Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operational Research

Model Selection & Model Averaging methods applied to Apple Inc. common stock data of historical quotes

Chris de Waal (534415)

Supervisor:	Xiaomeng Zhang
Second assessor:	Dr. Carlo Cavicchia
Date final version:	1st July 2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

The problem this paper aims to solve is how model averaging methods (MMA and FMA) compare to model selection methods (AIC and BIC) in a financial data application, such as closing stock prices of Apple Inc. common stock historical quotes of one year. Model selection methods based on AIC and BIC are used on financial data of the stock Apple Inc. of one year. Next to that, model averaging methods such as MMA and FMA are used on Apple Inc. stock data. After that, the results of these methods are compared to each other and evaluated according to some appropriate evaluation criteria to see how they relate to each other. This research broadens the empirical applications of these methods from its theoretical background. The main results are that by applying MA methods to the data instead of MS methods to account for model uncertainty, the models will have a higher predictive accuracy according to the evaluation criteria MSE. The data that is used is Apple Inc. common stock historical quotes of one year from the website of Nasdaq (2024).

1 Introduction

In this section, the theoretical background of model selection and model averaging methods is described as first, after that the research question is formulated, following that the motivation of the research problem is explained and finally the relevance for the subject is outlined.

To begin with, model selection (MS) and model averaging (MA) are two important methods in statistical modeling and predictive analytics. Model selection involves choosing the best 'single' model from a set of candidate models. Various model selection criteria have been suggested according to optimality factors including minimizing the estimated prediction risk such as AIC (Akaike, 1973). Another model selection criteria that is suggested according to optimality factors is asymptotically maximizing the posterior probability of a model such as BIC (Schwarz, 1978). Model selection relies on criteria that balance model fit and complexity. AIC (Akaike, 1973) and BIC (Schwarz, 1978), as an example for model selection criteria to choose a best model, penalize the likelihood of the model by the number of parameters to prevent over fitting. The selected model may not be robust to model uncertainty and may perform poorly if the chosen model is not the true model according to the research of Chatfield (1995), Draper (1995) and Yuan & Yang (2005). To address this issue, model averaging provides a robust alternative by combining predictions from multiple models, thereby enhancing performance and reliability in the presence of model uncertainty and also reduce modeling biases. Model averaging is based on the principle that taking the average of a set of models can mitigate the risk of selecting a poor model. This method can therefore improve predictive performance and reliability by accounting for model uncertainty and reducing the risk of over fitting which is a challenge and problem in model selection methods. There are a lot of different model averaging methods, such as Mallows Model Averaging (Hansen, 2007), Bayesian Model Averaging (Adrian E. Raftery & Hoeting, 1997) and Frequentist Model Averaging (Claeskens, 2003). This research is going to use two of them in a financial application that is related to stocks.

This paper is therefore going to investigate the following research question:

How do model averaging methods compare (improve predictive accuracy or not) to model selection methods in a financial application, such as closing stock prices from Apple Inc.?

Accurate stock price prediction is crucial in finance. This is due to the fact that market volatility and complexity exists. But, traditional model selection methods like AIC (Akaike, 1973) and BIC (Schwarz, 1978) often do not include model uncertainty which may result in sub-optimal predictions. An interesting alternative is offered by model averaging techniques where multiple models are used together to enhance forecasting performance such as Mallows Model Averaging according to the approach of Hansen (2007) and Frequentist Model Averaging (FMA) according to the paper of Claeskens (2003). However, we don't know enough about how well these methods work when applied to financial problems, specifically applied to Apple Inc.'s common stock historical quotes of the last year. The problem with this kind of research is that most research has looked at these methods, but only from theoretical or non-financial perspectives without making direct comparisons with real financial data. And also most of the studies have focused on static models while ignoring dynamic nature of financial markets influenced by various macroeconomic factors. Thus, it is important that further study is carried out to confirm through empirical analysis if model averaging techniques are more efficient than conventional procedures such as model selection methods for predicting stock prices. Better risk management and investment decisions require accurate forecasts. The practical application of the benefits and limitations of using model averaging techniques in the case of forecasting Apple's share price will be investigated using actual financial data from the website of Nasdaq (2024).

There are several reasons why this research is important and relevant. The first reason is that this research fills in the empirical comparison gap amongst the model selection and model averaging methods in academic literature. Next to that, this study can improve financial stability by increasing risk management and improve investment strategies in stocks such as Apple Inc. common stock. Lastly, this research is relevant because it develops advanced financial models that will go a long way to benefit in the wider finance industry.

2 Literature

Within this section, an overview of the existing literature pertaining to model selection and model averaging in empirical applications (also finance related) is provided. First, a summary is given of the results that have been obtained in previous research on this topic. Following that, a summary is given of how this research relates to the existing literature. After that, the structure of the remaining sections of this research are outlined.

Over time, an array of studies has delved into the model selection methods and model averaging methods. Most of the times, those studies were purely theoretical papers like the research of Peng & Yang (2022) that answers if model averaging methods offer any significant improvement over model selection methods in regression estimation without using an empirical application. The research of Johnson & Omland (2004) uses an empirical application for model selection methods that are well developed in certain fields, like in molecular systematics and mark recapture analysis. But, model selection methods are less used in other areas, like in evolution and ecology where those methods could be useful. By investigating this, researchers in ecology and evolution find a valuable alternative to traditional null hypothesis testing. Those aforementioned fields are a small subset of fields where the use of model selection methods is needed. So, other empirical applications that need model selection methods to reduce estimation risk instead of evolution and ecology could be in finance, economics or marketing. The study of Hartman & Groendyke (2013) uses model selection methods (based on AIC, BIC and DIC) and model averaging methods in an empirical application that is finance related. Namely, financial risk management where the MS and MA methods are applied to total return data from the S&P 500. Their research specifically considers MS methods that provide posterior probabilities being the best, enabling MA and providing deeper insights into the relationships between the financial risk management models. While model selection methods in empirical applications provide a robust framework for the best fitting models, model averaging methods offer an additional approach to take model uncertainty into account by integrating multiple models. So, while the literature that pertains to model selection methods in empirical applications have been discussed in this paragraph, the literature that pertains to model averaging methods in empirical applications will be provided in the next paragraph.

Another empirical application where model averaging methods are used for research is in the paper of Moral-Benito (2015). This paper reviews the literature on model averaging specifically on the applications in economics. Their main conclusion is that model averaging takes into account the model uncertainty surrounding the selection of controls in a natural manner in applications that relate to economics. To understand the role of model averaging in addressing model uncertainty even better (also in other applications than economics), the review of Wasserman (2000) offers an in-depth examination of objective Bayesian methods. These methods described by Wasserman (2000) highlight the use of non informative priors and the practical aspects of the implementation. The main points of the research of Wasserman (2000) are that the Bayesian model selection and averaging methods are straightforward and unified approaches. BIC (Schwarz, 1978) provides a useful approximation for well-behaved models and large sample sizes. Intrinsic Bayes factors are suggested for nonstandard problems and averaging predictions from multiple models is highlighted. Lastly, the paper emphasizes objective Bayesian methods instead of subjective priors and acknowledge robust Bayesian inference as a bridge between objective and subjective Bayesian methods. Building on the study of Wasserman (2000) that is a detailed analysis of Bayesian methods in MS and MA, the discussion is extended to practical approaches for addressing model uncertainty in linear regression models in the study of Adrian E. Raftery & Hoeting (1997). 'Occam's window' and a Markov chain Monte Carlo method, which are the Bayesian model averaging methods used in the paper, both improve predictive performance by considering a subset of models or directly approximate the exact solution. The conclusion of the paper of Adrian E. Raftery & Hoeting (1997) is that the choice between the two Bayesian model averaging methods depends on the application. 'Occam's window' is better and faster for inferring relationships between variables, while Markov chain Monte Carlo is better in making predictions or investigating posterior distributions when computation time is not a constraint. Both of the two Bayesian model averaging methods are flexible and can be used for inference and prediction for accounting model uncertainty. Following these explorations of Bayesian MA methods, the paper of Goodness Aye (2015) further explores advanced predictive models for financial applications, specifically focusing on variables of gold returns. Results of this study show that dynamic model selection (not dynamic model averaging) is the best across all different forecast horizons. The exchange rate has the strongest predictive power to predict the price of gold in the research of Goodness Aye (2015).

This research in model selection & model averaging methods applied to a financial application (historical quotes of Apple Inc. common stock) relates to the existing literature by broadening the empirical applications of these methods from its theoretical background. The theoretical paper of Peng & Yang (2022) discusses model selection and model averaging methods and no empirical application is provided (only related simulations that confirm what they theoretically argued in their study). Yet, studies from Johnson & Omland (2004) and Hartman & Groendyke (2013) show that MS and MA methods are being actively used in fields, such as ecology, evolution and financial risk management, in practice. The use of MS and MA methods to economic data and an in-depth examination of objective Bayesian methods, is guided by the insights from Moral-Benito (2015) and Wasserman (2000). The significance of managing model uncertainty and the usefulness of Bayesian techniques are the main topics of these studies. More recent papers that have investigated Apple stock data are papers from Rai et al. (2018) and Banerjee (2020). In the research of Rai et al. (2018) stock market movements are analyzed using several internet sources including stock prices and transaction volumes data of Apple and other big companies. The study of Banerjee (2020) forecasts Apple Inc. common stock prices using the S&P 500. So, Apple data is popular to use for research, however there are hardly any papers to find where MS and MA methods are applied to Apple stock data. That is where this research is breaking new ground. Furthermore, this study advances knowledge about the potential applications of MS and MA in the finance industry by examining their predictive power for stock returns. The results of Goodness Aye (2015) on gold returns and dynamic model selection are consistent with the present study. This demonstrates the potential for enhancing prediction accuracy and lowering estimation risk in a financial application (like Apple Inc. stock), in addition to validating the adaptability of MA methods in a variety of empirical applications.

To address the research question, the data is presented first in Section 3, accompanied by the summary statistics. Following that, the methodology is outlined in Section 4, where Subsection 4.1 is the methodology for the replication part and Subsection 4.2 the methodology for the extension part. The results are displayed in Section 5, where Subsection 5.1 are the results for the replication part and Subsection 5.2 the results for the extension part. The conclusion is presented at the end of this paper in Section 6.

3 Data

In this part, the data that is used is described and how to obtain that data. In Subsection 3.1 this is done for the data of the replication part of the research of Peng & Yang (2022). In Subsection 3.2 this is done for the data of the extension part of this research.

3.1 Replication

The data that is used in Section 5 for the simulation settings to compare MS with MA of the paper of Peng & Yang (2022)is not a certain dataset. The data is generated or produced according to a linear regression model, which is the data generating process (DGP):

$$y_i = \mu_i + e_i = \sum_{j=1}^{p_n} \theta_j x_{ji} + e_i, \quad i = 1, ..., n \tag{1}$$

where $p_n = \lfloor 5n^{2/3} \rfloor$, $x_{1i} = 1$, the leftover x_{ji}'s are independently generated from $N(0, 1)$, the random errors e_i are iid from $N(0, \sigma^2)$ and are independent of x_{ji}'s and n is the number of observations that varies with $n = 50, 150$ or 400 for Figures 1 and 2 in Section 5 of the research of Peng & Yang (2022) and n varies with $50, 100, 500, 1000$ and 5000 for Figures 3 and 4 in Section 5 of the research of Peng & Yang (2022). The population $R^2 = Var(\mu_i)/Var(y_i)$, which looks like Signal-to-Noise-Ratio (SNR), is regulated in the range of [0.1, 0.9] via the parameter σ^2. Two cases are considered with slowly decaying coefficients and fast decaying coefficients:

- Slowly decaying coefficients: $\theta_j = j^{-\alpha_1}$ and α_1 is set to be 1, 1.5 or 2.

- Fast decaying coefficients: $\theta_j = \exp(-\alpha_2 j)$ and α_2 is set to be 1, 1.5 or 2.

To obtain the data that is generated according to the previous specified process that comes from the paper of Peng & Yang (2022) simulations are done in a program like Python to generate the synthetic data. The data is simulated according to the theoretical model (DGP), a linear regression model as specified above.

3.2 Extension

To address the research question, data used in this analysis was obtained from the freely available website of Nasdaq (2024), which is a dataset about Apple Inc. common stock historical quotes of one year. It provides a daily record of Apple Inc. common stock performance over the period of 1 year, so from the 31th of May in 2023 till the 30th of May 2024 including detailed price movements of the stock and trading volume. The data includes the following column, one variable to predict (also column) and four covariates (which are also columns): 'Date', 'Close/Last', 'Volume', 'Open', 'High' and 'Low'.

The column 'Date' gives the specific trading date in the format month/day/year (American style). This column captures each trading day within the one-year period. So, all the weekend dates are missing out in this dataset, because these are no trading days. Next to that, all the US holiday calendar dates are also missing out in this dataset, because these are not trading days as well. To be precise, it is about New Year's Day (January 1, 2024), Martin Luther King Jr. Day (January 15, 2024), Washington's Birthday (February 19, 2024), Memorial Day (May 27, 2024), Independence Day (July 4, 2023), Labor Day (September 2, 2023), Columbus Day (October 14, 2023), Veterans Day (November 11, 2023), Thanksgiving Day (November 28, 2023) and Christmas Day (December 25, 2023). Because of this, the dataset has 252 observations (rows in Excel data document).

The target variable (to predict variable) 'Close/Last' captures the closing price of Apple Inc. stock on the given date (given row in dataset), represented in USD. The 'Close/Last' price in USD is the final price at which the stock traded during that day after regular trading hours.

The covariate (or predictor) ‘Volume’ stands for the total number of shares of Apple Inc. that were exchanged during that trading day. This column gives information about the trading activity and liquidity of the stock on a particular day.

The covariate (or predictor) ‘Open’ represents the opening price of Apple Inc. stock on the given date (given row in dataset), also represented in USD. The ‘Open’ price is the price in USD at which the stock first traded upon the market opening that day.

The covariate (or predictor) ‘High’ indicates the highest price of Apple Inc. stock that was reached during that particular trading day, in USD.

The covariate (or predictor) ‘Low’ indicates the same as the covariate ‘High’, but then for the lowest price of Apple Inc. stock that was reached during that specific day.

How do the covariates actually relate to each other? To begin with, the difference between the ‘High’ and ‘Low’ prices indicate the volatility of the Apple Inc. stock on the given day. Large differences between ‘High’ and ‘Low’ indicate high volatility. On top of that, there is a positive correlation between ‘Volume’ and the size of price changes (difference between ‘Open’ and ‘Close/Last’ or between ‘High’ and ‘Low’). This positive correlation can be seen in Table 7 in Section A.4. High trading volumes lead to larger price movements. Lastly, there is a relationship between ‘Open’ and ‘Close/Last’ prices. They show the overall trend of the Apple Inc. stock price during the day whether the price increased, decreased or stayed relatively stable. This overall trend can be seen in Table 8 in Section A.4.

Secondly, we will zoom in if the data contains missing observations or outliers that can impact the results when performing model selection and model averaging techniques. This dataset about historical quotes of the common stock Apple Inc. does not contain any missing values, so we do not have to have an approach to handle these missing observations as they are not there. Next, we will identify if there are any outliers present in the data. Common approaches to detect outliers are: empirical relations in normal distributions and IQR (Inter Quartile Range) in skewed distributions. In our case, we use IQR as an outlier detection method. For the reason that the data follows a slightly skewed distribution (See Table 1 for skewness of the target variable and the covariates of the data). The IQR for each covariate of the data is calculated and identified if they lie beyond the 1.5 times the IQR from the first and third quartiles. As a result in the dataset, only the covariate ‘Volume’ contains outliers, namely 16 observations. The other covariates in the data do not contain outliers. Common approaches to handle outliers are: trimming (remove extreme values), capping (replace outlier values with nearest non-outlier values), discretization (categorize outliers into specific group with same behavior), treating them as missing values or transform them (logarithmic transformation to reduce impact of them). Capping is the most appropriate approach in this case, because outliers in the ‘Volume’ column of the stock trading data of Apple Inc. can vary significantly and therefore handle these outliers in a proper way without losing valuable information. The outliers in the ‘Volume’ column have been capped at the 5th and 95th percentiles. Still, the number of outliers in that column stays the same, because the extreme values were replaced but are still considered as outliers according to the IQR method. Taking this all into account (missing observations and outliers), the data is cleaned and ready for analysis.

Finally, some relevant statistics about the data are provided in Table 1.

Table 1: Summary Statistics of Apple Inc. common stock (AAPL) historical quotes from 31th of May 2023 till the 30th of May 2024

	Mean	**Median**	**Standard Error**	**Skewness**	**Kurtosis**
Close/Last	182.39	182.66	0.53	-0.11	-1.13
Volume	58,258,780	53,261,680	1,184,039	2.06	6.02
Open	182.32	182.43	0.54	-0.09	-1.15
High	183.80	184.23	0.52	-0.11	-1.09
Low	180.93	181.15	0.54	-0.04	-1.18

4 Methodology

In this segment, the econometric methods and techniques that will be applied in this research are explained and why they are appropriate for this research specifically to answer the research question. In Subsection 4.1 this is done for the econometric methods of the replication part of the research of Peng & Yang (2022). In Subsection 4.2 this is done for the econometric methods of the extension part of this research.

4.1 Replication

Several simulation settings are considered to compare MS method(s) with MA method(s). This is for the reason to illustrate the theoretical results presented in the paper of Peng & Yang (2022). AIC (Akaike, 1973) and BIC (Schwarz, 1978) are chosen as MS methods and MMA (Hansen, 2007) is chosen as MA method as representative. The data of these replication methods of the study of Peng & Yang (2022) is described in Subsection 3.1. The method(s) that is/are used to replicate the results of the research of Peng & Yang (2022) is that M_n nested approximating models are considered that consists of the first s regressors for $1 \leq s \leq M_n$. After that, all the candidate models are estimated by ordinary least squares (OLS). Then, the precision of each procedure on the observed data is measured in terms of squared L_2 loss at 10000 new independently drawn covariates. This squared L_2 loss or risk has the following formula:

$$R(f, \hat{f}) = \frac{1}{n} E \left\| f - \hat{f} \right\|^2 = \frac{1}{n} E \sum_{i=1}^{n} \left[f(x_i) - \hat{f}(x_i) \right]^2 \tag{2}$$

where n in Equation 2 is the number of replications, $f(x_i)$ is the true estimator function (θx_i), $\hat{f}(x_i)$ is the estimated estimator function based on the model ($\hat{\theta} x_i$) and the x_i's are the new independently covariates that are drawn (10000 according to the research of Peng & Yang (2022)).

This procedure is replicated 1000 times ($n = 1000$ in Equation 2) to compute the risk functions of MS and MA methods. In each simulation, the risks of the MS methods and the MA methods are divided by the risk of MMA. So, the risk of the MA method (MMA in this case) is divided by itself.

In the simulation settings, the effects of n (sample size), $R^2 = Var(\mu_i)/Var(y_i)$ (which looks like SNR) and coefficient decaying order (slowly and fast) are investigated on relative performances of MS and MA in two different ways. The first way is presented in Figures 1 and 2, that implement the approach of Hansen (2007) to compare the risk as a function of the population R^2 for different n (n = 50, 150 and 400). When looking at the settings of Hansen (2007), M_n is set to be 11,16 and 22 under the three different sample sizes. The second way is displayed in Figures 3 and 4, that investigate the relative risks of the MS and MA methods as a function of n, where n increases from 50 to 5000 on a logarithmic scale. In Case 1, where we have slowly decaying coefficients M_n is set to be $3\left(\frac{n}{\sigma^2}\right)^{\frac{1}{2\alpha_1}}$. In Case 2, where we have fast decaying coefficients M_n is set to be $\frac{4.5}{\alpha_2}\log\left(\frac{n}{\sigma^2}\right)$ These are multiples of the optimal model size in each case of decaying coefficients. The second way (Figures 3 and 4) correspond more to asymptotic statements in the main theorems of the paper of Peng & Yang (2022). The first way (Figures 1 and 2) correspond to the information on the impact of the SNR at a given sample size.

4.1.1 Model selection method(s) based on AIC and BIC

The model selection methods that are executed in the research of Peng & Yang (2022) and on the financial data in this research (extension) are based on AIC and BIC. Here is an overview of these methods (description of symbols in Equations can be found in Section A.3):

Akaike Information Criterion (AIC) (Akaike, 1973):

$$2k - 2\log(L) \tag{3}$$

Bayesian Information Criterion (BIC) (Schwarz, 1978)

$$k\log(n) - 2\log(L) \tag{4}$$

4.1.2 Mallows Model Averaging

The model averaging method that is executed in the research of Peng & Yang (2022) and on the financial data in this research (extension) is Mallows Model Averaging (MMA) according to the approach of Hansen (2007). MMA is a Model Averaging (MA) technique that selects the weights of the model by minimizing a Mallows criterion, which is an estimate of the average squared error from the model average fit. Following that, the formulas are as follows (for the model average estimator in 5 of Θ_M in matrix notation and Mallows criterion for the model average estimator in 6 also in matrix notation, so not in vector notation θ_j like what is done in Subsection 3.1, this to present the Equations 5 and 6 in an easier and more compact way):

Model average estimator:

$$\hat{\Theta} = \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\Theta}_m \\ 0 \end{pmatrix} \tag{5}$$

where m are the approximating models, $M = M_n \leq n$ and $m \leq M$ where M is the max model size (so max amount of models that can be considered), w_m the weight that corresponds to the m-th approximating model and $\hat{\Theta}_m$ is the least squares estimate of Θ_m.

Mallows criterion for the model average estimator:

$$C_n(W) = (Y - X_M\hat{\Theta})'(Y - X_M\hat{\Theta}) + 2\sigma^2 k(W) \tag{6}$$

where $Y = X_m\Theta_m + b_m + e$ or $Y = (y_1, ..., y_n)'$, X_M is the $n \times k_M$ matrix with the with the ij-th element x_{ji}, $\hat{\Theta}$ is the model average estimator from 5, σ^2 unknown and $k(W)$ the effective number of parameters.

The σ^2 is going to be replaced with an estimate according to the approach of Hansen (2007). The Mallows criterion in Equation 6 may be used to select the weight vector W. First, the empirical Mallows selected weight vector is introduced:

$$\hat{W} = \arg\min_{W \in \mathcal{H}_n} C_n(W) \tag{7}$$

where $\mathcal{H}_n$ in Equation 7 is the non negativity and summation constraint that follows the following notation:

$$\mathcal{H}_n = \left\{ W \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\} \tag{8}$$

There is no closed form solution to Equation 7, so the weight vector must be found numerically. Therefore Equation 6 can be written in the following form:

$$C_n(W) = W'\bar{e}\bar{e}'W + 2\sigma^2 K'W \tag{9}$$

where $W = (w_1, ..., w_M)'$, $\bar{e} = (\hat{e_1}, ..., \hat{e_M})$ is the $n \times M$ matrix of residuals where $\hat{e_m}$ is the $n \times 1$ residual vector from the mth model and $K = (k_1, ..., k_M)'$ is the $M \times 1$ vector of the number of parameters in the M models.

Then Equation 9 is linear quadratic in W. The solution of Equation 7 minimizes Equation 9 subject to Equation 8. This is a typical quadratic problem for which numerical algorithms can be used that are available.

4.2 Extension

4.2.1 Frequentist Model Averaging

The model averaging method that is performed on the financial data of Apple stock in this research is Frequentist Model Averaging (FMA) according to the approach of Buckland & Augustin (1997) that is described in the paper of H. Wang & Zou (2009). They introduce the model averaging estimator of a parameter μ as follows:

$$\hat{\mu}_B = \sum_{k=1}^{K} \lambda_k \hat{\mu}_k \tag{10}$$

where k is the k-th candidate model, $\hat{\mu}_k$ is the estimator of μ of the k-th candidate model, λ_k the weight associated with $\hat{\mu}_k$ and all the weights (λ_k's) of the candidate models must sum up to 1 as a constraint.

The weights are in practice estimated according to the following information criteria:

$$I_k = -2\log(L_k) + \ell, \tag{11}$$

where L_k is the maximized likelihood function under the k-th model and ℓ is a penalty function for the number of parameters and/or observations. Then following the approach of Buckland & Augustin (1997), they recommend to use the following formula for the weights:

$$\lambda_k = \frac{\exp(-I_k/2)}{\sum_{i=1}^{K}\exp(-I_i/2)}, \quad k = 1, 2, \ldots, K. \tag{12}$$

If $\ell = 2p$ in 11, where p is the number of parameters, I_k would be AIC. This assumption is made in this paper because it is grounded in the theoretical principles of information theory, practical need to balance model fit and complexity and to simplify the Equation for the weights in 11 and in 12. Therefore the estimator is called smooth AIC estimator with Akaike weight. Theoretical study is not conducted on such estimators, but numerical examples are presented to demonstrate the reliability of them. The weights according to Equation 12 have been extensively used in the literature, for example in the research of Wan & Zhang (2009), Wagenmakers & Farrell (2004) and F.E. Turheimer & Cunningham (2003).

4.2.2 Evaluation criteria

The evaluation criteria that are used after having applied model selection and model averaging methods to evaluate the performance of the models are MSE, MAE and R^2, which can be seen in Table 6 and Table 5 in Section A.2. The formulas of these criteria can be found in Section A.2 (also how the data is split).

5 Results

In this section, the most important empirical findings and relevant results are presented. Next to that, necessary explanations, implications and interpretations are provided to validate the results. In Subsection 5.1 this is done for the results of the replication part of the research of Peng & Yang (2022). In Subsection 5.2 this is done for the results of the extension part of this research.

5.1 Replication

Firstly, the normalized risk functions for AIC, BIC and MMA when $\theta_j = j^{-\alpha_1}$ with $\alpha_1 = 1$ in first row, $\alpha_1 = 1.5$ in second row and $\alpha_1 = 2$ in third row (of the subplots) are presented in Figure 1. The subplots correspond to slowly decaying coefficients.

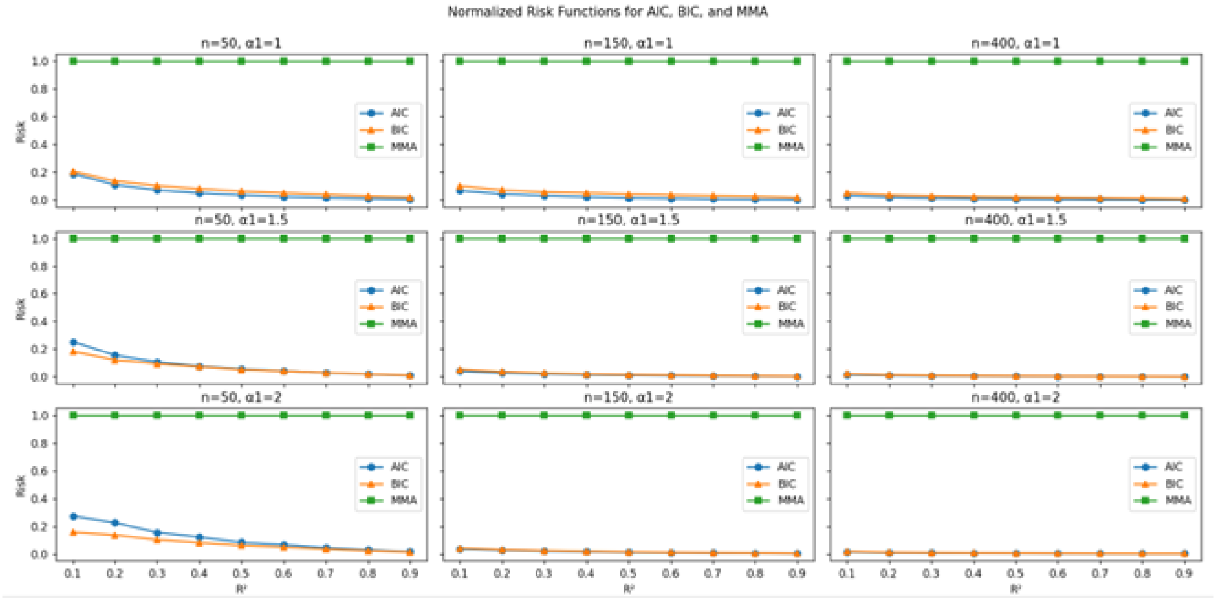

Figure 1: Normalized risk functions for AIC, BIC and MMA when $\theta_j = j^{-\alpha_1}$ with $\alpha_1 = 1$ in first row, $\alpha_1 = 1.5$ in second row and $\alpha_1 = 2$ in third row. The graphs correspond to slowly decaying coefficients.

The y-axis in each subplot represents 'Risk' (squared L_2 risk according to Equation 2) associated with each criterion (AIC, BIC and MMA). The x-axis in each subplot presents the info $R^2 = Var(\mu_i)/Var(y_i)$, which represents the Signal-to-Noise Ratio (SNR) in a slightly different way, controlled in the range of [0.1, 0.9] via the parameter σ^2 (unknown, but estimated according to the approach of Hansen (2007)). These criteria are compared in terms of risk, with AIC and BIC showing lower risks compared to MMA in each of the subplots of Figure 1 (different result than in research of Peng & Yang (2022)). BIC shows even a higher risk compared to AIC in in the first two subplots of Figure 1 for the small values of R^2 and then BIC becomes equal to AIC for bigger values of R^2. If the sample size n increases from $n = 50$ to $n = 400$, the trend is that the risks for AIC and BIC remain relatively similar but become slightly less volatile. MMA stays the same for different values of n, because the risk of MMA is normalized by itself. If the α_1 values increase from $\alpha_1 = 1$ to $\alpha_1 = 2$, the trend is that the risks of AIC and BIC decreases, especially when R^2 increases. However, MMA remains robust for different values of α_1. To conclude, AIC and BIC consistently outperforms MMA in terms of risk for different n and for different parameter values (α_1's). Therefore model selection is a more reliable and robust approach under various conditions (different n and different parameter values) compared to model averaging according to Figure 1 (different result than in research of Peng & Yang (2022)).

The results of Figure 1 do not match the published Figure 1 of the research of Peng & Yang (2022) exactly. Possible reasons for this could be differences in simulation parameters & data generation, differences in model specification & implementation, differences in software and differences in plotting & normalization. Firstly, with differences in simulation parameters it is meant that n, R^2, α_1 and simulation values should be exactly the same between the results of this research and the published figures. Even small deviations could lead to significant differences. This research uses n values of 50, 150 and 400, α_1 values of 1, 1.5 and 2, R^2 values that span the range from 0.1 to 0.9 in equal increments of 0.1 and 500 simulations. The only difference between this research in terms of parameters and the research of Peng & Yang (2022) is that this research uses 500 simulations instead of 1000, as 1000 simulations are challenging. This can cause the difference in results and is a limitation. Secondly, with differences in data generation it is meant that the DGP of this research should be exactly the same as the DGP from the published research. This research uses the linear regression model for response variable y in matrix notation, so $\mathbf{y} = X\Theta + \mathbf{e}$ instead of $y_i = \mu_i + e_i = \sum_{j=1}^{p_n} \theta_j x_{ji} + e_i$. But, this study uses the same p_n, first column to 1 as per the given DGP, same distribution for X and $\mathbf{e}$ and independent elements of X. Next to that, this research uses the same R^2 formula as published research and the σ^2 that is estimated by the formula in the paper of Hansen (2007) with the $\theta_j = j^{-\alpha_1}$ and α_1 is set to be 1, 1.5 or 2. So, the difference in linear regression model for response variable y in matrix notation can cause the difference in results and can be a limitation and the other settings not as they are exactly the same as in Peng & Yang (2022). Thirdly, there could be some mistakes in the Python code or in the order of doing things in the code. This study first introduces parameters in the code, then presents the function to calculate the risk with the given DGP from the research of Peng & Yang (2022) on page 251 in the beginning of Section 5, after that the σ^2 is estimated according to the formula in Hansen (2007) on page 1181 on line 21, following that the AIC and BIC calculations are done across different model complexities following the AIC and BIC formulas out of the paper Hansen (2007) on page 1182, thereafter the models are selected with minimum AIC and BIC, consequently MMA is introduced in the code, then the risk calculation(s) is/are parallelized and finally the results are reshaped and plotted in the code. This can cause different figures than in the pubished research. Fourthly, the MMA based on Mallow's criterion in this research can differ a little bit from the published research of Hansen (2007) because his Gauss procedure to compute MMA least squares estimates is not used in the code in this research. This can cause some different results than the study of Peng & Yang (2022). Finally, there could be some differences in used software, software versions and libraries. This could lead to different figures than the published figures. This research is performed on an Acer i7 laptop 16.0 GB (15.9 GB usable) of RAM, using Python 3.12 (64-bit) as a program. The research of Peng & Yang (2022) is probably done on another laptop or computer with another app or program to perform the results what can lead to minor differences.

Secondly, the normalized risk functions for AIC, BIC and MMA when $\theta_j = \exp(-\alpha_2 j)$ with $\alpha_2 = 1$ in first row, $\alpha_2 = 1.5$ in second row and $\alpha_2 = 2$ in third row are presented in Figure 2. The subplots correspond to fast decaying coefficients.

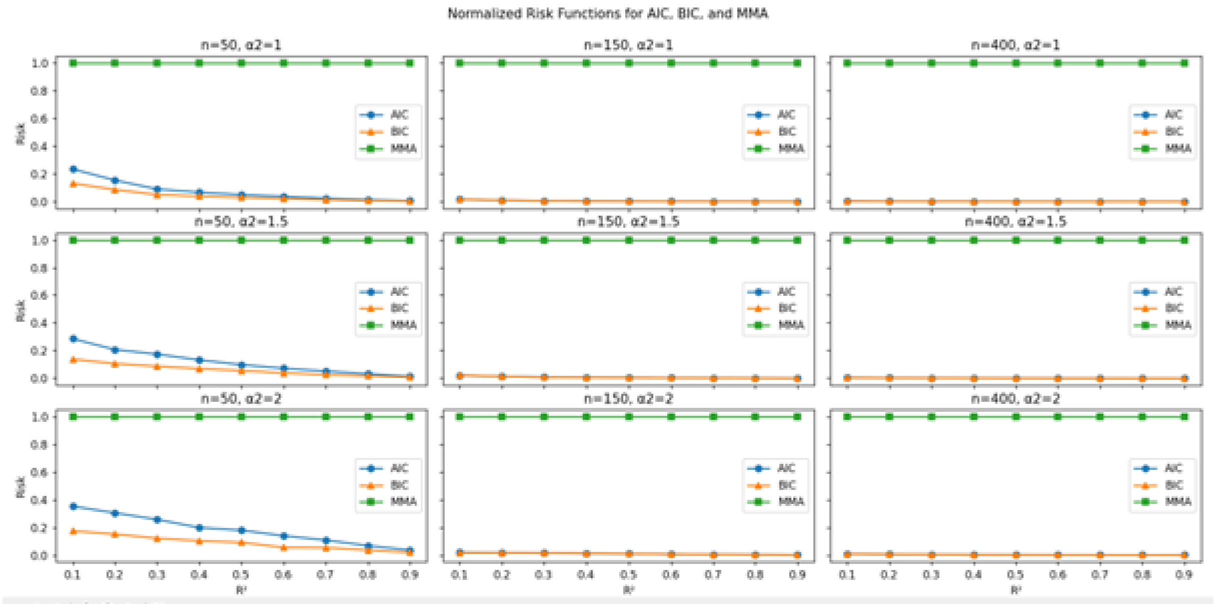

Figure 2: Normalized risk functions for AIC, BIC and MMA when $\theta_j = \exp(-\alpha_2 j)$ with $\alpha_2 = 1$ in first row, $\alpha_2 = 1.5$ in second row and $\alpha_2 = 2$ in third row. The graphs correspond to fast decaying coefficients.

The y-axis in each subplot represents the same as in Figure 1 and the x-axis represents the same as in Figure 1. The criteria are compared in terms of risk, with AIC and BIC showing lower risks compared to MMA in each of the subplots of Figure 2. AIC shows even a higher risk compared to BIC in the first three subplots of Figure 2 when $n = 50$ for small values of R^2 . If the sample size n increases from $n = 50$ to $n = 400$, the trend is that the risks for AIC and BIC remain relatively similar but become slightly less volatile and become eventually the same. MMA stays the same for different values of n, because the risk of MMA is normalized by itself. If the α_2 values increase from $\alpha_2 = 1$ to $\alpha_2 = 2$, the trend is that the risks of AIC and BIC increase, especially when R^2 values are small and AIC and BIC decrease, especially when R^2 values are large. However, MMA remains robust for different values of α_2. To conclude, AIC and BIC consistently outperforms MMA in terms of risk for different n and for different parameter values (α_2's). Therefore model selection is a more reliable and robust approach under various conditions (different n and different parameter values) compared to model averaging according to Figure 2 (different result than in research of Peng & Yang (2022)).

The results of Figure 2 do not match the published Figure 2 of the research of Peng & Yang (2022) exactly and also do not show the same trends in the subplots. Possible reasons for this are the same reasons mentioned that apply to Figure 1 as argued before.

Thirdly, the normalized risk functions for AIC, BIC and MMA when $\theta_j = j^{-\alpha_1}$ with $\alpha_1 = 1$ in first row, $\alpha_1 = 1.5$ in second row and $\alpha_1 = 2$ in third row are presented as a function of n (number of samples) for different values of R^2 (0.25, 0.5 and 0.75) and different values of α_1. The subplots correspond to slowly decaying coefficients.

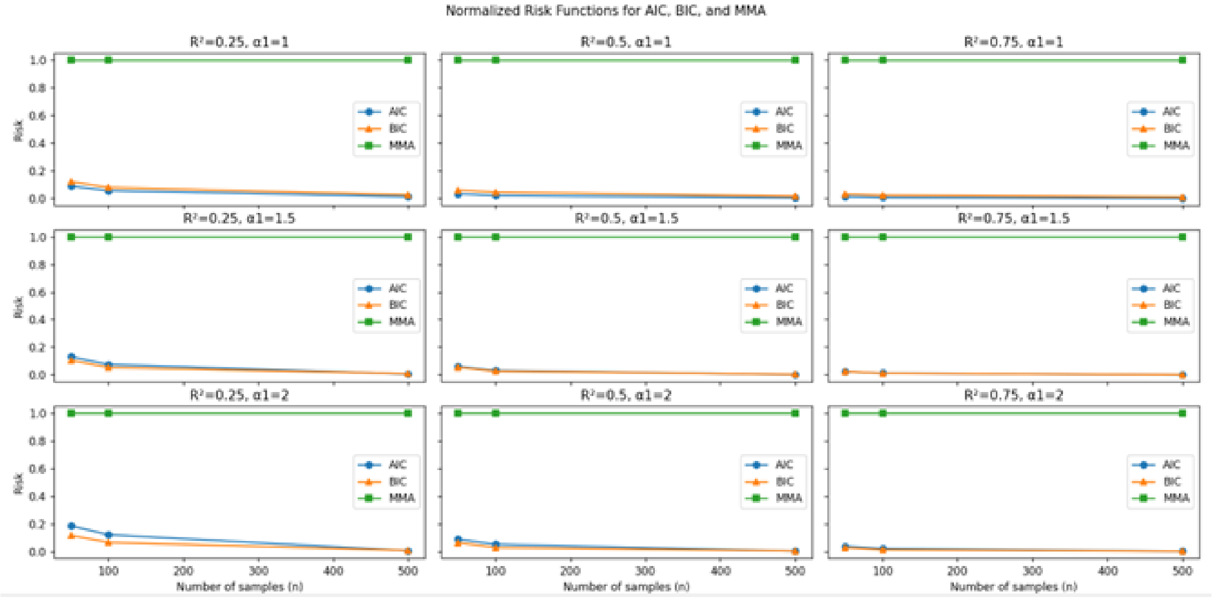

Figure 3: Normalized risk functions for AIC, BIC and MMA when $\theta_j = j^{-\alpha_1}$ with $\alpha_1 = 1$ in first row, $\alpha_1 = 1.5$ in second row and $\alpha_1 = 2$ in third row. The graphs correspond to slowly decaying coefficients. The risk functions for AIC, BIC and MMA are now shown as a function of n (number of samples) for different values of R^2 and α_1 instead of that the risk functions are shown as a function of R^2 in the range of [0.1, 0.9] for different values of n and α_1 like in 1.

The y-axis in each subplot represents 'Risk' (squared L_2 risk or loss according to Equation 2) associated with each criterion (AIC, BIC and MMA). The x-axis in each subplot represents the number of samples n (for n = 50, 100, 500). (That is different than in Figures 1 and 2 and different than in research of Peng & Yang (2022)). These criteria are compared in terms of risk, with AIC and BIC showing lower risks compared to MMA in each of the subplots of Figure 3. BIC shows even a higher risk compared to AIC in the first two subplots of Figure 3 for small values of n and then BIC becomes equal to AIC for bigger values of n. If R^2 (which looks like SNR) increases from $R^2 = 0.25$ to $R^2 = 0.75$, the trend is that the risks for AIC and BIC remain relatively similar but become slightly less volatile and converge to each other. MMA stays the same for different values of R^2, because the risk of MMA is normalized by itself. If the α_1 values increase from $\alpha_1 = 1$ to $\alpha_1 = 2$, the trend is that the risks of AIC and BIC increase a little bit, especially when n is small. However, MMA remains robust for different values of α_1. To conclude, AIC and BIC consistently outperforms MMA in terms of risk for different R^2 and for different parameter values (α_1's). Therefore model selection is a more reliable and robust approach under various conditions (different n and different parameter values) compared to model averaging according to Figure 3 (different result than in research of Peng & Yang (2022)).

The results of Figure 3 do not match the published Figure 3 of the research of Peng & Yang (2022) exactly and also do not show the same trends in the subplots. Possible reasons for this are the same reasons mentioned that apply to Figure 1 as argued before. Besides that, 500 repetitions are executed instead of 1000 which are challenging and a smaller sample size of

$n = 50, 100, 500$ is used instead of $n = 50, 100, 500, 1000$ and 5000 because of computation time issues. This can also cause different results and is a limitation of this study.

Lastly, the normalized risk functions for AIC, BIC and MMA when $\theta_j = \exp(-\alpha_2 j)$ with $\alpha_2 = 1$ in first row, $\alpha_2 = 1.5$ in second row and $\alpha_2 = 2$ in third row are presented as a function of n (number of samples) for different values of R^2 (0.25, 0.5 and 0.75) and different values of α_2 in Figure 4. The subplots correspond to fast decaying coefficients.

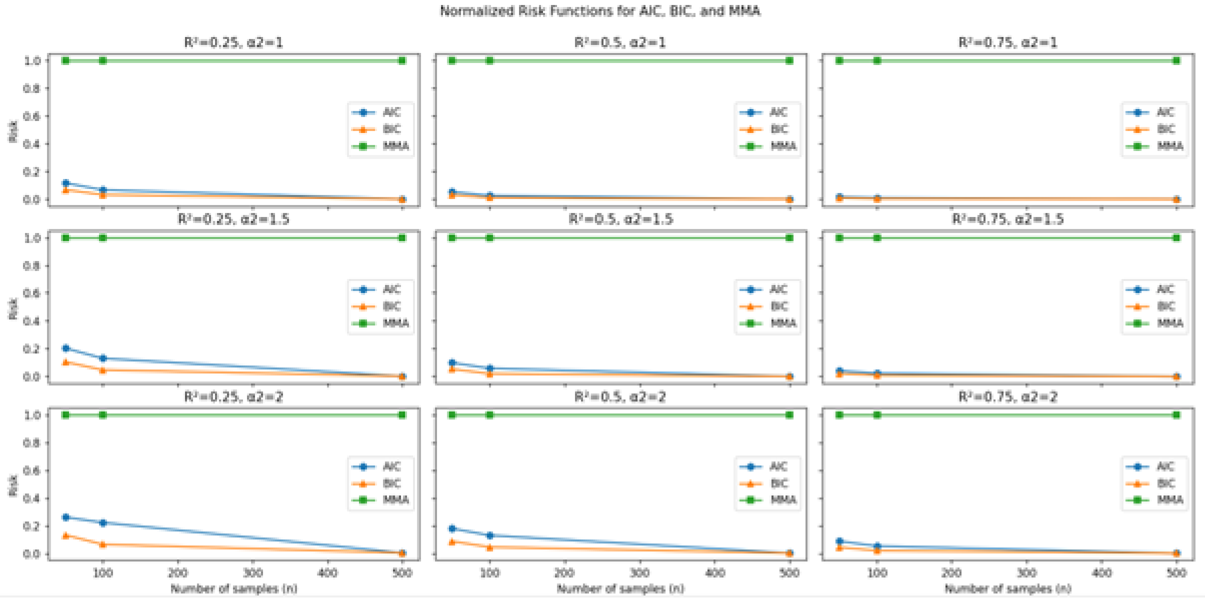

Figure 4: Normalized risk functions for AIC, BIC and MMA when $\theta_j = \exp(-\alpha_2 j)$ with $\alpha_2 = 1$ in first row, $\alpha_2 = 1.5$ in second row and $\alpha_2 = 2$ in third row. The graphs correspond to fast decaying coefficients. The risk functions for AIC, BIC and MMA are now shown as a function of n (number of samples) for different values of R^2 and α_2 instead of that the risk functions are shown as a function of R^2 in the range of [0.1, 0.9] for different values of n and α_2 like in 2.

The y-axis in each subplot represents the same as in Figure 3 and the x-axis represents the same as in Figure 3 (Different than in Figures 1 and 2). These criteria are compared in terms of risk, with AIC and BIC showing lower risks compared to MMA in each of the subplots of Figure 4. AIC shows even a higher risk compared to BIC in the first three subplots of Figure 4 when $R^2 = 0.25$ and for small values of n. If R^2 (which looks like SNR) increases from $R^2 = 0.25$ to $R^2 = 0.75$, the trend is that the risks for AIC and BIC remain relatively similar but become slightly less volatile and converge to each other. MMA stays the same for different values of R^2, because the risk of MMA is normalized by itself. If the α_2 values increase from $\alpha_2 = 1$ to $\alpha_2 = 2$, the trend is that the risks of AIC and BIC increase a little bit for every possible n ($n =$ 50, 100, 500). However, MMA remains robust for different values of α_2. To conclude, AIC and BIC consistently outperforms MMA in terms of risk for different R^2 and for different parameter values (α_2's). Therefore same conclusion can be drawn as in Figure 3.

The results of Figure 4 do not match the published Figure 4 of the research of Peng & Yang

(2022) exactly and also do not show the same trends in the subplots. Possible reasons for this are the same reasons mentioned that apply to Figure 1 and Figure 3 as argued before.

5.2 Extension

5.2.1 Model selection methods based on AIC and BIC

To begin with, the results of the model selection methods based on AIC and BIC criteria are presented in Table 2. The AIC and BIC values are given in the table of all the candidate models, which are shown in the rows of the table.

Table 2: AIC and BIC values for Candidate Models to select the best model

Model	Predictors Used	AIC	BIC
1	Volume	1421.70	1428.31
2	Open	813.20	819.81
3	High	655.01	661.62
4	Low	610.60	617.20
5	Volume, Open	811.49	821.40
6	Volume, High	629.79	639.70
7	Volume, Low	610.05	619.96
8	Open, High	648.95	658.86
9	Open, Low	594.13	604.04
10	High, Low	579.36	589.27
11	Volume, Open, High	613.51	626.72
12	Volume, Open, Low	586.78	600.00
13	Volume, High, Low	579.48	592.69
14	Open, High, Low	**480.78**	**493.99**
15	Volume, Open, High, Low	481.14	497.66

The best model to select out of all the candidate models to predict the closing stock prices of Apple Inc. in Table 2 based on AIC is the model with all the predictors in it except the predictor 'Volume' ('Model 14'), so the model that contains 'Open', 'High' and 'Low' as predictors. This model has the lowest AIC of 480.78, as can be seen from Table 2 in bold. The AIC value of 'Model 14' is the lowest compared to the other models, because this model with all the predictors except the predictor 'Volume' in it is the most likely to provide the best fit to the training data. Therefore this model achieves the highest log-likelihood ($\log(L)$) compared to all the other candidate models with other predictors, which results in the lowest AIC value when looking at Equation 3 in Section 4.1.1. While 'Model 14' has many parameters (high value of k in Equation 3), the improvement of fit of the model (high value of $\log(L)$ in Equation 3) outweighs the penalty for having more parameters. This therefore results in the lowest AIC value of 'Model 14' compared to the other models. The simpler models with less predictors used have higher AIC values, because they do not fit the data as well as 'Model 14' resulting in a significantly lower log-likelihood ($\log(L)$). The reduction in the complexity penalty ($2k$ in Equation 3) is not enough to compensate for the loss in fit (lower $\log(L)$ in Equation 3), leading to higher overall AIC values for the simpler models. 'Model 15' has a slightly higher AIC value than 'Model 14', because it includes an additional predictor increasing complexity.

The best model to select out of all the candidate models to predict the closing stock prices of Apple Inc. in Table 2 based on BIC is the model with all the predictors in it except the predictor ‘Volume’ (‘Model 14’), so the model that contains ‘Open’, ‘High’ and ‘Low’ as predictors. This model has the lowest BIC of 493.99, as can be seen from Table 2 in bold. The BIC value of ‘Model 14’ is the lowest compared to the other models, because ‘Model 14’ fits the data well but not as perfectly as ‘Model 15’ (so lower $\log(L)$ for ‘Model 14’ in Equation 4 than $\log(L)$ for ‘Model 15’). On top of that, BIC has a heavier penalty on ‘Model 15’ than on ‘Model 14’ because of the number of parameters that is penalized heavier (k in Equation 4 is higher for ‘Model 15’ than for ‘Model 14’). ‘Model 14’ has one fewer predictor (‘Volume’) in comparison with ‘Model 15’, which strikes the balance by providing a good fit without the penalty for the parameter that affects ‘Model 15’. So, that is the reason why ‘Model 14’ has the lowest BIC value and the other models have higher BIC values.

As can be seen from Table 5 in Section A.2, the MSE and MAE are slightly lower for ‘Model 15’ (second best model according to AIC and BIC). This indicates better predictive accuracy for ‘Model 15’ instead of ‘Model 14’ (chosen as best according to AIC and BIC values) when looking at these evaluation criteria. This is against expectation, because ‘Model 14’ should be the best in terms of MSE, MAE and R^2 according to AIC and BIC values. However, the extra predictor of ‘Model 15’ slightly improves fit (lower MSE and MAE) despite AIC and BIC penalties. This indicates that ‘Model 15’ captures more data variability and nuances than ‘Model 14’. Both models (‘Model 14’ and ‘Model 15’) have very high R^2 values, implying that both models explain a very high proportion of the variance in the target variable. Therefore both models show excellent predictive performance with low MSE and MAE values and high R^2 values. However, ‘Model 15’ (not selected by AIC or BIC) performs slightly better according to the evaluation criteria compared to ‘Model 14’ (selected by AIC and BIC). This suggests that ‘Model 15’ may be the better choice for the predictive accuracy in this context.

To manage the model uncertainty in this financial application of Apple Inc. stock, model averaging can be applied. This method combines the strength of multiple models, maybe leading to even better predictive performance than in the situation where we choose one single best model (model selection). For instance, ‘Model 15’ can be weighted alongside other strong models (‘Model 14’) to improve the predictive performance of Apple Inc. stock historical quotes of one year.

5.2.2 Mallows Model Averaging (MMA)

Secondly, the results of the model averaging method based on minimizing Mallows criterion are shown in Table 3. The results in Table 3 indicate that ‘Model 14’ has the highest weight (0.9637). This suggests that this model is the best balance between fit and complexity after having performed MMA as method (same conclusion as for AIC and BIC values across different candidate models but now not weight 1). The combined information from the predictors ‘Open’, ‘High’ and ‘Low’ provides an extensive overview of market dynamics, capturing both price movements and volatility effectively and is therefore most valuable for predicting the target variable ‘Close/Last’. Following that, ‘Model 11’ has a notable weight (0.0185) because adding ‘Volume’ to ‘Open’ and ‘High’ helps capturing trading activity, which might be crucial in predicting the target variable. Moreover, ‘Model 9’ has a notable weight as well (0.0177) because it provides valuable and complementary information that improves the predictive accuracy when combined with the other models. ‘Model 1’ with a very small weight of 0.000115 indicates that ‘Volume’ has some predictive power but is relatively weak compared to other predictors. So, these models are included to some extent but not as influential (reason why the weights are very small). The reason why a lot of models have weight 0 or close to zero is due to the fact that they do nothing in terms of adding information since it shows no improvement in overall fit or they provide redundant info (reason why left out as models in Table 3). This redundancy occurs when different models include highly correlated predictors (between ‘Volume’ and size of price changes as can be seen in Table 7 in Section A.4).

Table 3: Model Weights after performing Mallows Model Averaging (MMA)

Model	**Predictors Used**	**Model Weights of MMA**
1	Volume	0.000115
9	Open, Low	0.017722
11	Volume, Open, High	0.018475
14	Open, High, Low	0.963687

To summarize, models with better predictive performance on the training data are favored (that is why they have higher weights), simpler models that achieve similar predictive performance to more complex models are preferred to avoid over fitting and models with redundant information (models that include predictors that are already well presented in higher weighted models) are given quite low weights.

In Table 6 in Section A.2, the MSE, MAE and R^2 values can be seen for each model individually on the 20% test set after having performed MMA. ‘Model 15’ has the lowest MSE and MAE of all candidate models. On top of that, ‘Model 15’ has the highest R^2 value of all candidate models. ‘Model 14’ has the second lowest MSE and MAE and second highest R^2 value across all the different candidate models. The risk in terms of MSE for the method MMA on the 20 % test set of the data with the given weights from Table 3 is 0.5706. This MSE is smaller than the MSE of ‘Model 14’ (chosen as best according to AIC and BIC) and bigger than the MSE of ‘Model 15’ (chosen as best according to MSE and MAE). So, MMA with the given weights per model in Table 3 shows a small improvement in MSE compared to choosing only ‘Model 14’, but not as good as the MSE of ‘Model 15’ seperately. Therefore MMA does improve the predictive accuracy in terms of MSE when choosing ‘Model 14’ only as benchmark (according to AIC and BIC) So, MMA lowers the MSE risk compared to the MSE risk of choosing only ‘Model 14’ based on AIC and BIC (0.5706 is smaller than 0.5737). However, MMA does not improve the predictive accuracy in terms of MSE when choosing ‘Model 15’ only as benchmark. So, MMA does not lower the MSE risk compared to the MSE risk of choosing only ‘Model 15’ (0.5706 is bigger than 0.5688).

5.2.3 Frequentist Model Averaging

Thirdly and lastly, the results of the frequentist model averaging method are shown in Table 4. The findings in Table 4 indicate that ‘Model 14’ has the highest weight (0.544879) and ‘Model

15' has the second highest weight (0.455121). 'Model 14' has the highest weight after having performed FMA, because this model has the lowest AIC value (480.78 in Table 2) as this model is the best balance between fit and simplicity (explained in 5.2.1). Therefore in Equation 11 I_k (actually AIC) is the lowest across all the other candidate models. This results in the highest weight for 'Model 14' according to Equation 12. This model is thus the most preferred by a significant margin because of the relatively high weight of 0.544879. 'Model 15' has the second highest weight after having performed FMA, because this model has the second lowest AIC value (481.14 in Table 2). Therefore in Equation 11 I_k (actually AIC) is the second lowest across all the other candidate models. This results in the second highest weight for 'Model 15' according to Equation 12. This model is a good model (weight of 0.455121), but significantly less preferred than 'Model 14' (weight of 0.544879). All the remaining models have very high AIC values (See Table 2) relative to 'Model 14' and 'Model 15', leading to extremely low weights (effectively zero according to Equation 12). These models are less preferred according to the best models (and left out in Table 4).

Table 4: Model Weights after performing Frequentist Model Averaging (FMA)

Model	**Predictors Used**	**Model Weights of FMA**
14	Open, High, Low	0.544879
15	Volume, Open, High, Low	0.455121

The MSE, MAE and R^2 values for each model individually on the 20% test set after having executed FMA are the same as in Table 6 in Section A.2 after having performed MMA. The risk in terms of MSE for the method FMA on the 20% test set of the data with the given weights from Table 4 is 0.5723. This MSE is a little bit smaller than the MSE of 'Model 14' (chosen as best according to AIC and BIC) and bigger than the MSE of 'Model 15' (chosen as best according to MSE and MAE). So, FMA with the given weights per model in Table 4 shows a small improvement in MSE compared to choosing only 'Model 14' (0.5723 is smaller than 0.5737), but not as good as the MSE of 'Model 15' (0.5723 is bigger than 0.5688). So, FMA lowers the MSE risk compared to the MSE risk of choosing only 'Model 14' based on AIC and BIC. However, FMA does not lower the MSE risk compared to the MSE risk of choosing only 'Model 15'. This is the same reasoning and conclusion that can be drawn after having performed MMA as method. The MSE risk of FMA of 0.5723 is bigger than the MSE risk of MMA of 0.5706. So, MMA is preferred over FMA as method when looking at MSE risk in this given empirical application.

6 Conclusion

In the first place, this paper has worked on the problem regarding model selection methods (based on AIC and BIC) and model averaging methods (MMA and FMA) to choose the best model or average with weights across more models in a financial application like Apple Inc. common stock historical quotes of one year. After that, these methods are evaluated based on some appropriate evaluation criteria (MSE and MAE, MSE mainly) to compare the methods (MS and MA) with each other (MA improve predictive accuracy or not).

Following that, the first important result of this research is that according to the model selection method based on AIC the model with the predictors ‘Open’, ‘High’ and ‘Low’ in it (‘Model 14’ in Table 2) is the best model to select. Besides that, another important result using the BIC criterion for model selection, the model with the predictors ‘Open’, ‘High’ and ‘Low’ in it (‘Model 14’ in Table 2) is the best model to select based on BIC values. The MSE and MAE values of ‘Model 15’ are the lowest compared to ‘Model 14’ and all the other candidate models when looking at Table 5. So, ‘Model 15’ is the best choice for predictive accuracy based on MSE and MAE values and ‘Model 14’ is the best according to AIC and BIC model selection methods.

According to the MMA method, ‘Model 14, 11, 9 and 1’ are weighted with each other to account for model uncertainty in Table 3. The important result of this method is that the MSE risk of MMA does not improve MSE risk of choosing ‘Model 15’ only, but does improve MSE risk of choosing ‘Model 14’ only. Following the FMA method, ‘Model 14 and 15’ are weighted with each other to account for model uncertainty in Table 4. The most relevant outcome of this method is that the MSE risk of FMA does not improve the MSE risk of MMA and choosing ‘Model 15’ only, but does improve the MSE risk of choosing ‘Model 14’ only.

The study shows that model averaging, whether MMA or FMA, reduces model selection uncertainty and leads to better predictive accuracy when compared to individual models like ‘Model 14’ but not necessarily with respect to the best model chosen according to MSE and MAE (‘Model 15’). This has a very important practical ramification: whereas comprehensive models, like ‘Model 15’, are necessary in financial applications for accuracy, model averaging provides insulation against the selection of poorer models.

Some interesting question to be investigated for future work is do this research using other stocks or financial instruments in order to validate its generalizability across different cases. These vary greatly depending on time period covered including daily, weekly, monthly or more than one year data periods. Shorter durations can be considered to capture high-frequency trading dynamics and longer periods can be taken into account to figure out long-term trends and cycles. More research needs to be done where different model averaging techniques as well as more advanced machine learning algorithms can be checked against it. The predictive power of the models could be improved by including macroeconomic factors and exogenous variables, such as interest rates and GDP growth.

To summarize and to conclude, in this paper we have shown that MMA and FMA improve predictive accuracy of ‘Model 14’ (chosen based on AIC and BIC) but MMA and FMA do not improve predictive accuracy of ‘Model 15’ (lowest MSE and MAE) in a financial application such as closing stock prices from Apple Inc. of one year. The limitation of this research and work is that it is one empirical application of one stock during one year with only four predictors where MSE risk is used as evaluation criteria.

A Appendix

A.1 Programming code

Will be available in Python documents in the zip file that is handed in on the 1st of July 2024 via sin online in Thesis Hub.

The first Python document is called datasummarystats. When running this Python document, the output in Table 1 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The second Python document is called risk_simulation_figure1_aicbicrisks. When running this Python document, the output in Figure 1 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The third Python document is called risk_simulation_figure2_aicbicrisks. When running this Python document, the output in Figure 2 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The fourth Python document is called risk_simulation_figure3_aicbicrisks. When running this Python document, the output in Figure 3 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The fifth Python document is called risk_simulation_figure4_aicbicrisks. When running this Python document, the output in Figure 4 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The sixth Python document is called MSResultsAICBIC. When running this Python document, the output in Table 2 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The seventh Python document is called MMAFINANCIALDATAREVISEDMSEPERMODEL. When running this Python document, the output in Table 5, Table 6 and the risk MSE for MMA on the test set can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The eight Python document is called mmafinancialdatarevisedversion. When running this Python document, the output in Table 3 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The ninth Python document is called FMAResultsFinancialData. When running this Python document, the output in Table 4 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The tenth Python document is called FMAEvaluationMSEPerModel. When running this Python document, the output in Table 6 can be obtained and the MSE, MAE and R^2 can be obtained of the FMA model performance. In the document, the code is well-documented enough to provide a short explanation of the code.

The eleventh Python document is called CorrelationVolume. When running this Python document, the output in Table 7 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

The twelfth Python document is called RelationshipOpenCloseLastPrices. When running this Python document, the output in Table 8 can be obtained. In the document, the code is well-documented enough to provide a short explanation of the code.

A.2 Evaluation criteria tables, equations and data split

To start with, MSE has the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{13}$$

where n is the number of observations, y_i is the actual value of the model and $\hat{y_i}$ is the predicted value that is estimated based on the model. The lower the value of the MSE, the better the performance of the model. The higher the value of the MSE, the worser the performance of the model.

Secondly, MAE has the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{14}$$

where n is the number of observations, y_i is the actual value of the model and $\hat{y_i}$ is the predicted value that is estimated based on the model. The lower the value of the MAE, the better the performance of the model. The higher the value of the MAE, the worse the performance of the model.

Lastly, R^2 which indicates the goodness of fit of a model has the following expression:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{15}$$

where n is the number of observations, y_i is the actual value of the model and $\hat{y_i}$ is the predicted value that is estimated based on the model, $\bar{y}$ is the mean of the observed data. R^2 is the proportion of variance in the dependent variable (Close/Last in this paper) that is predictable from the independent variables (Volume, High, Low and Open in this paper). The closer the R^2 value to 1, the better the fit is of the model.

The data is preprocessed and split at first. Then the dates are ordered before splitting the dataset. After that, models are created based on the predictors mentioned (for ‘Model 14’: Open, High and Low and for ‘Model 15’: Volume, Open, High and Low). At this point, we have training sets (80 %) and test sets (20 %) of the data. Based on this data split, the evaluation criteria are calculated for each model. So, we can fit linear regression models now for ‘Model 14’ and ‘Model 15’ using their respective predictors and also for the other candidate models with their predictors. The evaluation criteria of how the models (especially ‘Model 14’, which is the best according to AIC and BIC values and ‘Model 15’ which is the second best according to AIC and BIC values) performed are shown in Table 5 in this Section.

Table 5: Evaluation Criteria after applying Model Selection Methods based on AIC and BIC

Evaluation criteria	Model 14	Model 15
MSE	0.5737	**0.5688**
MAE	0.5944	**0.5929**
R^2	0.9930	**0.9931**

Table 6: MSE, MAE and R^2 values of each model individually

Model	**Predictors Used**	**MSE**	**MAE**	$\mathbf{R}^2$
1	Volume	86.921326	7.614943	-0.057678
2	Open	2.858476	1.279722	0.965217
3	High	0.981008	0.765765	0.988063
4	Low	1.333712	0.866787	0.983771
5	Volume, Open	3.011932	1.300197	0.963350
6	Volume, High	0.854371	0.682570	0.989604
7	Volume, Low	1.277140	0.846875	0.984459
8	Open, High	1.012111	0.776344	0.987684
9	Open, Low	1.261749	0.838418	0.984647
10	High, Low	0.941784	0.747966	0.988540
11	Volume, Open, High	0.782258	0.633201	0.990481
12	Volume, Open, Low	1.153164	0.797645	0.985968
13	Volume, High, Low	0.934432	0.745020	0.988630
14	Open, High, Low	0.573685	0.594359	0.993019
15	Volume, Open, High, Low	**0.568807**	**0.592910**	**0.993079**

A.3 Description symbols equation(s)

The description of the symbols in Equation 3 in Section 4.1.1 is as follows: where k in 3 is the number of parameters in the model and L the log likelihood of the model.

The description of the symbols in Equation 4 in Section 4.1.1 is as follows: where k in 4 is the number of parameters in the model, n the sample size and L the log likelihood of the model.

A.4 Tables of how covariates relate to each other

Table 7: Correlation between Volume and Price Differences

	Volume and Open-Close Difference	Volume and High-Low Difference
Correlation	0.08	0.52

Table 8: Trend of Apple Inc. Stock Prices (Open vs Close)

Trend	Count
Increase	131
Decrease	121
Stable	0

References

Adrian E. Raftery, D. M. & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179-191.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Institute of Statistical Mathematics*, 267-281.

Banerjee, T. (2020). Forecasting apple inc. stock prices using sp500– an ols regression approach with structural break. In *2020 ieee 1st international conference for convergence in engineering (icce)* (p. 306-310). doi: 10.1109/ICCE50343.2020.9290495

Buckland, B. & Augustin. (1997). Model selection: An integral part of inference. *Biometrics*, *53*(2), 603-618.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *158*(3), 419-466.

Claeskens, N. L. H. . G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*(464), 879-899.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 45-70.

F.E. Turheimer, R. H. & Cunningham, V. (2003). On the undecidability among kinetic models: From model selection to model averaging. *Journal of Cerbral Blood Flow Metabolism*, *23*, 490-498.

Goodness Aye, S. H. W. J. K., Rangan Gupta. (2015). Forecasting the price of gold using dynamic model averaging. *International Review of Financial Analysis*, *41*, 257-266.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175-1189.

Hartman, B. M. & Groendyke, C. (2013). Model selection and averaging in finacial risk management. *North American Actuarial Journal*, *17*(3), 216-228.

H. Wang, X. Z. & Zou, G. (2009). Frequentist model averaging estimation: a review. *Journal of Systems Science and Complexity*, *22*, 732-748.

Johnson, J. B. & Omland, K. S. (2004). Model selection in ecology and evolution. *TRENDS in Ecology and Evolution*, *19*(2), 101-108.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys*, *29*(1), 46-75.

Nasdaq. (2024). *Apple inc. (aapl) historical quotes.* `https://www.nasdaq.com/market-activity/stocks/aapl/historical?page=1&rows_per_page=10&timeline=y1`. (Accessed: 2024-05-30)

Peng, J. & Yang, Y. (2022). On improvability of model selection by model averaging. *Journal of Econometrics*, *229*(2), 246-262.

Rai, B., Kasturi, M. & Huang, C.-y. (2018). Analyzing stock market movements using news, tweets, stock prices and transactions volume data for apple (aapl), google (goog) and sony (sne). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3243250.3243263` doi: 10.1145/3243250.3243263

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Wagenmakers, E.-J. & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic Bulletin and Review*, *11*, 192-196.

Wan, A. T. & Zhang, X. (2009). On the use of model averaging in tourism research. *Annals of Tourism Research*, *36*(3), 525-532.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92-107.

Yuan, Z. & Yang, Y. (2005). Combining linear regression models. *Journal of the American Statistical Association*, *100*(472), 1202-1214.