIC will not perform well and enter in "overfitting".

The red dots return average values of AIC which form an indistinct curve and flatter compared to BIC one. Therefore, BIC is more resistant to overfitting than the AIC. We also have a similar interpretation indicated in the BIC case for models with more than 170 predictors.

4.3 Consistency in Information Criteria

4.3.1 Problem and Solution

1. **Problem.**

   - Theoretically, as mentioned in Section 2.4.2, BIC is consistent. That means BIC is guaranteed to select the "true" model as the sample size n increases to infinity. In contrast, AIC can be considered to have weak consistency.
   - There are some assumptions to be concerned about:
     - ✓ The "true" model is included the list of candidate models.
     - ✓ The number of predictors p or the number of parameters is finite and existent.
     - ✓ When the sample size n increases, the dimension of the true model still remained.
     - ✓ The sample size n and p–number of variables must ensure that $p \leq n-2$.

2. **Particular solution by simulating data from polynomials.**
   Our data are chosen from a polynomial of degree 4

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 + \epsilon, \tag{4.5}$$

   where x denotes the predictor, y denotes the response variable, and ϵ, the random error.

   More specific,

   - List of coefficients $[\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4]$ are chosen as $[-2, -1, 1, 2, 3]$.
   - The sample size n is changed in range of list $[8, 16, 32, 64, 128, 256, 512, 1024]$
   - By each sample size n_i chosen, we generate n_i randomly predictors x that are distributed as a uniform distribution between -2 and 2 ($X \sim Uni(-2, 2)$).
   - The random error "ϵ" is considered to be a standard normal distribution $N(0, 1)$.
   - Six candidate models are used to fit the data. Starting from the polynomial of degree one and up to six.

- We repeated this process 2000 times and counted the number of times when AIC and BIC indicated the "true" model with degree four. We converted this number to an estimated probability of selecting the "true" model.
- These probabilities are used to observe the consistency property.

3. **Particular solution by simulating data from multiple linear regression.**
   We consider a set of data generated by linear regression:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \tag{4.6}$$

   where y is the response, X_1, X_2, X_3 are three predictors, and ϵ the random error. In this simulation, we generate three predictors and ϵ from normal distribution $N(0, 3)$. By a very similar way, we add 3 more predictors X_4, X_5, X_6 that are independent from y. Our set of candidate models has $2^6 = 64$ models.

   We carry out the simulation study as following:

   - Step 1: Let n equal each of the values in list $[8, 16, 32, 64, 1268, 256, 612, 1024]$
   - Step 2: For each value of n, generate n observations of each $X_1, \ldots, X_6$ and ϵ. Choose $[\beta_0, \beta_1, \beta_2, \beta_3] = [1, 2, 3, 4]$ and generate y according (4.6).
   - Step 3: For the generated dataset, fit 2^6 models and select ones with minimum AIC and BIC.
   - Step 4: If the ones found by AIC or BIC are the set of X_1, X_2, X_3 then AIC or BIC is considered as choosing the "true" model.
   - Step 2-4 are repeated 2000 times to find the approximate probability.

4.3.2 Results and Discussion

1. **Simulating data from polynomials.**
   As can be seen from the table in Figure 13, the sample size generally determines whether the candidate model with the true degree was chosen for both AIC and BIC. Secondly, the BIC is more consistent with the "best" candidate model compared to the AIC. It can be justified that AIC prefers 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 the consistency property of the BIC to select a true model. However, for sample sizes 32, 64, 128, 256, 512, 1024, 2048 the probability of AIC is hovering a round 75%. But it still holds that the BIC is more consistent compared to the AIC for these simulations. Besides, in the first row, there is little to choose between two metrics and both results are very low. A possible explanation may lie in the small sample size. That's why in practice, we often choose the sample size of at least 40 times larger than the number of parameters.

Figure 13 also graphically summarizes the estimated probability of AIC and BIC selecting the correct model. By observing, we can see clearly the consistency property of BIC.

Sample Size	AIC	BIC
8	0.1600	0.1630
16	0.5890	0.6930
32	0.7205	0.8900
64	0.7630	0.9520
128	0.7645	0.9615
256	0.7800	0.9760
512	0.7955	0.9860
1024	0.7820	0.9885
2048	0.7790	0.9965

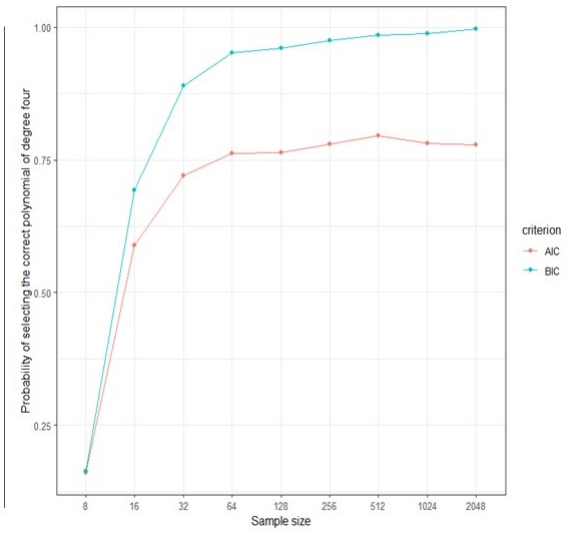

Figure 13: Probability of AIC and BIC selecting the correct polynomial with degree four among six candidate polynomials.

2. **Simulating data from multiple linear regression.**

   Figure 14 evidence results which are similar to previous simulation for polynomial. The BIC selects the true model consistently, and its probability of doing so rapidly approaches 1. The AIC is inconsistent in selecting the true model. Despite increasing double the sample size n at each stage, the probability of selecting the

actual model remains at about 0.60.

Sample Size	AIC	BIC
8	0.0170	0.0175
16	0.1495	0.1525
32	0.3230	0.3545
64	0.4890	0.6055
128	0.5675	0.8460
256	0.5695	0.9335
512	0.5970	0.9625
1024	0.5955	0.9700
2048	0.5950	0.9795

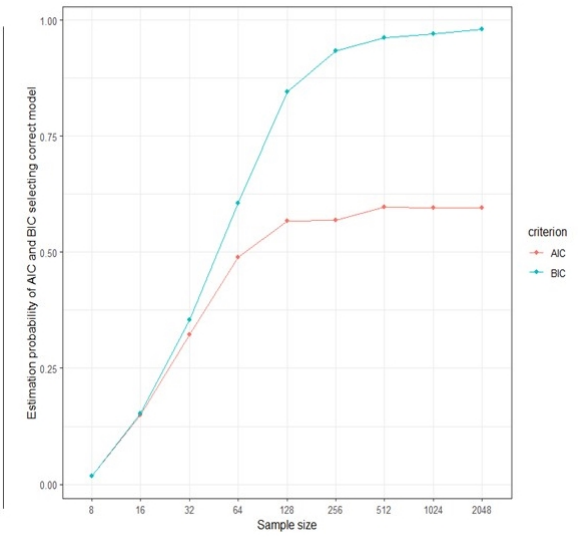

Figure 14: Visualization of the estimation probability of AIC and BIC selecting the correct linear regression with three predictors $X1, X2, X3$.

4.4 Tapering and Strong Effects

4.4.1 Problem and Solution

1. **Problem.**

   According to Burnham and Anderson [7], there are a wide range of simulation studies in the existing statistical literature that focus on either AIC or BIC alone, many of them aim to make comparisons between the two models and make suggestions for selection. It is worth noticing that these simulations are often too simplistic or biased against AIC, because a simple model is used without tapering effects. However, we usually witness tapering effect sizes ([7]). Thus, Burnham and Anderson argue that inference under model selection can first be enhanced by building a philosophy about models and data analysis, followed by identifying an appropriate model selection criterion. In such a philosophy, the critical question to be addressed is whether the information extracted by the model in the set is simple (a few big effects - later referred to as strong effects) or complex (many tapering effects-coefficients have close values). It is concluded that in such a context, AIC performs better than BIC. AIC is recommended if this

is real data analysis. Otherwise, BIC should be used in case of there are only a few big effects and all others are zero. Although this is not a widely held belief, many authors have recommended BIC as the criterion of choice such as Lawrence (2008)–[5]. Our simulations are used to examine this issue in some simplistic ways: uncorrelated covariates and correlated covariates.

2. **Solution**

The simulations are modified versions from [14]. We create 500 datasets generated from multivariate normal distribution

$$X \sim N(\mu_X, \Sigma_X).$$

A linear model is specified as

$$Y = \alpha + \beta X + \epsilon,$$

where ϵ has a normal distribution where its elements have mean zero and variance equals 1. We assume in case of tapering effects:

- Our model has 21 real effects, coefficients starting from $\beta_1 = 3$ and tapering off quickly on a log scale to very small values $\beta_{21} = 0.005$. In details the array of $\beta_1, \beta_2, \ldots, \beta_{21}$ is assigned as

$$\beta = \exp(\text{seq}(\log(3), \log(0.005), \text{length} = 21))$$

$$\text{The true model} : Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{21} X_{21} + \epsilon.$$

- Suppose that we can only collect information of $X_1, X_6, X_{11}, X_{16}, X_{21}$, the other variables are unknown or omitted, simply the data set we consider is a set of $Y, X_1, X_6, X_{11}, X_{16}, X_{21}$. In the context of real data analysis, the model is made up of factors that we do not know, do not have information about, or cannot collect data from those. That's why in this simulation, we create some variables that are used to form the true model but are omitted in the dataset. Therefore, the model which we consider for analyzing is

$$Y = \alpha + \beta_1 X_1 + \beta_6 X_6 + \beta_{11} X_{11} + \beta_{16} X_{16} + \beta_{21} X_{21} + \epsilon.$$

For strong effects, the coefficients of $X_1, X_2, \ldots, X_{10}$ are supposed to equal 3 as 10 strong effect sizes. Similar to the case of tapering effects, we can only collect information of $X_1, \ldots, X_5$ that will be supplied as data, others denoted omitted variables, which are not available in the considered model.

We select the "best" regression model from each of these datasets using two model selection procedures: AIC and BIC. We repeat this procedure for different sample sizes ($n = 5, 10, \ldots, 150$). Making comparisons between the average RMSE of these model selection procedures and the true data-generating processes helps to examine the extent to which each strategy includes the correct variables and excludes the irrelevant.

4.4.2 Results and Discussion

For the uncorrelated covariates, Σ_X is the identity matrix. The averages RMSE for strong effects and tapering effects, for each of AIC and BIC, for each sample size from the simulation are shown in Figure 15.

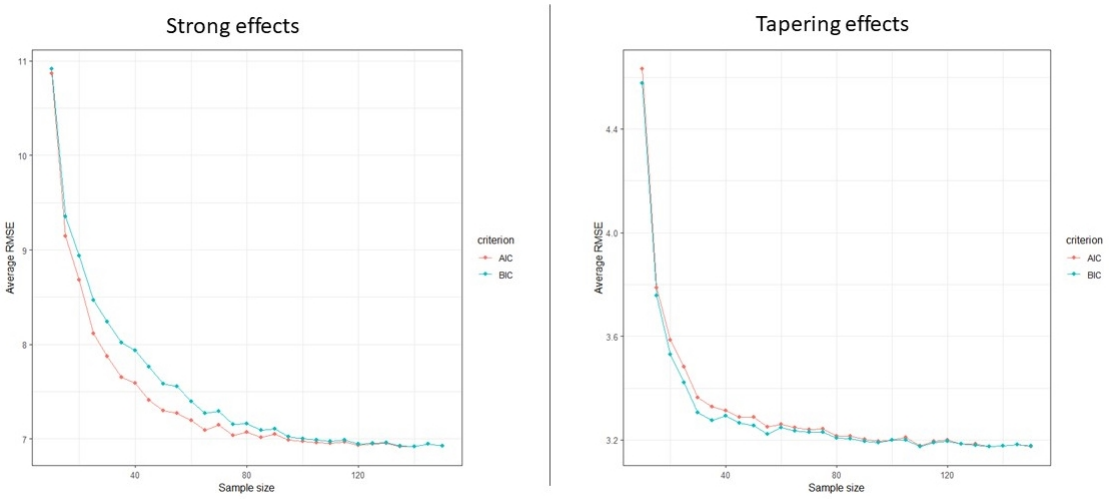

Figure 15: Averages RMSE when using AIC and BIC selecting regression models for strong efects and tapering effects in case the covariance matrix is identity. Sample size is in range of $n = 5, 10, \ldots, 150$..

We can observe that there is minimal difference between the two measures for tapering effects in terms of RMSE, only a small advantage for BIC on small sample

sizes. This is in conflict with the conclusion that Burnham and Anderson have stated on [7]-pages 25, 33, AIC performs better than BIC in the context of tapering effects. One explanation for this is that the AIC often chooses very weak effects. In case sample sizes are small, AIC might estimate poorly coefficients of weak effects. For example, these coefficients would be supposed to estimate to be positive but the model estimated them to be negative, in these cases, choosing coefficient to equal zero should be better. In case of strong effects, for small samples, there is a slight advantage of AIC and the difference is more visible than in the tapering effects ones. For larger samples, the two criteria perform equally well both in strong and tapering effects. Our finding turns out that AIC is not a preference for all cases of tapering effects.

Next, we consider the covariance matrix Σ_X not to be identified as previous, which means the data is not independently generated. All elements not in the diagonal are set equal to 0.5 and the elements on the diagonal are 1. Given theory, in the tapering-effects context, AIC performs better than BIC and this can be seen in Figure 16. There is a strong advantage for AIC compared to BIC in terms of RMSE. For strong effect, there is no difference between the two simulations. We can see that the correlation does not affect the performance of the two criteria AIC and BIC in terms of RMSE in the case of strong effects.

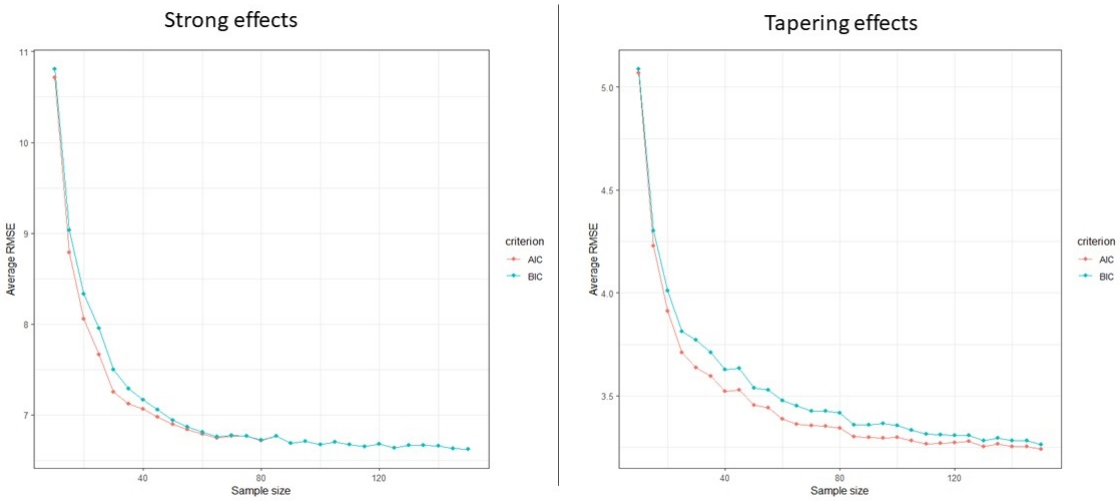

Figure 16: Averages RMSE when using AIC and BIC selecting regression models for strong efects and tapering effects in case the covariance matrix is **not** identity. Sample size is in range of $n = 5, 10, \ldots, 150$.

Conclusion

The thesis aims to take an in-depth look at the two well known information criteria, the AIC and BIC. Initially, we revisited some statistical concepts, namely the K-L information used to measure the loss information, regression analysis used to investigate the relationships between a response and (multiple) independent variables, likelihoods, loss and risk functions, and the Bayes theory.

AIC and BIC are grounded on theoretical studies in regression. AIC tends to reveal unknown models with the most predictive ability. Meanwhile, BIC often chooses the true model which is existing in the set of candidate models. Unlike AIC, BIC penalizes free parameters more strongly. In the context of limited simulations, AIC is likely to pick models with high predictive power. When considering the choice of selecting AIC or BIC we need to pay attention to the existence of the true model.The next issue we looked at was that there was a significant difference in the performance of AIC and BIC in the context of increasingly large data sets. With small datasets, the accuracy of BIC is more significant than that of AIC, that is why we should use adjusted AIC instead of AIC. The larger the data set, the more efficient AIC is. This work has also shown that BIC is consistent when choosing the true model while the true model is in the list of the candidate models. However, it is necessary to differentialte the consistency and efficiency. BIC does not always give an exact and clear choice. Even if the true model is in the set of considered models, BIC can still perform worse than AIC.

Another factor that many researchers are interested in is the problem of overfitting and underfitting of the model. Obviously, we should avoid letting our model fall into either of these scenarios. Our simulations confirmed that AIC tends to overfit while BIC tends to underfit models. The overfit of AIC is even more obvious when the model is complicated and the differentiating factors are not obvious. From this inference, we made a next simulation observing the performance of AIC and BIC with the model affected by strong effect and tapering effect. In case of strong effects, we created a very simple model and the difference in levels of variables is very large. The results were not unexpected as AIC performed better than BIC in the case of strong effects, even when the number of observations was small. With tapering effects, AIC is expected

to be better than BIC however we have obtained a simulation with results in favor of BIC in the case of the covariance matrix is identity.

Simulation studies are still in their infancy. There are many factors and aspects of model selection that have not been covered as well as the application of the information criterion for time series forecasting model selection and for other models.

Bibliography

[1] Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, 95(3), 631–636, doi: 10.1890/13-1452.1.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in B. N. Petrov and S. Caski, editors. *Proceedings of the Second International Symposium on Information Theory.* Akademiai Kaido.

[3] Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, 30(1), 9–14, doi: 10.1007/bf02480194.

[4] Bhansali, R. & Downham, D. Y. (1977). Some properties of order of an autoregressive model selected by a generalization of Akaikes FPE criterion. *Biometrika* 64(3), 547–551, doi: 10.2307/2345331.

[5] Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution,* 7(6), 678–692, doi: 10.1111/2041-210x.12541.

[6] Burnham, K. P., & Anderson, D. R. (2002). *Model selection and Multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer Science & Business Media.

[7] Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2), 261–304, doi: 10.1177/0049124104268644.

[8] Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging.* Cambridge University Press.

[9] Dziak J. J., Coffman D. L., Lanza S. T., & Li R., Lars S Jermiin (2020). Sensitivity and specificity of information criteria, *Briefings in Bioinformatics*, 21(2), 553—565, doi: 10.1093/bib/bbz016.

[10] Eid, M., & Langeheine, R. (1999). The measurement of consistency and occasion specificity with latent class models: A new model and its application to the

measurement of affect. *Psychological Methods*, 4(1), 100–116, doi: 10.1037/1082-989X.4.1.100.

[11] 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, doi: 10.1111/j.2517-6161.1979.tb01072.x.

[12] Hjort, N., & Pollard, D. (2011). (PDF) *Asymptotics for Minimisers of convex processes.*, available online (17.12.2021): https://arxiv.org/abs/1107.3806v1

[13] Hurlin, C. (2013). *Chapter 2: Maximum Likelihood Estimation* (PDF). HEC Lausanne, available online (16.08.2021): https://www.scribd.com/document/341757248/mle

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

[15] Katch, F. I., & McArdle, W. D. (1994). *Introduction to nutrition exercise, and health.* Fourth edition. The Endocrinologist.

[16] Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling.* Springerverlag.

[17] Kullback, S. (1959). *Information Theory and Statistics.* Wiley.

[18] McQuarrie, A. D., & Tsai, C. (1998). *Regression and time series model selection.* World Scientific.

[19] Newey, W. K., & McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. *Handbook of Econometrics*, 2111–2245, doi: 10.1016/s1573-4412(05)80005-4.

[20] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics.*, 758—765.

[21] Raffalovich, L. E., Deane, G. D., Armstrong, D., & Tsao, H. (2008). Model selection procedures in social research: Monte-Carlo simulation results, *Journal of Applied Statistics*, 35(10), 1093–1114, doi: 10.1080/03081070802203959.

[22] Schwarz, Gideon E. (1978), Estimating the dimension of a model, *Annals of Statistics*, 12(2), 461—464, doi: 10.1214/aos/1176346522.

[23] Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics*, 35(3), 415–423, doi: 10.1007/bf02480998.

[24] Sugiura, N. (1978) *Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections.* Communications in Statistics—Theory and Methods, 13–26.

[25] Tong, H. (1975). Autoregressive model fitting with noisy data by Akaike's information criterion (Corresp.). *IEEE Transactions on Information Theory*, 21(4), 476–480, doi: 10.1109/TIT.1975.1055402.

[26] Takeuchi, K. (1976). Distributions of information statistics and criteria for adequacy of models. *Mathematical Science*, 153, 230–245.

[27] Umberto, T. (2014). *Maximum Likelihood Statistical Inference* [PDF]. Corsi di Studio di Economia, available online (07.01.2022): https://ec.univaq.it/fileadmin/user_upload/Economia/docenti/Triacca_Umberto/4674Newtrinity.pdf

[28] Wikipedia, the free encyclopedia. Hirotogu Akaike, available online (16.08.2021) https://en.wikipedia.org/wiki/Hirotugu_Akaike

[29] Zellner, A., Keuzenkamp H. A., and McAleer M. (eds.) (2001). *Simplicity, Inference and Modelling.* Cambridge University Press, 83–119.