國立交通大學

統計學研究所

碩士論文

AIC、BIC 和 EBIC 之回顧

Review of AIC, BIC and EBIC

研 究 生：顏妤樺
指導教授：洪慧念　教授

中華民國　一〇三　年　六　月

AIC、BIC 和 EBIC 之回顧

Review of AIC, BIC and EBIC

研究生：顏妤樺　　Student: Yu-Hua Yen

指導教授：洪慧念　博士　　Advisor: Dr. Hui-Nien Hung

國立交通大學理學院

統計學研究所

碩士論文

A Thesis

Submitted to

Institute of Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2014

Hsinchu, Taiwan, Republic of China

中華民國　一〇三　年　六　月

AIC、BIC 和 EBIC 之回顧

研究生：顏妤樺　　指導教授：洪慧念　博士

國立交通大學統計學研究所碩士班

摘　要

自資訊爆炸以來，利用統計方法分析資料漸漸成為一種常態。而我們所面對的問題也從過去的大樣本資料分析逐漸轉變成高維度資料分析。如何找出這些資料的最適模型是我們最重要的課題。在這篇文章中，我們將 Chen & Chen (2008) 提出之針對高維度模型選取方法 EBIC 與常見的模型選取方法 AIC、BIC 做比較，並利用模擬的方式說明這些方法的差異與優劣。

關鍵詞：高維度模型、模型選取、AIC、BIC、EBIC

Review of AIC, BIC and EBIC

Student: Yu-Hua Yen Advisor: Dr. Hui-Nien Hung

Institute of Statistics

National Chiao Tung University

Hsinchu 30010, Taiwan

Abstract

Since the information explosion, analyzing data by using statistical methods progressively becomes norm. Nowadays, the problem we are faced with large sample size analysis gradually transformed into high dimensional model analysis. How to find the optimal model for the data is our most important issue. In our study, we compare EBIC, which proposed by Chen & Chen (2008) for high dimensional model, with common model selection methods, AIC and BIC, and use simulations illustrating the difference and the pros and cons of these methods.

Key words: High Dimensional Model, Model Selection, AIC, BIC, EBIC

誌　謝

首先，誠摯的感謝我的指導教授　洪慧念博士，感謝老師在指導我的論文上所花費的時間與精力，並且不時的指點與引導我正確的研究方向，在這個過程中，除了更瞭解模型選擇的方法之外，我學到了一件最重要的事：透過舉例的方式將抽象的概念具體表達出來。另外，感謝口試委員　徐南蓉教授、黃信誠教授和王維菁教授的建議與指教，使得本論文更為完整。

兩年裡的日子，研究室裡共同的生活點滴，學術上的討論、言不及義的閒扯、趕作業的革命情感等，感謝所有同學的互相勉勵，因為有你們讓兩年的研究所生活變得絢麗許多，恭喜我們都順利走過這兩年並且即將畢業。

感謝好友賴怡方、白永馨、徐光縈無時無刻給予我心靈上最強大的支持，也感謝室友邱晴瑜、劉晏羽與我交流他們在各自的領域上對各種人事物的想法與觀點，讓我的思維更加的多元化，另外特別感謝過去的學長、現在的學弟張景弦三不五時提供我別人的八卦豐富我的生活。最後，感謝我的父母提供我安穩的生活與最棒的避風港，給予我最穩固的支持。

謹以此論文獻給我的家人，以及所有關心我的師長與朋友們。

Contents

List of Tables

1 Introduction

Since the information explosion, information science is flourishing and the data volume owned by humans is increasing exponentially. For example, according to Technorati, an internet search engine for searching blogs, the number of blogs doubles about every 6 months with a total of 35.3 million blogs as of April 2006.

Today, we are faced with the era of "big data". So we are most concerned about an issue of how to analyze data using statistical methods. One of the problem is an appropriate model for a given data set. For example, in financial world, enterprises use a variety of its value creation information for building a financial model, so as to complete such as analysis, prediction and assessment of the financial performance of the enterprise.

There are two ordinary model selection methods, Akaike Information Criterion, AIC (Akaike, 1974) and Bayesian Information Criterion, BIC (Schwarz, 1978). In many areas, we can see examples of using AIC or BIC for model selection, such as in finance, use for stock-recruitment model selection (Wang & Liu, 2006) and in bioinformatics, use for mixed graphical model selection (Edwards, Abreu & Labouriau, 2010). Unfortunately, the problems we are faced with changing from large sample data analysis gradually to high dimensional model analysis today. In order to solve it, Chen & Chen (2008) proposed a new model selection method, Extended Bayesian Information Criterion (EBIC), which is particularly useful in genome-wide association studies.

In the following, we introduce three model selection methods mentioned above, AIC, BIC and EBIC in Section 2. For AIC and BIC, we refer the book "Information Criterion and Statistical Modeling" (Konishi & Kitagawa, 2008), which also includes GIC, TIC, PIC, DIC, etc., but in our study, we only focus on AIC, BIC and the new method, EBIC. In Section 3, we compare these three methods by simulation under general linear model case and AR(1) model case. Furthermore, we also consider the high dimensional model to illustrate the difference and the pros and cons of these three methods. Finally, we will give a conclusion about which is the best method in these three methods in Section 4.

2 Model Selection Methods

In this section, we describe two model selection methods, Akaike Information Criterion (AIC, 1974) and Bayesian Information Criterion (BIC, 1978), and introduce a new method, Extended Bayesian Information Criterion (EBIC, 2008), which is particularly useful in high dimensional model analysis.

2.1 Akaike Information Criterion

In the middle of the 20th century, a new financial instrument — "stock" rise, and the stock market is booming. The old statistical method, hypothesis testing, has been insufficient to analyze such time-series data sets. In 1974, Hirotugu Akaike first proposed the Akaike Information Criterion (AIC), which is designed for the purpose of statistical identification. In statistics, a model must be identifiable so as to infer its possible properties accurately. That is, we can use AIC to select a better model.

When we build a model by data, we assume that the data $\mathbf{x} = \{x_1, \cdots, x_n\}$ are generated from the true distribution $f(x)$. In order to capture the structure of the given phenomena, we assume a k-dimensional parametric model $\{g(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^k\}$ and we estimate it by the maximum likelihood method. In other words, we construct a statistical model $g(x|\hat{\boldsymbol{\theta}})$ by replacing the unknown parameter $\boldsymbol{\theta}$, which contained in the probability distribution, with the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$.

Kullback-Leibler information $I(f, \hat{g})$ is the information lost when statistical model $\hat{g} = g(x|\hat{\boldsymbol{\theta}})$ is used to approximating true distribution $f = f(x)$; it is defined as the integral

$$I(f, \hat{g}) = \int f(x) \log \left(\frac{f(x)}{g(x|\hat{\boldsymbol{\theta}})} \right) dx. \tag{2.1.1}$$

Obviously, the best model loses the least information relative to other models in the set; this is equivalent to minimizing $I(f, \hat{g})$ over $\hat{g}$. Furthermore, K-L information also can be conceptualized as a "distance" between true distribution and a statistical model.

Equation (2.1.1) can be expressed as

$$\begin{aligned} I(f,\hat{g}) &= \int f(x)\log f(x)dx - \int f(x)\log g(x|\hat{\boldsymbol{\theta}})dx \\ &= \int \log f(x)dF(x) - \int \log g(x|\hat{\boldsymbol{\theta}})dF(x) \end{aligned}$$

or

$$I(f,\hat{g}) = E_F[\log f(X)] - E_F[\log g(X|\hat{\boldsymbol{\theta}})],$$

where the expectations are taken with respect to true distribution $F(x)$ and the quantity $E_F[\log f(X)]$ is a constant (say C) across models. Hence,

$$I(f,\hat{g}) = \mathrm{C} - E_F[\log g(X|\hat{\boldsymbol{\theta}})],$$

where

$$\mathrm{C} = \int \log f(x)dF(x)$$

does not depend on the data or the statistical model. Thus, only relative expected K-L information, $E_F[\log g(X|\hat{\boldsymbol{\theta}})]$, needs to be estimated for each model in the set.

One such estimator is

$$\begin{aligned} E_{\hat{F}}[\log g(X|\hat{\boldsymbol{\theta}})] &= \int \log g(x|\hat{\boldsymbol{\theta}})d\hat{F}(x) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\log g(x_i|\hat{\boldsymbol{\theta}}) \\ &= \frac{1}{n}\log g(\mathbf{x}|\hat{\boldsymbol{\theta}}), \end{aligned}$$

in which the unknown probability distribution F contained in the expected log-likelihood is replacing with an empirical distribution function $\hat{F}$. So the log-likelihood $\log g(\mathbf{x}|\hat{\boldsymbol{\theta}})$ is an estimator of the expected log-likelihood $nE_F[\log g(X|\hat{\boldsymbol{\theta}})]$.

The bias of the log-likelihood as an estimator of the expected log-likelihood $E_F[\log g(X|\hat{\boldsymbol{\theta}})]$ is defined by

$$bias(F) = E_{F(\mathbf{x})}[\log g(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{x})) - nE_{F(x)}[\log g(X|\hat{\boldsymbol{\theta}}(\mathbf{x}))]],$$

where the expectation $E_{F(\mathbf{x})}$ and $E_{F(x)}$ are taken with respect to the joint distribution, $\prod_{i=1}^{n} F(x_i) = F(\mathbf{x})$, of the sample $\mathbf{x}$ and true distribution $F(x)$ respectively, $\mathbf{x}$ and x are independent.

According to Konishi & Kitagawa (2008), suppose that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$ when $n \to \infty$, then the bias can be decomposed

as follow:

$$\begin{aligned}
&E_{F(\mathbf{x})}[\log g(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{x})) - nE_{F(x)}[\log g(X|\hat{\boldsymbol{\theta}}(\mathbf{x}))]] \\
=\ &E_{F(\mathbf{x})}[\log g(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{x})) - \log g(\mathbf{X}|\boldsymbol{\theta}_0)] \\
+\ &E_{F(\mathbf{x})}[\log g(\mathbf{X}|\boldsymbol{\theta}_0) - nE_{F(x)}[\log g(X|\boldsymbol{\theta}_0)]] \\
+\ &E_{F(\mathbf{x})}[nE_{F(x)}[\log g(X|\boldsymbol{\theta}_0)] - nE_{F(x)}[\log g(X|\hat{\boldsymbol{\theta}}(\mathbf{x}))]].
\end{aligned}$$

By writing $\ell(\boldsymbol{\theta}) = \log g(\mathbf{x}|\boldsymbol{\theta})$ and applying a Taylor series expansion around the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, we obtain

$$\ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left.\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \cdots . \tag{2.1.2}$$

Here, the quantity $\hat{\boldsymbol{\theta}}$ satisfies the equation $\left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} = 0$ and the quantity

$$-\frac{1}{n}\left.\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}} = -\frac{1}{n}\left.\frac{\partial^2 \log g(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}}$$

converges in probability to

$$J(\boldsymbol{\theta}_0) = -E_{F(x)}\left[\left.\frac{\partial^2 \log g(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_0}\right] = -\int f(x) \left.\frac{\partial^2 \log g(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_0} dx \tag{2.1.3}$$

when n tends to ∞. Then we can obtain the approximation

$$\ell(\boldsymbol{\theta}_0) - \ell(\hat{\boldsymbol{\theta}}) \approx -\frac{n}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})$$

for (2.1.2). Based on this result, we obtain approximately

$$\begin{aligned}
E_{F(\mathbf{x})}[\log g(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{x})) - \log g(\mathbf{X}|\boldsymbol{\theta}_0)] &= \frac{n}{2}E_{F(\mathbf{x})}[(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})] \\
&= \frac{n}{2}E_{F(\mathbf{x})}[\mathrm{tr}\{J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T\}] \\
&= \frac{n}{2}\mathrm{tr}\{J(\boldsymbol{\theta}_0)E_{F(\mathbf{x})}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T]\}.
\end{aligned}$$

By substituting the asymptotic variance covariance matrix

$$E_{F(\mathbf{x})}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T] = \frac{1}{n}J(\boldsymbol{\theta}_0)^{-1}I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1} \tag{2.1.4}$$

of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, where

$$\begin{aligned}
I(\boldsymbol{\theta}_0) &= E_{F(\mathbf{x})}\left[\left.\frac{\partial \log g(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \log g(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_0}\right] \\
&= \int f(x) \left.\frac{\partial \log g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \log g(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_0} dx,
\end{aligned} \tag{2.1.5}$$

we have

$$E_{F(\mathbf{x})}[\log g(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{x})) - \log g(\mathbf{X}|\boldsymbol{\theta}_0)] = \frac{1}{2}\mathrm{tr}\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\}. \tag{2.1.6}$$

Now we evaluate the easiest part

$$E_{F(\mathbf{x})}[\log g(\mathbf{X}|\boldsymbol{\theta}_0) - nE_{F(x)}[\log g(X|\boldsymbol{\theta}_0)]],$$

which does not contain an estimator. It can easily be seen that

$$\begin{aligned} & E_{F(\mathbf{x})}[\log g(\mathbf{X}|\boldsymbol{\theta}_0) - nE_{F(x)}[\log g(X|\boldsymbol{\theta}_0)]] \\ &= E_{F(\mathbf{x})}\left[\sum_{i=1}^{n} \log g(\mathbf{X}_i|\boldsymbol{\theta}_0)\right] - nE_{F(x)}[\log g(X|\boldsymbol{\theta}_0)] \\ &= 0. \end{aligned} \tag{2.1.7}$$

The final part

$$E_{F(\mathbf{x})}[nE_{F(x)}[\log g(X|\boldsymbol{\theta}_0)] - nE_{F(x)}[\log g(X|\hat{\boldsymbol{\theta}}(\mathbf{x}))]]$$

can be calculated approximately as follows:

$$\begin{aligned} & E_{F(\mathbf{x})}[nE_{F(x)}[\log g(X|\boldsymbol{\theta}_0)] - nE_{F(x)}[\log g(X|\hat{\boldsymbol{\theta}}(\mathbf{x}))]] \\ &\approx nE_{F(\mathbf{x})}[\frac{1}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})] \\ &= \frac{n}{2}E_{F(\mathbf{x})}[\mathrm{tr}\{J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T\}] \\ &= \frac{n}{2}\mathrm{tr}\{J(\boldsymbol{\theta}_0)E_{F(\mathbf{x})}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T]\}. \end{aligned}$$

By the asymptotic variance covariance matrix (2.1.4) of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, we have

$$E_{F(\mathbf{x})}[nE_{F(x)}[\log g(X|\boldsymbol{\theta}_0)] - nE_{F(x)}[\log g(X|\hat{\boldsymbol{\theta}}(\mathbf{x}))]] = \frac{1}{2}\mathrm{tr}\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\}. \tag{2.1.8}$$

Therefore, combining (2.1.6), (2.1.7) and (2.1.8), the bias resulting from the estimation of the expected log-likelihood of the model is asymptotically obtained as

$$bias(F) = \frac{1}{2}\mathrm{tr}\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\} + 0 + \frac{1}{2}\mathrm{tr}\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\} = \mathrm{tr}\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\},$$

where $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$ are respectively given in (2.1.5) and (2.1.3).

Now assume that the true distribution $f(x)$ can be expressed as $f(x) = g(x|\boldsymbol{\theta}_0)$ for properly specified $\boldsymbol{\theta}_0 \in \Theta \subset R^k$. Under this assumption, the equality $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ holds for the $k \times k$ matrix $I(\boldsymbol{\theta}_0)$ given in (2.1.5) and $J(\boldsymbol{\theta}_0)$ given in (2.1.3). Therefore, the bias of the log-likelihood is asymptotically given by

$$bias(F) = \mathrm{tr}\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\} = \mathrm{tr}\{I_k\} = k,$$

where I_k is the identity matrix of dimension k. Hence, the AIC

$$AIC = -2\sum_{i=1}^{n} \log g(x_i|\hat{\boldsymbol{\theta}}) + 2k$$

can be obtained by correcting the asymptotic bias k of the log-likelihood.

Then we give an example to calculate the value of its AIC. Suppose there is a linear model

$$\mathbf{Y} = 1 \cdot \mathbf{X}_1 + 1 \cdot \mathbf{X}_2 + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)^T$, $\mathbf{X}_1 = (X_{11}, X_{21}, \cdots, X_{n1})^T$, $\mathbf{X}_2 = (X_{12}, X_{22}, \cdots, X_{n2})^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$, $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$, $i = 1, 2, \cdots, n$. We use `rnorm()` in R to generate the data of covariates $\mathbf{X}_1$ and $\mathbf{X}_2$, each covariate contains 50 records. We also generate the data of $\boldsymbol{\epsilon}$ by `rnorm()`. And we use `lm()` and `AIC()` to compute the AIC value of the simulated data of the linear model. Consider the following two models:

$$M_1 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \boldsymbol{\epsilon}$$
$$M_2 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \boldsymbol{\epsilon},$$

and let M_2 be the true model, the coefficients $\beta_1 = \beta_2 = 1$. By the function `AIC()` in R, we get the AIC values for M_1 and M_2 equal to 189.0057 and 161.9824 respectively. Therefore, in this simulation, we will prefer the true model M_2 rather than the model M_1.

2.2 Bayesian Information Criterion

The maximum likelihood principle in some cases, such as the choice of degree for a polynomial regression and the choice of order for a multi-step Markov chain, invariably leads to choosing the highest possible dimension, but not the "right" dimension. Although there is a general model selection method, AIC, which is an extension of the maximum likelihood principle, Schwarz (1978) proposed an alternative method, Baysian Information Criterion (BIC), especially for this problem. It is derived as follows.

According to Konishi & Kitagawa (2008), let $M_1, M_2, \cdots, M_r$ be r candidate models, and assume that each model M_i is characterized by a parametric distribution $g_i(x|\boldsymbol{\theta}_i)$ ($\boldsymbol{\theta}_i \in \Theta_i \subset R^{k_i}$) and the prior distribution $\pi_i(\boldsymbol{\theta}_i)$ of the k_i-dimensional parameter vector $\boldsymbol{\theta}_i$. When n observations $\mathbf{x} = \{x_1, \cdots, x_n\}$ are given, then, for the ith model M_i, the marginal distribution or probability of $\mathbf{x}$ is given by

$$m(\mathbf{x}|M_i) = \int g_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i, \tag{2.2.1}$$

This quantity can be considered as the likelihood of the ith model and is referred to as the marginal likelihood of the data.

According to Bayes' theorem, if we suppose that the prior probability of the ith model is $p(M_i)$, the posterior probability of the ith model is given by

$$p(M_i|\mathbf{x}) = \frac{m(\mathbf{x}|M_i)p(M_i)}{\sum_{j=1}^{r} m(\mathbf{x}|M_j)p(M_j)}, \qquad i = 1, 2, \cdots, r. \tag{2.2.2}$$

This posterior probability indicates the probability of the data being generated from the ith model when data $\mathbf{x}$ are observed. Therefore, if one model is to be selected from r models, it would be natural to adopt the model that has the largest posterior probability. This principle means that the model that maximizes the numerator $m(\mathbf{x}|M_i)p(M_i)$ must be selected, since all models share the same denominator in (2.2.2).

If we further assume that the prior probabilities $p(M_i)$ are equal in all models, it follows that the model that maximizes the marginal likelihood $m(\mathbf{x}|M_i)$ of the data must be selected. Therefore, if an approximation to the marginal likelihood expressed in terms of an integral in (2.2.1) can readily be obtained, the need to compute the integral on a problem-by-problem basis will vanish, thus making the BIC suitable for use as a general model selection criterion.

Equation (2.2.1) may be written as

$$m(\mathbf{x}|M) = \int \exp\{\log g(\mathbf{x}|\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2.2.3}$$

The Laplace approximation (Laplace, 1774) takes advantage of the fact that when the number n of observations is sufficiently large, the integrand is concentrated in a neighborhood of the mode of $\log g(\mathbf{x}|\boldsymbol{\theta})$ or, in this case, in a neighborhood of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, and that the value of the integral depends on the behavior of the function in this neighborhood.

Since $\left.\frac{\partial \log g(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$ holds for the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, the Taylor expansion of the log-likelihood function $\log g(\mathbf{x}|\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}}$ yields

$$\log g(\mathbf{x}|\boldsymbol{\theta}) = \log g(\mathbf{x}|\hat{\boldsymbol{\theta}}) - \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \cdots, \tag{2.2.4}$$

where

$$J(\hat{\boldsymbol{\theta}}) = -\frac{1}{n}\left.\frac{\partial^2 \log g(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Similarly, we can expand the prior distribution $\pi(\boldsymbol{\theta})$ in a Taylor series around the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ as

$$\pi(\boldsymbol{\theta}) = \pi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left.\frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \cdots . \tag{2.2.5}$$

Substituting (2.2.4) and (2.2.5) into (2.2.3) and simplifying the results lead to the approximation of the marginal likelihood as follows:

$$\begin{aligned} m(\mathbf{x}|M) = \int \exp\left\{\log g(\mathbf{x}|\hat{\boldsymbol{\theta}}) - \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \cdots\right\} \\ \cdot \left\{\pi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left.\frac{\partial \pi(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \cdots\right\} d\boldsymbol{\theta} \\ \approx \exp\{\log g(\mathbf{x}|\hat{\boldsymbol{\theta}})\}\pi(\hat{\boldsymbol{\theta}})(2\pi)^{\frac{k}{2}} n^{-\frac{k}{2}} \left|J(\hat{\boldsymbol{\theta}})\right|^{-\frac{1}{2}}. \end{aligned}$$

Taking the logarithm of this expression and multiply it by -2, we obtain

$$-2\log m(\mathbf{x}|M) \approx -2\log g(\mathbf{x}|\hat{\boldsymbol{\theta}}) + k\log n + \log\left|J(\hat{\boldsymbol{\theta}})\right| - k\log(2\pi) - 2\log\pi(\hat{\boldsymbol{\theta}}). \tag{2.2.6}$$

Then the following model evaluation criterion BIC can be obtained by ignoring terms with order less than $O(1)$ with respect to the sample size n.

Let $g(\mathbf{x}|\hat{\boldsymbol{\theta}})$ be a statistical model estimated by the maximum likelihood method. Then the Bayesian information criterion BIC is given by

$$BIC = -2\log g(\mathbf{x}|\hat{\boldsymbol{\theta}}) + k\log n.$$

For example, under the same assumptions of the example in Subsection 2.1, but we use `BIC()` instead of `AIC()` to compute the BIC value of the simulated data in here. Then, we get the BIC values for M_1 and M_2 (true model) equal to 190.1232 and 159.6058 respectively. Therefore, in this simulation, we will prefer the true model M_2 rather than the model M_1.

From the above argument, it can be seen that, BIC is an evaluation criterion for models estimated by using the maximum likelihood method and that the criterion is obtained under the condition that the sample size n is made sufficiently large. We also see that it was obtained by approximating the marginal likelihood associated with the posterior probability of the model by Laplace's method for integrals and that it is not an information criterion, leading to an unbiased estimation of the K-L information.

2.3 Extended Bayesian Information Criterion

In a typical genome-wide association study with single-nucleotide polymorphisms, the number of covariates is of the order of tens or hundreds or thousands while the sample size is only in the hundreds. To solve the problem with a moderate sample size but with a huge number of covariates, a new model selection method, Extended Bayesian Information Criterion (EBIC), proposed by Chen & Chen (2008).

Suppose that the number of covariates under consideration is $P = 1000$. The class of models containing a single covariate, S_1, has size 1000, while the class of models containing two covariates, S_2, has size $1000 \times 999/2$. The constant prior behind BIC amounts to assigning probabilities to the S_k proportional to their sizes. Thus, the probability assigned to S_2 is $999/2$ times that assigned to S_1. The size of S_k increases as k increases to $k = P/2 = 500$, so that the probability assigned to S_k by the prior increases almost exponentially. Models with a larger number of covariates, 50 or 100 say, receive much higher probabilities than models with fewer covariates. This is obviously unreasonable, being strongly against the principle of parsimony.

This re-examination of BIC prompts us to consider other reasonable priors over the model space in the Bayesian approach. Assume that the model space $\mathbf{S}$ is partitioned into $\cup_{k=1}^{P} S_k$, such that models within each S_k have equal dimension. Let $\tau(S_k)$ be the size of S_k. For example, if S_k is the collection of all models with k covariates, then $\tau(S_k) = \begin{pmatrix} P \\ k \end{pmatrix}$. We assign the prior distribution over $\mathbf{S}$ as follows. For each model M in the same subspace S_k, assign an equal probability, i.e. $pr(M|S_k) = 1/\tau(S_k)$ for any $M \in S_k$. This implies that all the models in S_k are equally plausible. Then, instead of assigning probabilities $pr(S_k)$ proportional to $\tau(S_k)$, as in the ordinary BIC, we assign $pr(S_k)$ proportional to $\tau^{\xi}(S_k)$ for some ξ between 0 and 1. This results in the prior probability $p(M)$ for $M \in S_k$ being proportional to $\tau^{-\gamma}(S_k)$, where $\gamma = 1 - \xi$. This type of prior distribution on the model space gives rise to an extended BIC family

$$BIC_\gamma(M) = -2 \log L\{\hat{\boldsymbol{\theta}}(M)\} + k \log n + 2\gamma \log \tau(S_k), \ 0 \le \gamma \le 1,$$

where $\hat{\boldsymbol{\theta}}(M)$ is the maximum likelihood estimator of $\boldsymbol{\theta}(M)$ given model M and k is the number of components in M. The first two terms in $BIC_\gamma(M)$ are the Laplace approximation to $-2 \log m(\mathbf{x}|M)$ and the last term is $-2 \log p(M)$ up to a common constant. The criterion BIC_γ is referred to as an extended Bayes information criterion.

Let's give an example to calculate its EBIC value. Suppose there is a model which contained 50 covariates, but we only have 30 records of this model. Consider the following two models:

$$M_1 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \boldsymbol{\epsilon}$$

$$M_2 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \boldsymbol{\epsilon},$$

and let M_2 be the true model, the coefficients $\beta_1 = \beta_2 = 1$. Using the formula of EBIC to calculate the EBIC values, and we get its for M_1 and M_2 equal to 123.8225 and 109.6918 respectively. Since the EBIC value of M_2 is less than M_1's, we may think that M_2 is the true model rather than M_1.

In the targeted application, P can be very large but the cardinality of the candidate models is small. If some of the covariates are heavily collinear, the effective number of different models might be smaller than that indicated by $\tau(S_k)$, and one might fear that our method is affected. Consider an extreme case in which half of the covariates are duplicates. Thus, in considering $\tau(S_k)$, P should be replaced by $P/2$. However, it is easy to see that, when P is replaced by $P/2$, the change in $\gamma \log \tau(S_k)$ is of a smaller order than the order $\log n + \log P$ of the leading terms. Thus, some adjustment might be helpful but the effect will not be important when n or P is large.

3 Comparison of AIC, BIC and EBIC

3.1 Large Sample Size ($n > p$)

3.1.1 Linear Model

Suppose there are three covariates $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ in a data, but the true model is

$$\mathbf{Y} = 1 \cdot \mathbf{X}_1 + 1 \cdot \mathbf{X}_2 + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)^T$, $\mathbf{X}_i = (X_{1i}, X_{2i}, \cdots, X_{ni})^T$, $i = 1, 2, 3$ and each component of $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$ is normally distributed independent with mean 0 and variance 1. We generate the data of covariates $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ separately from standard normal distribution, that is, $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ are independent standard normally distributed.

Consider all possible models:

M_1 : $\mathbf{Y} = \beta_1 \mathbf{X}_1 + \boldsymbol{\epsilon}$

M_2 : $\mathbf{Y} = \beta_2 \mathbf{X}_2 + \boldsymbol{\epsilon}$

M_3 : $\mathbf{Y} = \beta_3 \mathbf{X}_3 + \boldsymbol{\epsilon}$

M_4 : $\mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \boldsymbol{\epsilon}$ (true)

M_5 : $\mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_3 \mathbf{X}_3 + \boldsymbol{\epsilon}$

M_6 : $\mathbf{Y} = \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \boldsymbol{\epsilon}$

M_7 : $\mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \boldsymbol{\epsilon}$,

and we compute the value of AIC, BIC and EBIC, finding which information criterion has the best performance. Since the difference between calculation results of function in R and the original formula are insignificant (see Table 3.1), we will use function in R to compute the value of AIC, BIC and EBIC in the following. (Suppose that the prior distributions of the coefficients β_1, β_2 and β_3 are independent exponential distribution with $\lambda = 1$ in BIC.)

First, we compare three information criterions under different sample size n. In each case, we simulate 1000 times and fix the parameter γ of EBIC to 1.0. In Table 3.2, when n is large enough ($n \geq 30$), the performances of three information criterions are good, and in this time, the result of BIC better than AIC is more significant than when n is not large enough. In addition, three information criterions indeed exist the large models tendency mentioned in Schwarz (1978) and Chen & Chen (2008). And no matter n is how much, the results of EBIC are worse than BIC, even worse than AIC in the case of

Table 3.1: AIC, BIC and EBIC values of one simulated data set ($n = 100$) of the true model M_4 calculated under the original formula, approximative formula and function in R. The original values of AIC, BIC and EBIC are calculated by the joint distribution of the sample $\mathbf{x}$, (2.2.6) and (2.2.6) plus the the correction term, respectively. The approximations are calculated by the formula for the three criterions.

Criterion	Original Value	Approximation	function in R
AIC	283.7878	281.1892	283.6349
BIC	286.8539	286.3996	294.0556
EBIC	289.0512	288.5968	296.2528

Table 3.2: Probability of model selection respectively using AIC, BIC and EBIC within M_1 to M_7 under different sample size n. Each case simulated 1000 times and the parameter γ of EBIC fixed to 1.0.

n	Criterion	M_1	M_2	M_3	M_4	M_5	M_6	M_7
10	AIC	0.071	0.06	0.014	0.562	0.024	0.03	0.239
	BIC	0.084	0.07	0.017	0.565	0.026	0.031	0.207
	EBIC	0.065	0.054	0.014	0.164	0.005	0.005	0.693
20	AIC	0.006	0.009	0	0.775	0.001	0.005	0.204
	BIC	0.02	0.014	0	0.843	0.003	0.007	0.113
	EBIC	0.019	0.012	0	0.542	0	0.002	0.425
30	AIC	0.001	0	0	0.808	0	0	0.191
	BIC	0.001	0.002	0	0.909	0	0.001	0.087
	EBIC	0.001	0.001	0	0.674	0	0	0.324
50	AIC	0	0	0	0.824	0	0	0.176
	BIC	0	0	0	0.932	0	0	0.068
	EBIC	0	0	0	0.796	0	0	0.204
100	AIC	0	0	0	0.811	0	0	0.189
	BIC	0	0	0	0.951	0	0	0.049
	EBIC	0	0	0	0.85	0	0	0.15

$n = 10$. Therefore, we compare information criterions under different γ, the parameter involved in the correction term of EBIC, in the following.

Because γ only involve in the correction term of EBIC, it has nothing to do with AIC and BIC. So we only focus on comparing the impact of different γ on EBIC. In Table 3.3, the performance of EBIC is good but not better than BIC, and the larger γ, the worse performan-

ce of EBIC. Since EBIC is applied suitably in the situation p greater than n, this result is expectable and acceptable.

Table 3.3: Probability of model selection respectively using AIC, BIC and EBIC within M_1 to M_7 under different parameter γ. Each case simulated 1000 times and the sample size n is 100.

γ	Criterion	M_1	M_2	M_3	M_4	M_5	M_6	M_7
0.1	AIC	0.000	0.000	0.000	0.841	0	0	0.159
	BIC	0	0	0	0.973	0	0	0.027
	EBIC	0	0	0	0.968	0	0	0.032
0.5	AIC	0	0	0	0.84	0	0	0.16
	BIC	0	0	0	0.96	0	0	0.04
	EBIC	0	0	0	0.929	0	0	0.071
1.0	AIC	0	0	0	0.82	0	0	0.18
	BIC	0	0	0	0.963	0	0	0.037
	EBIC	0	0	0	0.852	0	0	0.148

3.1.2 Autoregressive Model

Suppose that the true model is an AR(2) model with the paramters $\boldsymbol{\phi} = (\phi_1, \phi_2) = (0.6, 0.3)$, it can be written as

$$X_t = 0.6 \cdot X_{t-1} + 0.3 \cdot X_{t-2} + \epsilon_t, \ t = 2, 3, 4, \cdots,$$

where ϵ_t is followed standard normal distribution. Given $X_0 = 0$ and $X_1 = 0$, and then we generate the data of the model by above formula.

Table 3.4: Probability of model selection respectively using AIC, BIC and EBIC under different model sets. Each case simulated 1000 times and the sample size n is 100, the parameter γ of EBIC fixed to 1.0.

Max Order	Criterion	M_1	M_2	M_3	M_4	M_5	M_6	M_7
2	AIC	0	0.01	0.99				
	BIC	0	0.158	0.842				
	EBIC	0	0.067	0.933				
3	AIC	0	0.004	0.092	0.904			
	BIC	0	0.128	0.618	0.254			
	EBIC	0	0.086	0.183	0.731			
4	AIC	0	0.003	0.02	0.079	0.898		
	BIC	0	0.127	0.539	0.185	0.149		
	EBIC	0	0.101	0.174	0.005	0.72		
5	AIC	0	0	0.005	0.014	0.094	0.887	
	BIC	0	0.11	0.49	0.169	0.112	0.119	
	EBIC	0	0.134	0.186	0.021	0	0.659	
6	AIC	0	0	0.002	0.008	0.019	0.092	0.879
	BIC	0	0.107	0.502	0.127	0.109	0.068	0.087
	EBIC	0	0.138	0.206	0.015	0.001	0	0.64

Consider the following models:

$$M_1 : \ X_t = \epsilon_t$$

$$M_2 : \ X_t = \phi_1 X_{t-1} + \epsilon_t$$

$$M_3 : \ X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t \text{ (true)}$$

$$M_4 : \ X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \epsilon_t$$

$$M_5 : \ X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \phi_4 X_{t-4} + \epsilon_t$$

$$M_6 : \ X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \phi_4 X_{t-4} + \phi_5 X_{t-5} + \epsilon_t$$

$$M_7 : \ X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \phi_4 X_{t-4} + \phi_5 X_{t-5} + \phi_6 X_{t-6} + \epsilon_t,$$

and we use function in R to compute the value of AIC, BIC and EBIC, discovering which information criterion has the best performance.

In Table 3.4, since the maximum order of model is 2, (i.e. the true model has the maximum order in the model set,) the performance of AIC is the best, but if the maximum order of model is larger than 2, the performance of AIC is the worst, and it has the maximum order tendency in the model sets. Therefore, we are unable to determine whether the best performance of AIC, when the maximum order equals 2, is based on the maximum order tendency or not. EBIC also has the same problem, maximum order tendency, but not so serious, better than AIC a little bit. Overall, BIC has the best performance, but not very good, when the order is greater than 3, only about half of the correct model selection rate.

3.1.3 Log-Normal Distribution vs. Exponential Distribution

Suppose we have a data set $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$, which is generated from the log-normal distribution $\ln N(0,1)$. We want to use criterions to help us find the true distribution. Since our problem is finding the true distribution, it is not involved in the models of different dimension, we only consider the comparison of models by AIC and BIC, regardless of EBIC.

Consider the following two models:

$M_1 : \ X_i \sim \ln N(\mu, \sigma^2), \ i = 1, 2, \cdots, n$ (true)

$M_2 : \ X_i \sim Exp(\lambda), \ i = 1, 2, \cdots, n,$

and we use the formulas of AIC and BIC to help us determine which model is true distribution.

For M_1, the probability density function of a log-normal distribution is

$$f_X(x|\mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\},$$

where $\mu = \ln(E(X)) - \frac{\sigma^2}{2}$ and $\sigma^2 = \ln\left(1 + \frac{Var(X)}{[E(X)]^2}\right)$. Let $E(X) = \bar{X}$ and $Var(X) = s_X^2$, then we use the estimators of μ, $\hat{\mu} = \ln(\bar{X}) - \frac{\hat{\sigma}^2}{2}$, and σ^2, $\hat{\sigma}^2 = \ln\left(1 + \frac{s_X^2}{\bar{X}^2}\right)$, to replace the parameters μ and σ^2 respectively. For M_2, the probability density function of an

exponential distribution is

$$f_X(x|\lambda) = \lambda \exp\{-\lambda x\},\ x \geq 0,$$

where $\lambda = \frac{1}{E(X)}$. Let $E(X) = \bar{X}$, then we have the estimator of λ, $\hat{\lambda} = \frac{1}{\bar{X}}$, and use it to replace the parameter λ similarly.

In Table 3.5, when n is not large enough, either AIC or BIC are only about half of the correct model selection rate. With n greater, the correct model selection rates of AIC and BIC will increase, and when $n = 1000$, the correct model selection rates of AIC and BIC are almost 1.

Table 3.5: Probability of model selection respectively using AIC and BIC within M_1 and M_2. Each case simulated 1000 times.

n	Criterion	Log-N	Exp	n	Criterion	Log-N	Exp
10	AIC	0.458	0.542	100	AIC	0.842	0.158
	BIC	0.412	0.588		BIC	0.805	0.195
20	AIC	0.549	0.451	200	AIC	0.947	0.053
	BIC	0.448	0.552		BIC	0.923	0.077
30	AIC	0.631	0.369	500	AIC	0.992	0.008
	BIC	0.528	0.472		BIC	0.989	0.011
50	AIC	0.697	0.303	1000	AIC	0.999	0.001
	BIC	0.606	0.394		BIC	0.999	0.001

3.2 High Dimensional Model ($p > n$)

Suppose there are p covariates $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ in a data, but the true model is

$$\mathbf{Y} = 1 \cdot \mathbf{X}_1 + 1 \cdot \mathbf{X}_2 + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)^T$, $\mathbf{X}_i = (X_{1i}, X_{2i}, \cdots, X_{ni})^T$, $i = 1, 2, \cdots, p$ and each component of $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$ is normally distributed independent with mean 0 and variance 1. The data of covariates $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ are generated separately from standard normal

distribution.

Consider the following model sets:

$$S_1 = \{M_j : \mathbf{Y} = \beta_j \mathbf{X}_j + \boldsymbol{\epsilon},\ j = 1, 2, \cdots, p\}$$

$$S_2 = \{M_{j,r} : \mathbf{Y} = \beta_j \mathbf{X}_j + \beta_r \mathbf{X}_r + \boldsymbol{\epsilon},\ j, r = 1, 2, \cdots, p,\ j \neq r\} \text{ (true)}$$

$$S_3 = \{M_{j,r,s} : \mathbf{Y} = \beta_j \mathbf{X}_j + \beta_r \mathbf{X}_r + \beta_s \mathbf{X}_s + \boldsymbol{\epsilon},\ j, r, s = 1, 2, \cdots, p,\ j \neq r \neq s \neq j\},$$

and we compute the value of AIC, BIC and EBIC of each model in the above model sets, finding which model sets has the model of minimum value.

Table 3.6: Probability of model selection respectively using AIC, BIC and EBIC within S_1 to S_3 and the status of model selection if given the model set S_2 or S_3. Each case simulated 100 times and the sample size n is 30, the parameter γ of EBIC fixed to 1.0.

P	Criterion	S_1	S_2	S_3	pr($M_T \vert S_2$)	pr($M_T \subset M_C \vert S_3$)
30	AIC	0	0	1	0	1
	BIC	0	0.08	0.92	1	1
	EBIC	0.04	0.71	0.25	1	1
35	AIC	0	0	1	0	1
	BIC	0	0.04	0.96	1	1
	EBIC	0.08	0.69	0.23	1	1
40	AIC	0	0	1	0	1
	BIC	0	0	1	0	1
	EBIC	0.08	0.72	0.2	1	1
45	AIC	0	0	1	0	1
	BIC	0	0.04	0.96	1	1
	EBIC	0.09	0.71	0.2	1	1
50	AIC	0	0	1	0	1
	BIC	0	0.02	0.98	1	1
	EBIC	0.12	0.64	0.24	0.969	1

In Table 3.6, the result of EBIC with respect to the AIC and BIC is pretty good; BIC only has less than one tenth correct rate, AIC completely tends to large model set

(the probability of S_2 being chosen is 0). In addition, we further discuss the situation of model selection if S_2 or S_3 being chosen. If S_2 being chosen, then the rate of choosing true model is greater than 96.9%; while in the case of S_3 being chosen, the model which be chosen must include the true covariates (X_1 and X_2), that is, the covariates of the model is $\{X_1, X_2, X_3\}$, $\{X_1, X_2, X_4\}$, etc.

Table 3.7: Probability of model selection respectively using AIC, BIC and EBIC within M_1 to M_5 under different model space p. Each case simulated 1000 times and the sample size n is 100, the parameter γ of EBIC fixed to 1.0.

p	Criterion	M_1	M_2	M_3	M_4	M_5
100	AIC	0	0.746	0.131	0.085	0.038
	BIC	0	0.953	0.045	0.002	0
	EBIC	0	0.998	0.002	0	0
200	AIC	0	0.764	0.12	0.077	0.039
	BIC	0	0.962	0.035	0.003	0
	EBIC	0	0.999	0.001	0	0
300	AIC	0	0.751	0.14	0.075	0.034
	BIC	0	0.964	0.033	0.003	0
	EBIC	0	1	0	0	0
400	AIC	0	0.749	0.136	0.078	0.037
	BIC	0	0.958	0.039	0.003	0
	EBIC	0	0.999	0.001	0	0
500	AIC	0	0.757	0.119	0.077	0.047
	BIC	0	0.965	0.032	0.003	0
	EBIC	0	1	0	0	0

Now we consider the similar situation with previous case $n > p$. Consider the following models:

$$M_1 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \boldsymbol{\epsilon}$$

$$M_2 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \boldsymbol{\epsilon} \text{ (true)}$$

$$M_3 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \boldsymbol{\epsilon}$$

$$M_4 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \beta_4 \mathbf{X}_4 + \beta_5 \mathbf{X}_5 + \boldsymbol{\epsilon}$$

$$M_5 : \ \mathbf{Y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \cdots + \beta_{10} \mathbf{X}_{10} + \boldsymbol{\epsilon},$$

and we compute the value of AIC, BIC and EBIC, finding which information criterion has the best performance.

In Table 3.7, even in the situation p greater than n ($n = 100$), AIC, BIC and EBIC still have very good performance, EBIC even achieve almost error-free result. This means that the additional correction term of EBIC, based on BIC, indeed can eliminate the large models tendency.

4 Conclusion

In the situation n greater than p, three information criterions mentioned in our study still have the large models tendency, especially when the data is generated from autoregressive model. Inversely, in the situation p greater than n, it seems that we get a good solution when the data is generated from standard normal distribution. We maybe compare the data which is generated from other distributions, or furthermore, comparing and analyzing the real data in the future.

References

[1] Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, **19**, 716-723.

[2] Burnham, K. P. and Anderson, D. R. (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Method & Research*, **33**, 261-304.

[3] Chen, J. and Chen, Z. (2008), "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, **95**, 759-771.

[4] Edwards, D., Abreu, G. C. G. and Labouriau, R. (2010), "Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests," *Edwards et al. BMC Bioinformatics*, **11**:18.

[5] Konishi, S. and Kitagawa, G. (2008), Information Criteria and Statistical Modeling, *New York: Springer*.

[6] Laplace, P. S. (1774), "Memoir on the probability of causes of events," *Laplace's Oeuvres complètes*, **8**, 27-65.

[7] Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, **6**, 461-464.

[8] Shumway, R. H. and Stoffer, D. S. (2011), Time Series Analysis and Its Applications: With R Examples, 3rd ed, *New York: Springer*.

[9] Wang, Y. and Liu, Q. (2006), "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock-recruitment relationships," *Fisheries Research*, **77**, 220-225.