CRITICAL STUDY OF AIC MODEL SELECTION TECHNIQUES.

BY

MING YANG JEREMY TAN

DISSERTATION
Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physics in the Graduate College of the University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:
Associate Professor Karin A Dahmen, Chair
Professor Yoshitsugu Oono, Director of Research
Professor Jon J Thaler
Professor James E Wiss
Abstract

Akaike Information Criterion (AIC) has been used widely as a statistical criterion to compare the appropriateness of different parametric models underlying a particular dataset. Under suitable conditions, AIC is an unbiased estimator of the Kullback-Leibler divergence \( D(T \parallel A) \) of a candidate model \( A \) with respect to the truth \( T \), where we have defined \( T \) as an underlying process consisting of a signal with stochastic noise; a particular empirical data is a single realization of this process. Thus, a model with a smaller AIC is ranked as a better model, since it has a smaller Kullback-Leibler discrepancy with \( T \). However, it is an important question whether the difference between the AIC values for two candidate models is statistically significant. This was partially addressed in terms of a probability of models by using Akaike weights. It is also important to remember that the AIC itself is statistically estimated with the aid of available data. We explored the impact of the possible errors of the estimated AIC in the context of comparing models underlying the late time acceleration of the universe, using SNIa, using a parametric bootstrap technique to study the reliability of the estimated AIC. From the specific example that we study, we find that the estimator uncertainty in the AIC differences can be significant. Therefore, AIC-based studies should not only pay attention to Akaike weight based criterion of reliability, but should also consider the impact of estimator uncertainty of AIC. We also examined the AIC model selection strategy and compared it to other model selection techniques. In our comparison, we showed that the AIC-based \( 2k \) cost term does not properly account for model complexity, in contrast to other model selection techniques. Besides complexity, an alternative model selection criterion based on model stability was proposed and studied.
‘It is better to be vaguely right than precisely wrong.’

- Carveth Read
Acknowledgments

First and foremost, I would like to thank Yoshi for being extremely supportive of my endeavors. Being in the group allowed me the freedom to pursue and explore my own agenda in information theory. Without Yoshi, I would not have been introduced to AIC or other model selection schemes that led to this thesis. Secondly, I would like to thank Rahul for his input and help on observational cosmology. He was also extremely helpful in providing Figures 7.1, 7.2, 7.3 and 7.4 for this dissertation. Without his help, a significant part of this thesis would not be possible. I would next like to thank Satwik. The introduction in his preliminary exam gave me the idea of using model stability as a modification to existing model selection schemes. I would also like to thank Adam, Aruna, Mehmet and all other members of the group for many helpful discussions. I am also grateful to the department for providing financial support for my graduate education and creating a conducive atmosphere for research. I also appreciate the many close friendships (Michael, Onyeama, George, Patricio, Ayo, Benjamin, both Nicks, Stan, Zhun Yong, Eun Yeong and many others) I have made throughout the 6+ years I spent in Champaign. Finally, I would like to thank my parents for supporting and understanding my decision to pursue my dreams.
# Table of Contents

List of Tables ........................................................................................................ vii
List of Figures .......................................................................................................... ix
List of Abbreviations ................................................................................................. xi
List of Symbols .......................................................................................................... xii

Chapter 1 Introduction .............................................................................................. 1
  1.1 The issue of model selection ............................................................................. 1
  1.1.1 Over-fitting .................................................................................................. 1
  1.1.2 Model selection techniques ......................................................................... 9
  1.2 Akaike information criterion .......................................................................... 10
  1.3 Observational cosmology ............................................................................... 10
  1.4 Different model selection criteria ................................................................... 11
  1.5 Roadmap .......................................................................................................... 11

Chapter 2 Kullback-Leibler Divergence ................................................................. 12
  2.1 Introduction ...................................................................................................... 12
  2.2 Shannon entropy of a probability distribution .............................................. 12
  2.3 Modeling a process with a ‘wrong’ distribution ............................................ 13
  2.4 Mismatch between two probability distributions ........................................... 13
  2.5 General form of KL divergence ..................................................................... 14
  2.6 The role of KL divergence in large deviation theory ...................................... 14
  2.7 Geometric interpretation of KL divergence for small perturbation ............. 17

Chapter 3 Model Class Comparison ...................................................................... 18
  3.1 Introduction ...................................................................................................... 18
  3.2 Regression models .......................................................................................... 18
  3.3 Derivation of AIC from the KL divergence ...................................................... 19

Chapter 4 Confidence Levels from the AIC Differences ........................................ 23
  4.1 Introduction ...................................................................................................... 23
  4.2 The interpretation of AIC differences ............................................................ 23
  4.3 Reinterpretation of the AIC differences ......................................................... 24
  4.4 Akaike weights ............................................................................................... 26
  4.5 Akaike threshold ............................................................................................ 26

Chapter 5 Cosmological Model Selection ............................................................ 28
  5.1 Introduction ...................................................................................................... 28
  5.2 Cosmological models ..................................................................................... 29
  5.3 Type Ia SN data ............................................................................................. 30
  5.3.1 Background .............................................................................................. 30
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3.2 Obtaining observational data</td>
<td>32</td>
</tr>
<tr>
<td>5.3.3 Relating data to the model</td>
<td>32</td>
</tr>
<tr>
<td>5.4 Getting a $\chi^2_{ML}$ fit from SNIa</td>
<td>33</td>
</tr>
<tr>
<td>5.5 Deriving the AIC results</td>
<td>34</td>
</tr>
<tr>
<td>Chapter 6 Finite Datasets</td>
<td>36</td>
</tr>
<tr>
<td>6.1 Impact of AIC uncertainties in finite datasets</td>
<td>36</td>
</tr>
<tr>
<td>6.2 Bootstrap</td>
<td>36</td>
</tr>
<tr>
<td>6.2.1 Non-parametric bootstrap</td>
<td>37</td>
</tr>
<tr>
<td>6.2.2 Parametric bootstrap</td>
<td>37</td>
</tr>
<tr>
<td>6.3 Simulating data</td>
<td>38</td>
</tr>
<tr>
<td>Chapter 7 Test for Reliability</td>
<td>41</td>
</tr>
<tr>
<td>7.1 Introduction</td>
<td>41</td>
</tr>
<tr>
<td>7.2 Statistical self-consistency</td>
<td>41</td>
</tr>
<tr>
<td>7.3 Numerical study of the threshold</td>
<td>43</td>
</tr>
<tr>
<td>7.4 Error bars</td>
<td>45</td>
</tr>
<tr>
<td>Chapter 8 Further Criticism of AIC</td>
<td>52</td>
</tr>
<tr>
<td>8.1 Introduction</td>
<td>52</td>
</tr>
<tr>
<td>8.2 $2k$ term as an inadequate measure of complexity</td>
<td>52</td>
</tr>
<tr>
<td>8.3 Issues with the analytic derivation of the AIC cost term</td>
<td>53</td>
</tr>
<tr>
<td>8.4 Comparison of AIC to other model selection criterion</td>
<td>54</td>
</tr>
<tr>
<td>8.4.1 Bayesian Information Criterion</td>
<td>54</td>
</tr>
<tr>
<td>8.4.2 Minimum Description Length</td>
<td>54</td>
</tr>
<tr>
<td>8.4.3 Takeuchi Information Criterion</td>
<td>55</td>
</tr>
<tr>
<td>8.4.4 Bootstrap estimate of AIC-like criteria</td>
<td>56</td>
</tr>
<tr>
<td>Chapter 9 Stability as an Alternative Consideration to Complexity</td>
<td>58</td>
</tr>
<tr>
<td>9.1 Non-bootstrap perturbation</td>
<td>59</td>
</tr>
<tr>
<td>9.2 Quantifying model instability</td>
<td>60</td>
</tr>
<tr>
<td>9.2.1 Using KL divergence</td>
<td>60</td>
</tr>
<tr>
<td>9.2.2 Using an $l_p$-inspired Distance Measure</td>
<td>62</td>
</tr>
<tr>
<td>9.3 Demonstration of a stability information criterion (SIC)</td>
<td>62</td>
</tr>
<tr>
<td>9.3.1 SIC based on KL divergences</td>
<td>62</td>
</tr>
<tr>
<td>9.3.2 SIC based on $l_2$ distance measure</td>
<td>63</td>
</tr>
<tr>
<td>9.4 Summary</td>
<td>64</td>
</tr>
<tr>
<td>Chapter 10 Conclusions</td>
<td>65</td>
</tr>
<tr>
<td>10.1 Alternative interpretation of AIC difference</td>
<td>65</td>
</tr>
<tr>
<td>10.2 Numerical studies of AIC in the context of cosmological model selection</td>
<td>65</td>
</tr>
<tr>
<td>10.3 Stability as an alternative to complexity</td>
<td>67</td>
</tr>
<tr>
<td>Appendix A Bayesian Interpretation of Akaike Differences</td>
<td>69</td>
</tr>
<tr>
<td>Appendix B Confusion Probability</td>
<td>71</td>
</tr>
<tr>
<td>Appendix C Minimization Procedures</td>
<td>74</td>
</tr>
<tr>
<td>C.1 Gauss-Newton algorithm</td>
<td>74</td>
</tr>
<tr>
<td>C.2 Correlated sum of squares</td>
<td>75</td>
</tr>
<tr>
<td>C.3 Hybrid grid-gradient search</td>
<td>75</td>
</tr>
<tr>
<td>C.4 Integration</td>
<td>76</td>
</tr>
<tr>
<td>C.5 Implementation</td>
<td>76</td>
</tr>
<tr>
<td>References</td>
<td>77</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Both datasets were computed from the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d.. Due to the different realization of the noise for \( \epsilon_i \), the datasets are different. The respective curves were calculated by taking the MLE fit against the original dataset (middle column). We compute the residual sum of squares (RSS) of these curves against the two dataset using the formula: \( \text{RSS} = \sum_i (y_i - f(x_i))^2 \), where \( f \) is the respective polynomial curve. The table shows that while the 20th order polynomial provided the best fit against the original dataset, it did not do so for the new dataset. 5

1.2 Both datasets were computed from the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 3\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d.. Due to the different realization of the noise for \( \epsilon_i \), the datasets are different. The respective curves were calculated by taking the MLE fit against the original dataset (middle column). We compute the residual sum of squares (RSS) of these curves against the two dataset using the formula: \( \text{RSS} = \sum_i (y_i - f(x_i))^2 \), where \( f \) is the respective polynomial curve. The table shows that while the 20th order polynomial provided the best fit against the original dataset, it did not do so for the new dataset. 5

1.3 Both datasets were computed from the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 5\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d.. Due to the different realization of the noise for \( \epsilon_i \), the datasets are different. The respective curves were calculated by taking the MLE fit against the original dataset (middle column). We compute the residual sum of squares (RSS) of these curves against the two dataset using the formula: \( \text{RSS} = \sum_i (y_i - f(x_i))^2 \), where \( f \) is the respective polynomial curve. The table shows that while the 20th order polynomial provided the best fit against the original dataset, it did not do so for the new dataset. 5

1.4 Both datasets were computed from the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 10\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d.. Due to the different realization of the noise for \( \epsilon_i \), the datasets are different. The respective curves were calculated by taking the MLE fit against the original dataset (middle column). We compute the residual sum of squares (RSS) of these curves against the two dataset using the formula: \( \text{RSS} = \sum_i (y_i - f(x_i))^2 \), where \( f \) is the respective polynomial curve. The table shows that while the 20th order polynomial provided the best fit against the original dataset, it did not do so for the new dataset. 5

5.1 Different model classes considered in this thesis: We show the evolution of the Hubble function \( H(z) \) with redshift \( z \), the Hubble constant \( H_0 \) and other free parameters in the models. \( k \) is the number of free parameters. Note that these models assume a flat universe. 29

5.2 Different flat model classes considered in this thesis: We indicate the free parameters obtained from the Constitution compilation [40] using the maximum likelihood estimation. \( k \) is the number of free parameters. The respective AIC values were also obtained from the Constitution compilation. 35
7.1 Percentage of cases where the AIC method with a threshold of 0 selects the correct model over other candidate models considered. We defined the difference and percentage using the following convention. If we define the bootstrap data as being produced by the reference model $X$ (rows) and we are comparing it against model $Y$ (columns), the difference is defined as the AIC value of $X$ minus the AIC value of $Y$ and is denoted by the symbol $\Delta_{X,Y}^d$. The table counts the percentage of negative $\Delta_{X,Y}^d, d \in X_n$. Note that a value greater than 50% indicates that the correct model is chosen a majority of the time. We use this cut-off of 50% or more as a simple definition for statistical self-consistency.

7.2 Failure Rate for $\Delta_{\text{threshold}} = 2$: The $(\text{f}_{\text{ind}}, \text{f}_{\text{all}}^{\text{false}}, \text{f}_{\text{det}}^{\text{false}})$ values in the parentheses show $\text{f}_{\text{ind}}$, the percentage of total cases where the AIC procedure has an indeterminate result, $\text{f}_{\text{all}}^{\text{false}}$, the fraction of total cases where the AIC procedure results in an incorrect model selection; and $\text{f}_{\text{det}}^{\text{false}}$, the percentage of determinate cases where the AIC wrongly selects a candidate model $Y$ over the reference model $X$.

7.3 Failure Rate for $\Delta_{\text{threshold}} = 5$: The $(\text{f}_{\text{ind}}, \text{f}_{\text{all}}^{\text{false}}, \text{f}_{\text{det}}^{\text{false}})$ values in the parentheses show $\text{f}_{\text{ind}}$, the percentage of total cases where the AIC procedure has an indeterminate result, $\text{f}_{\text{all}}^{\text{false}}$, the fraction of total cases where the AIC procedure results in an incorrect model selection; and $\text{f}_{\text{det}}^{\text{false}}$, the percentage of determinate cases where the AIC wrongly selects a candidate model $Y$ over the reference model $X$. '-' indicates an indeterminate result.
List of Figures

1.1 Top Left: A raw dataset was generated from the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0,1) \) and i.i.d.. Top Right: Linear curve fitted to the data (Under-fitting). Middle Left: Quadratic curve fitted to the data (Under-fitting). Middle Right: Cubic curve fitted to the data (Optimal fit). Bottom Left: Quartic curve fitted to the data (Slight over-fitting). Bottom Right: 20th order polynomial fitted to the data (over-fitting).

1.2 Top Left: Another raw dataset that was generated from the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0,1) \) and i.i.d.. The subsequent curves used were taken from Fig. 1.1. They were generated by taking the MLE fit against the dataset used in Fig. 1.1. The purpose is to see whether curves generated in one dataset would provide a good fit for another similar dataset. Top Right: Linear curve superposed over the data. Middle Left: Quadratic curve superposed over the data. Middle Right: Cubic curve superposed over the data. Bottom Left: Quartic Curve superposed over the data. Bottom Right: 20th order polynomial superposed over the data.

1.3 Comparison of two linear curves that were fitted to two different datasets. Each dataset was generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0,1) \) and i.i.d.. Figure is taken from [32] and reproduced with the permission of Prof. Alexei Filippenko (author).

1.4 Comparison of two quadratic curves that were fitted to two different datasets. Each dataset was generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0,1) \) and i.i.d.. Figure is taken from [32] and reproduced with the permission of Prof. Alexei Filippenko (author).

1.5 Comparison of two cubic curves that were fitted to two different datasets. Each dataset was generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0,1) \) and i.i.d.. Figure is taken from [32] and reproduced with the permission of Prof. Alexei Filippenko (author).

1.6 Comparison of two quartic curves that were fitted to two different datasets. Each dataset was generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0,1) \) and i.i.d.. Figure is taken from [32] and reproduced with the permission of Prof. Alexei Filippenko (author).

1.7 Comparison of two 20th order polynomial curves that were fitted to two different datasets. Each dataset was generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0,1) \) and i.i.d.. Figure is taken from [32] and reproduced with the permission of Prof. Alexei Filippenko (author).

5.1 Spectra of the major types of supernovae. Line (a) represents a Type Ia event. Line (b) represents a Type II event. Line (c) represents a Type Ic event. Line (d) represents a Type Ib event. Figure is taken from [32] and reproduced with the permission of Prof. Alexei Filippenko (author).

5.2 Type Ia SN Data. 371 data points were used from the Constitution Compilation [40]. The vertical axis indicates the distance moduli while the horizontal axis indicates the redshift. Vertical error bars are given for the distance moduli. Horizontal error bars are negligible and not indicated.
7.1 Probabilities of different candidate models (Cand) being selected over different reference models (Ref) by using AIC for different values of $\Delta_{\text{threshold}}$ for the case of (a) (top left) the candidate model DGP being picked over the reference model ACDM, (b) (top right) the candidate model ACDM being picked over the reference model wCDM, and (c) (bottom) the candidate model ACDM being picked over the reference model CPL. The solid blue curve shows the number of incorrect results as a fraction $f_{\text{false}}$ of the cases where the procedure returns a determinate result. The black dashed curve shows the number of incorrect results as a fraction of the total number of simulations. The fraction $f_{\text{false}}$ is extremely noisy and should be ignored when the fraction of indeterminate cases $f_{\text{ind}}$ (red, dotted) is large. The plots show that increasing $\Delta_{\text{threshold}}$ always decreases $f_{\text{false}}$. $f_{\text{false}}$ does not necessarily decrease. For a comparison of the models used, this shows that AIC tends to incorrectly select models with a lower number of parameters. In the comparison of nested models, $f_{\text{false}}$ drops sharply at twice the difference in the number of free parameters of the models compared. 

7.2 Probability of the candidate model (Cand) CPL being selected over the reference model (Ref) ACDM for different values of the threshold. In this case, $f_{\text{false}}$ actually increases sharply at about twice the difference in free parameters between these models, showing that in this case AIC tends to incorrectly select the model with a higher number of parameters.

7.3 Probability distributions of the AIC differences between ACDM and DGP ($\Delta_{\text{ACDM,DGP}}^d$) for different reference model X. X is used to generate the respective bootstrap samples and is written in the right upper corner of the figures. The horizontal axis indicates the $\Delta_{\text{ACDM,DGP}}^d$ value while the vertical axis indicates their relative frequency. If the process underlying our observations was really the best fit model of class X, then the values of the AIC differences under different realizations of noise would have the histogram distribution $\{\Delta_{\text{ACDM,DGP}}^d | d \in C_371\}$ shown. Vertical lines show the respective AIC differences that were derived from the observed data.

7.4 Probability distributions of the AIC differences between ACDM and CPL ($\Delta_{\text{ACDM,CPL}}^d$) for different reference model X. X is used to generate the respective bootstrap samples and is written in the right upper corner of the figures. The horizontal axis indicates the $\Delta_{\text{ACDM,CPL}}^d$ value while the vertical axis indicates their relative frequency. If the process underlying our observations was really the best fit model of class X, then the values of the AIC differences under different realizations of noise would have the histogram distribution $\{\Delta_{\text{ACDM,CPL}}^d | d \in C_371\}$ shown. The exponential-like distributions observed are due to the fact that ACDM and CPL are nested models. Vertical lines show the respective AIC differences that were derived from the observed data.

9.1 $\gamma_1$ vs. model order. 50 data points were generated from the function $x^2 + 2x + 3 + w$, where $w \sim N(0, 1)$, at intervals of 0.1 from $x = -3.5$ to $x = 1.4$. We fitted several different order polynomial models to the data. For example, an order 2 model is a quadratic curve. We calculated the corresponding instability term $\gamma_1$ from (9.4) and plot the relationship.
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion.</td>
</tr>
<tr>
<td>BAO</td>
<td>Baryon Acoustic Oscillations.</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion.</td>
</tr>
<tr>
<td>CMB</td>
<td>Cosmic Microwave Background.</td>
</tr>
<tr>
<td>CPL</td>
<td>Chevaliar-Polarski-Linder.</td>
</tr>
<tr>
<td>DGP</td>
<td>Dvali-Gabadadze-Poratti.</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>Independent and Identically Distributed Random Variables.</td>
</tr>
<tr>
<td>KL</td>
<td>Kullback-Leibler.</td>
</tr>
<tr>
<td>ACDM</td>
<td>Lambda Cold Dark Matter.</td>
</tr>
<tr>
<td>LDT</td>
<td>Large Deviation Theory.</td>
</tr>
<tr>
<td>MLCS</td>
<td>Multicolor Light Curve Shapes.</td>
</tr>
<tr>
<td>MDL</td>
<td>Minimum Description Length.</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimation.</td>
</tr>
<tr>
<td>OM</td>
<td>Original Model.</td>
</tr>
<tr>
<td>PM</td>
<td>Perturbed Model.</td>
</tr>
<tr>
<td>RSS</td>
<td>Residual Sum of Squares.</td>
</tr>
<tr>
<td>SIC</td>
<td>Stability Information Criterion.</td>
</tr>
<tr>
<td>SN</td>
<td>Supernova.</td>
</tr>
<tr>
<td>SNIa</td>
<td>Type Ia Supernova.</td>
</tr>
<tr>
<td>TIC</td>
<td>Takeuchi Information Criterion.</td>
</tr>
<tr>
<td>wCDM</td>
<td>w Cold Dark Matter.</td>
</tr>
<tr>
<td>WMAP</td>
<td>Wilkinson Microwave Anisotropy Probe.</td>
</tr>
</tbody>
</table>
List of Symbols

\( m \)  Apparent Magnitude.
\( M \)  Absolute Magnitude.
\( \chi^2 \)  Chi-Square (Fit).
\( \mu \)  Distance Moduli.
\( H_0 \)  Hubble Parameter.
\( d_L \)  Luminosity Distance.
\( z \)  Red-shift.
\( \Omega_k \)  Term describing the curvature of the universe.
\( c \)  Speed of light.
Chapter 1

Introduction

1.1 The issue of model selection

1.1.1 Over-fitting

Suppose we wish to choose the best\(^1\) model from a set of theoretical models (or theories) of a natural phenomenon with the aid of the relevant empirical data. How can we objectively accomplish this goal? One solution is to choose the model that gives the closest fit to the data. Such a strategy inevitably favors a model with an excessive number of free parameters. In the best case scenario, where the data is not noisy, this results in a model with too many redundant free parameters. In the worst case scenario, where the data is noisy, this would result in a model with poor predictive power. This means that a good model is not merely the one that fits the existing data closely but, more importantly, the one that is both manageable (few redundancies) and has good predictive powers.

As an illustrative example of over-fitting, we generate a set of 46 data points \(\{(x_1, y_1), (x_2, y_2), \ldots, (x_{46}, y_{46})\}\) by adding stochastic noise to the cubic relation: \(y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i\), where \(\epsilon_i\) is a spatially uncorrelated stochastic term obeying a normal distribution of mean 0 and standard deviation 1 (i.e., \(\epsilon_i \sim N(0, 1)\)). \(x_i\) consisted of a set of points starting at \(x = -1.5\) and ending at \(x = 3\), spaced at intervals of 0.1.

We next choose five different models to fit the data generated (Fig. 1.1). The models chosen were a linear model (Fig. 1.1, Top Right), a quadratic model (Fig. 1.1, Middle Left), a cubic model (Fig. 1.1, Middle Right), a quartic model (Fig. 1.1, Bottom Left) and a 20th order polynomial (Fig. 1.1, Bottom Right), and they were fitted to the generated data via the maximum likelihood (MLE) method \(^2\).

\(^{1}\)We define the best model as the model with the highest predictive power. That means it would give the smallest mismatch with future data on average.

\(^{2}\)We will describe the method in Section 3.2.
discrepancy of the model against the data using the Residual Sum of Squares (RSS):

\[
\text{RSS} = \sum_i (y_i - f(x_i))^2,
\]

where \( f \) is the fitted model.

A 20th order polynomial provides the best fit to the data used. However, when we generate another set of data points from the original stochastic function \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), using a different set of randomly generated \( \epsilon_i \), we find that the originally fitted 20th order polynomial solution provides a worse fit compared to the cubic solution. This is shown in Fig. 1.2 and Table 1.1. We also note that the cubic solution has the best predictive power as noted by it having the lowest value in the second column of Table 1.1. In this case, the issue of over-fitting occurs when the proposed model (20th order polynomial) tries to fit not only the underlying cubic structure but also the noise structure generated by \( \epsilon_i \).

We can repeat our example for different noise levels to see the respective behavior of the polynomial curve fit. We generate data using \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + \lambda\epsilon_i \), where \( \lambda \) indicates the noise level of the data. We use this to further construct Tables 1.2-1.4. We see that the cubic model has the best predictive power for \( \lambda = 1.5 \) and 3. However, when \( \lambda = 5 \) and 10, the noise overwhelms the underlying structure of the data and the model with the best predictive power is the linear model.\(^3\) If the noise gets too large, lower order models become favored. This means that sometimes the best model to predict the data need not correspond to the original underlying structure of the data.

Due to the random fluctuations of the \( \epsilon_i \) term, over-fitting the noise structure reduces the predictive power of the 20th order polynomial model. Over-fitting the noise structure also has the added problem of unstable solutions. Revisiting the \( \lambda = 1.5 \) case, we see this in Fig. 1.3-1.7. In these five cases, the respective models were fitted to two different datasets and then compared. The two datasets were generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \) with different realizations of \( \epsilon_i \). We see that the 20th order polynomial is unstable under different realizations of \( \epsilon_i \) noise. This seems to suggest that model stability is an important consideration in guarding against over-fitting. We will explore this in Chapter 9.

The consideration of the issue of over-fitting is important because we are looking for models with good predictive powers. This highlights a need for a model selection procedure that tries to fit the underlying data structure but not its noise term, as the underlying structure does not change among the different datasets but the noise term does. The issue of over-fitting assumes the existence of this underlying data structure.

\(^3\)In our models considered, we did not consider the lowest order fit, which would be the \( y \) average of the generated data.
Figure 1.1: Top Left: A raw dataset was generated from the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d.. Top Right: Linear curve fitted to the data (Under-fitting). Middle Left: Quadratic curve fitted to the data (Under-fitting). Middle Right: Cubic curve fitted to the data (Optimal fit). Bottom Left: Quartic curve fitted to the data (Slight over-fitting). Bottom Right: 20th order polynomial fitted to the data (over-fitting).
Figure 1.2: Top Left: Another raw dataset that was generated from the curve: $y_i = x^3_i - 3x^2_i + x_i - 1 + 1.5\epsilon_i$, where $\epsilon_i \sim N(0, 1)$ and i.i.d.. The subsequent curves used were taken from Fig. 1.1. They were generated by taking the MLE fit against the dataset used in Fig. 1.1. The purpose is to see whether curves generated in one dataset would provide a good fit for another similar dataset. Top Right: Linear curve superposed over the data. Middle Left: Quadratic curve superposed over the data. Middle Right: Cubic curve superposed over the data. Bottom Left: Quartic Curve superposed over the data. Bottom Right: 20th order polynomial superposed over the data.
Table 1.1: Both datasets were computed from the curve: $y_i = x_i^3 - 3x_i^2 + x_i - 1 \pm 1.5 \epsilon_i$, where $\epsilon_i \sim N(0,1)$ and i.i.d.. Due to the different realization of the noise for $\epsilon_i$, the datasets are different. The respective curves were calculated by taking the MLE fit against the original dataset (middle column). We compute the residual sum of squares (RSS) of these curves against the two datasets using the formula: $RSS = \sum_i \{y_i - f(x_i)\}^2$, where $f$ is the respective polynomial curve. The table shows that while the 20th order polynomial provided the best fit against the original dataset, it did not do so for the new dataset.

<table>
<thead>
<tr>
<th>Order</th>
<th>Original Dataset</th>
<th>New Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>273.05</td>
<td>271.50</td>
</tr>
<tr>
<td>Quadratic</td>
<td>201.06</td>
<td>232.87</td>
</tr>
<tr>
<td>Cubic</td>
<td>89.755</td>
<td>133.56</td>
</tr>
<tr>
<td>Quartic</td>
<td>89.35</td>
<td>136.40</td>
</tr>
<tr>
<td>20th Order Polynomial</td>
<td>49.88</td>
<td>171.07</td>
</tr>
</tbody>
</table>

Table 1.2: Both datasets were computed from the curve: $y_i = x_i^3 - 3x_i^2 + x_i - 1 + 3\epsilon_i$, where $\epsilon_i \sim N(0,1)$ and i.i.d.. Due to the different realization of the noise for $\epsilon_i$, the datasets are different. The respective curves were calculated by taking the MLE fit against the original dataset (middle column). We compute the residual sum of squares (RSS) of these curves against the two datasets using the formula: $RSS = \sum_i \{y_i - f(x_i)\}^2$, where $f$ is the respective polynomial curve. The table shows that while the 20th order polynomial provided the best fit against the original dataset, it did not do so for the new dataset.

<table>
<thead>
<tr>
<th>Order</th>
<th>Original Dataset</th>
<th>New Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>514.32</td>
<td>599.41</td>
</tr>
<tr>
<td>Quadratic</td>
<td>434.12</td>
<td>589.62</td>
</tr>
<tr>
<td>Cubic</td>
<td>359.02</td>
<td>534.24</td>
</tr>
<tr>
<td>Quartic</td>
<td>357.41</td>
<td>545.60</td>
</tr>
<tr>
<td>20th Order Polynomial</td>
<td>199.514</td>
<td>684.292</td>
</tr>
</tbody>
</table>

Table 1.3: Both datasets were computed from the curve: $y_i = x_i^3 - 3x_i^2 + x_i - 1 + 5\epsilon_i$, where $\epsilon_i \sim N(0,1)$ and i.i.d.. Due to the different realization of the noise for $\epsilon_i$, the datasets are different. The respective curves were calculated by taking the MLE fit against the original dataset (middle column). We compute the residual sum of squares (RSS) of these curves against the two datasets using the formula: $RSS = \sum_i \{y_i - f(x_i)\}^2$, where $f$ is the respective polynomial curve. The table shows that while the 20th order polynomial provided the best fit against the original dataset, it did not do so for the new dataset.

<table>
<thead>
<tr>
<th>Order</th>
<th>Original Dataset</th>
<th>New Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>1126.98</td>
<td>1464.79</td>
</tr>
<tr>
<td>Quadratic</td>
<td>1035.14</td>
<td>1498.53</td>
</tr>
<tr>
<td>Cubic</td>
<td>997.278</td>
<td>1483.99</td>
</tr>
<tr>
<td>Quartic</td>
<td>992.814</td>
<td>1515.56</td>
</tr>
<tr>
<td>20th Order Polynomial</td>
<td>554.206</td>
<td>1900.81</td>
</tr>
</tbody>
</table>
Table 1.4: Both datasets were computed from the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 10\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d.. Due to the different realization of the noise for \( \epsilon_i \), the datasets are different. The respective curves were calculated by taking the MLE fit against the original dataset (middle column). We compute the residual sum of squares (RSS) of these curves against the two datasets using the formula: \( \text{RSS} = \sum_i (y_i - f(x_i))^2 \), where \( f \) is the respective polynomial curve. The table shows that while the 20th order polynomial provided the best fit against the original dataset, it did not do so for the new dataset.

<table>
<thead>
<tr>
<th>Order</th>
<th>Original Dataset</th>
<th>New Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>4113.52</td>
<td>5769.06</td>
</tr>
<tr>
<td>Quadratic</td>
<td>3989.13</td>
<td>5936.96</td>
</tr>
<tr>
<td>Cubic</td>
<td>3989.11</td>
<td>5935.98</td>
</tr>
<tr>
<td>Quartic</td>
<td>3971.26</td>
<td>6062.22</td>
</tr>
<tr>
<td>20th Order Polynomial</td>
<td>2216.82</td>
<td>7603.25</td>
</tr>
</tbody>
</table>

Figure 1.3: Comparison of two linear curves that were fitted to two different datasets. Each dataset was generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d..
Figure 1.4: Comparison of two quadratic curves that were fitted to two different datasets. Each dataset was generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d..

Figure 1.5: Comparison of two cubic curves that were fitted to two different datasets. Each dataset was generated by the curve: \( y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i \), where \( \epsilon_i \sim N(0, 1) \) and i.i.d..
Figure 1.6: Comparison of two quartic curves that were fitted to two different datasets. Each dataset was generated by the curve: $y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i$, where $\epsilon_i \sim N(0,1)$ and i.i.d..

Figure 1.7: Comparison of two 20th order polynomial curves that were fitted to two different datasets. Each dataset was generated by the curve: $y_i = x_i^3 - 3x_i^2 + x_i - 1 + 1.5\epsilon_i$, where $\epsilon_i \sim N(0,1)$ and i.i.d.
structure. However, unlike our illustrative example used, this underlying structure is unknown \textit{a priori} in reality (otherwise we would not need statistics in the first place). This begs the question of finding the most optimal method to achieve this goal of finding good predictive models.

### 1.1.2 Model selection techniques

Several statistical techniques have been proposed to address this issue of model selection. Examples of such techniques include the Akaike Information Criterion (AIC) \cite{2}, Bayesian Information Criterion (BIC) \cite{68} and Minimum Description Length (MDL) \cite{9}. Apparently, these techniques share the following form:

\[
\text{Criterion} = \text{Discrepancy} + \text{Cost},
\]  

(1.2)

where a smaller criterion value indicates a better model. The discrepancy term indicates the amount of mismatch between the model and the data. Its smaller value indicates a smaller mismatch with the data and hence a higher likelihood of the model fitting the underlying structure. However, a smaller value in the mismatch would also increase the risk of over-fitting the noise structure. This can be mitigated by the cost term which indirectly indicates the amount of over-fitting. This usually involves a term that contains a crude estimate of the model ‘complexity’ (i.e., number of free parameters) since there is a rough relationship between model ‘complexity’ and over-fitting, especially for noisy datasets. We saw this in Section 1.1.1 where the 20th order polynomial model had a higher number of free parameters compared to the lower order models (e.g., quadratic models, linear models). We should always remember that the ultimate aim of the strategy (of having two competing terms) is to find the model with the best predictive power.

Both the low discrepancy value and the low cost value are highly desired in a good model. However, these are competing terms (i.e., an over-fitted model has a low discrepancy value but a high cost value), so a good model is one which finds a compromise between both competing terms to give a good (small) criterion value. Despite their superficial similarity, the three criteria mentioned above are derived from different underlying principles. AIC is derived from the Kullback-Leibler divergence (KL) \cite{24}, BIC is derived from ideas in Bayesian inference \cite{15} and MDL is derived from the concept of Kolmogorov complexity \cite{22, 46, 73, 74}. We will compare and contrast the other techniques to AIC in Section 8.4.
1.2 Akaike information criterion

In the first part of our thesis, we study the use and validity of the AIC model selection technique. AIC is based on the idea of Shannon entropy [24], offering a relative estimate of the information loss (Kullback-Leibler divergence) when using a candidate model to describe reality. Minimizing this divergence between the candidate model and the underlying data generating truth can be interpreted as minimizing the fit between the model against any future datasets, favoring a model with a high predictive power in the process. The use of AIC has been extended through the use of AIC differences to include a measure of confidence in the AIC procedure [3, 4, 17]. However, such an analysis does not take into account the statistical uncertainty of the AIC estimate. Through the use of bootstrap simulations, we developed tools to study the effect of estimator uncertainty, paying particular attention to the shape of the distribution of the AIC difference. When taking the effect of estimator uncertainty into account, the original analysis of AIC differences was shown to be inadequate and sometimes wrong. This will be shown in the thesis using the practical example of observational cosmology.

1.3 Observational cosmology

There are two ways to collect empirical data. The first is to use experiments to recreate processes that produce data. The second way is to observe the processes occurring in nature and record them. We can call the latter observational sciences and the former experimental sciences. Examples of the observational sciences include astronomy, seismology and ecology. The observational sciences occur when it becomes extremely impractical to recreate the actual physical processes in a lab. Due to its reliance on naturally occurring processes, data is not as easily available. Thus, there is a need for model selection techniques to infer the best model given the scarcity of empirical data.

In this thesis, we will use the practical example of observational cosmology to explore the limitations of AIC. In particular, we will apply AIC to Type Ia supernova (SN) events and use it to compare between four late time acceleration expansion models of the universe. This would be referred to as cosmological model selection in this thesis. Our focus is not necessarily on the ‘best’ cosmological model but on the reliability of using AIC in comparing between them. Thus, we did not consider all the possible models found in the literature nor did we use the most exhaustive datasets.\(^4\)

\(^4\)We included the $\Lambda$CDM model in our set of candidate models. This is considered the leading model to describe the late time accelerated expansion of the universe.
1.4 Different model selection criteria

In the second part of our thesis, we will look beyond its philosophy and compare AIC to other model selection approaches. We will also introduce a new criterion for model selection that is based on model stability. This is philosophically different (though related) to the idea of using model complexity. A lower model complexity is generally favored in some model selection strategies. However, this may not necessarily be the best approach. For example, adding error correcting codes to a model increases both its complexity and predictive power. Thus, we do not necessarily need to penalize complexity in our cost term. In our proposed approach, we will associate model stability, instead of model complexity, with better data prediction. While the first part of the thesis focuses on the validity of AIC under the assumption that it is conceptually a good technique for model selection, the second part would critically assess those assumptions.

1.5 Roadmap

The thesis is organized as follows. In Chapter 2, we briefly review the information theoretic origin of AIC by introducing the Kullback-Leibler (KL) divergence. This idea is explored further in Chapter 3 by demonstrating its use in model class comparison, deriving the AIC formula in the process. The use of AIC is developed further by the concept of confidence levels [16, 17], which we will review in Chapter 4, exploiting the differences between different AIC values. In Chapter 5, we provide some background to cosmological model selection and use it as a demonstration of the AIC method. The limits of the existing approach by AIC is discussed in Chapter 6 and addressed in Chapter 7. We would next further explore the philosophical misconceptions of the technique in Chapter 8. In this chapter, we also compare AIC to other model selection techniques. In Chapter 9, we introduce an alternative type of cost term which penalizes stability instead of complexity. Finally, we summarize our findings in Chapter 10 and discuss the implications of our results.
Chapter 2

Kullback-Leibler Divergence

2.1 Introduction

The Kullback-Leibler (KL) divergence is a commonly used quantity that measures the discrepancy of one probability distribution with respect to another probability distribution. This is not a metric because it is non-symmetric and does not obey the triangle inequality and is derived from the information theoretic concept of Shannon entropy [24].

2.2 Shannon entropy of a probability distribution

Suppose we have a discrete random variable $X$ which obeys the probability distribution $p(x)$, where we will define $\mathcal{X}$ as the set of all outcomes for $X$ and $x$ as a particular element of this set. We define the Shannon entropy $H(X)$ as the following:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x).$$

If we sample the distribution many times, this term asymptotically quantifies the average uncertainty that results from each sampling of the random variable $X$. As an illustration, consider two coins. The first is a fair coin that has a 50% chance of either a HEADS or TAILS outcome during a toss while the other has HEADS on both sides. The Shannon entropy of a toss using the first coin is 1 bit while that of a toss using the second coin is 0 bit, indicating the greater uncertainty when using the first coin.

$H(X)$ has the following properties [24]:

1. Nonnegativity: $H(X) \geq 0$.

2. Upperbound: $H(X) \leq \ln(N)$, where $N$ is the cardinality of $\mathcal{X}$.


12
2.3 Modeling a process with a ‘wrong’ distribution

If we use a different probability distribution model $q$ to model a random variable $X$ with an underlying probability distribution $p$, the uncertainty of modeling $X$ is quantified by $H[p, q] = -\sum_{x \in X} p(x) \log q(x)$ [24]. This quantity is greater or equal to $H[p] = H(X) = -\sum_{x \in X} p(x) \log p(x)$. Using $\log x \leq x - 1$, we have

$$H[p, q] - H[p] = -\sum_{x \in X} p(x) \log \frac{q(x)}{p(x)} \geq -\sum_{x \in X} p(x) \left( \frac{q(x)}{p(x)} - 1 \right) = -\sum_{x \in X} q(x) + \sum_{x \in X} p(x) = 0.$$ 

Thus, $H[p, q] \geq H[p]$. The lowest uncertainty that we can have about $X$ is when we use its underlying distribution $p$ to model $X$. However, in a lot of cases, we do not know $p$ and try to guess $q$ as a close approximation of $p$. This approximation of $p$ is guessed by looking at the empirical distribution. This could result in a larger uncertainty when $p \neq q$. Hence, from the modeler’s point of view, $H(X)$ indicates the lower bound to the uncertainty of $X$. This uncertainty can be reinterpreted as the code length required to describe the signal from $X$ using some pre-defined coding scheme.\(^1\) This lower bound is achieved when the probability distribution used to predict and model the outcome of $X$ matches its underlying probability $p$.

2.4 Mismatch between two probability distributions

We can define the extra code length due to using an incorrect model as the KL divergence $D(p \parallel q)$:

$$D(p \parallel q) = H[p, q] - H[p] = -\sum_{x \in X} p(x) \log \frac{q(x)}{p(x)} \quad (2.1)$$

It has the following properties [24]:

1. Nonnegativity: As shown in Section 2.3, $D(p \parallel q) \geq 0$ with equality iff $p = q$.

2. Convexity: $D(p \parallel q)$ is convex with respect to the pair $(p, q)$.

\(^1\)This thesis assumes some basic knowledge of coding theory. For more information, we refer the interested reader to [24].
3. Non-symmetric: $D(p \parallel q) \neq D(q \parallel p)$ if $p \neq q$.

4. Does not obey triangle inequality $D(p \parallel q) \leq D(p \parallel r) + D(r \parallel q)$.

Due to its asymmetry, $D(p \parallel q)$ cannot be used as a distance metric\(^2\) between two probability distributions. However, we can use it to quantify the extra code length (uncertainty) to describe elementary events when we use the wrong distribution $q$ to model the empirically observed data that is generated by the reference probability distribution $p$, which we call the truth. We can compare a set of candidate models $\{q_1, q_2, \ldots, q_n\}$ against $p$ by computing and comparing their respective $D(p \parallel q)$. This allows us to compare the faithfulness of the various candidate models against the truth and is an important idea in deriving AIC.

### 2.5 General form of KL divergence

It should be noted that $D(p \parallel q)$ is not restricted to the discrete form. We can rewrite the KL divergence as $D(p \parallel q) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{h(x)} dx$, where $f(x)$ and $h(x)$ are probability density functions.

This continuous KL divergence is used to quantify the mismatch between two probability densities. Probability density functions are used in regression modeling. Hence, we would use this form when comparing between models. It should be noted that, using the language of probability measures, both the discrete and continuous forms have a unified expression: If $Q$ and $P$ are probability measures over a set $\mathcal{X}$, and if $P$ is absolutely continuous with respect to $Q$, then

$$D(P \parallel Q) = \int_{\mathcal{X}} \log \frac{dP}{dQ} dP,$$

(2.2)

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative [71].

### 2.6 The role of KL divergence in large deviation theory

Large Deviation Theory (LDT) [80] is a tool to study large deviation or how many samples one needs to achieve the weak Law of Large Numbers. If we treat the observed sequence as the empirical probability

---

\(^2\)There have been attempts to symmetrize the KL divergence. For example, we could define a quantity $\frac{1}{2} \{D(q \parallel p) + D(p \parallel q)\}$. However, such a quantity would not obey the triangle inequality. To prove by counter example, we can consider the case of a three outcome discrete probability distribution. We define the following three distributions: $a = \{0.3, 0.3, 0.4\}$, $b = \{0.25, 0.35, 0.4\}$, $c = \{0.16, 0.33, 0.51\}$ and show that the symmetrized KL divergence does not obey the triangle inequality.
distribution, LDT\textsuperscript{3} would tell us about the statistical behavior of the empirical probability distribution. This is also known as Sanov’s Theorem [24].

To prove Sanov’s Theorem, we refer to the demonstration by Oono [55]. This is reproduced almost verbatim for the convenience of the reader. Let us first state \{X_i\} as a sequence of \(N\) i.i.d. random variables distributed according to the density distribution function \(f\). The empirical density \(f_N(x)\) is given by

\[
f_N(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(X_i - x).
\]

(2.3)

We want to study the probability that \(f_N\) asymptotically behaves like distribution \(g\). We will denote this as \(P(f_N \sim g)\), where \(\sim\) indicates the asymptotic relation. The Level 2 large deviation principle reads

\[
P(f_N \sim g) \sim e^{-NI^{(2)}(g)},
\]

(2.4)

where the rate functional \(I^{(2)}(g)\) is a function on the space of distributions, and satisfies:

1. \(I^{(2)}(g)\) is a convex nonnegative function.

2. \(I(f) = 0\) is a unique minimum.

We refer the reader to [55] for more background information.

Sanov’s Theorem tells us that

\[
I^{(2)}(g) = D(g \| f) = \int dx \, g(x) \log \frac{g(x)}{f(x)}.
\]

(2.5)

To show this, we use the characteristic functional technique. First we notice that

\[
P(f_N \sim g) = E(\Delta(f_N - g)),
\]

(2.6)

where \(\Delta\) is the functional delta function.\textsuperscript{4} Therefore, we get the ‘partition function’

\[
Z_N(\xi) = \int \delta[g] P(f_N \sim g) e^{\int \xi(x)g(x)dx} = E \left[ \exp \left( \sum \xi(X_i) \right) \right] = E \left( e^{\xi(X_i)} \right)^N = \left[ \int dx f(x)e^{\xi(X_i)} \right]^N.
\]

(2.7)

\textsuperscript{3}This is known as Level 2 in LDT. For this thesis, we do not consider Level 1 and Level 3 of LDT.

\textsuperscript{4}A function \(f(\alpha)\) may be interpreted as the \(\alpha\)-component of the vector \(f\). Functional analysis may be informally interpreted as the analysis of functions regarded as such vectors. The \(N\)-dimensional \(\delta\)-function is defined as \(\delta(v) = \delta(v_1) \cdots \delta(v_N)\). Analogously, \(\Delta(f) = \delta(f(\alpha)) \cdots\), where the product is the continuous product.
On the other hand, by definition

\[ Z_N(\xi) = \int \delta[g] e^{N \left[ \int \xi(x) g(x) dx - I^{(2)}(g) \right]} \sim e^{N \sup_g \left[ \int \xi(x) g(x) dx - I^{(2)}(g) \right]} . \]  

Therefore, introducing the ‘free energy’ (thermodynamic potential), we obtain

\[ A(\xi) = \lim_{N \to \infty} \frac{1}{N} \log Z_N(\xi) = \log \int dx f(x) e^{\xi(x)} = \sup_g \left[ \int dx \xi(x) g(x) - I^{(2)}(g) \right] . \]  

Due to the Legendre-Fenchel transformation [65],

\[ I^{(2)}(g) = \sup_{\xi} \left[ \int dx \xi(x) g(x) - A(\xi) \right] . \]  

To calculate the supremum, we functional-differentiate \( [ \ ] \):

\[ \frac{\delta}{\delta \xi(x)} \left[ \int dx \xi(x) g(x) - A(\xi) \right] = g(x) - \frac{f(x) e^{\xi(x)}}{\int dx f(x) e^{\xi(x)}} . \]  

Equating it to zero, we obtain: \( g(x) = f(x) e^{\xi(x)}/e^{A(\xi)} \). This means that \( \xi = \log(g/f) + A \) and putting it into (2.10), we get Sanov’s Theorem.

The implication of Sanov’s Theorem is that it asymptotically relates the KL divergence \( D(P \parallel Q) \) to the probability of a sequence, generated by \( Q \), mimicking the empirical distribution \( P \). We can call this the probability of confusing one distribution \( P \) for another \( Q \) or likelihood to produce \( P \) from \( Q \). For the sake of brevity, we will simply refer to it as confusion probability for the rest of the thesis. This confusion probability can also be derived via Stein’s lemma (Appendix B), which is framed in the language of hypothesis testing. We will later see the significance of this confusion probability in assessing confidence levels to AIC selected models.
2.7 Geometric interpretation of KL divergence for small perturbation

Let $\theta_0$ represent the set of Maximum Likelihood Estimation (MLE) parameters for a probability distribution $f$, and let $\Delta \theta$ represent a small perturbation in parameter space about $\theta_0$.

\[
D(f(x|\theta_0) \parallel f(x|\theta_0 + \Delta \theta)) = D(f(x|\theta_0) \parallel f(x|\theta_0)) + \sum_i \left\{ \frac{\partial D(f(x|\theta_0) \parallel f(x|\theta))}{\partial \theta^i} \bigg|_{\theta = \theta_0} \Delta \theta^i \right\} \\
+ \frac{1}{2} \sum_{i,j} \left\{ \frac{\partial^2 D(f(x|\theta_0) \parallel f(x|\theta))}{\partial \theta^i \partial \theta^j} \bigg|_{\theta = \theta_0} \Delta \theta^i \Delta \theta^j \right\} + O(\Delta \theta^3).
\]

The first term is zero from the definition of KL (2.1). The second term goes to zero because $\theta_0$ is the MLE of the parameters. Neglecting the higher order deviations and treating $\frac{\partial^2 D(f(x|\theta_0) \parallel f(x|\theta))}{\partial \theta^i \partial \theta^j} \bigg|_{\theta = \theta_0}$ as a metric $I_{ij}$, we see that the KL divergence behaves approximately as a distance for small perturbations.

\[
D(f(x|\theta_0) \parallel f(x|\theta_0 + \Delta \theta)) \approx \frac{1}{2} \sum_{i,j} I_{ij} \Delta \theta^i \Delta \theta^j,
\]  

(2.12)

where $I_{ij}$ is known as the Fisher Information matrix (metric) [6]. We also see that for small perturbations $D(P \parallel Q) \approx D(Q \parallel P)$. 

17
Chapter 3

Model Class Comparison

3.1 Introduction

We define a model class as the totality of model probability distributions with the same parametric form (but with different parameter values). For example, the function \( f(x) = 3x^2 - 6x + 3 \) is a model and the set of all quadratic functions \( F \) is the corresponding model class; \( f \in F \). Within each model class, there is a set of parameters (the ‘best’ model) that gives the lowest KL divergence with respect to the truth \( T \). Thus, to choose the ‘best’ model class we must first choose the ‘best’ model (parameter set) from a particular model class as the representative of the model class. This is done by MLE. The model class selection strategy is thus obtained by comparing the KL divergence of the representative models of the individual model classes.

However, the truth is unknown \textit{a priori}, so for a given representative model \( A \), \( D(T \| A) \) cannot be evaluated directly. The AIC method tries to solve this problem by computing a term known as the AIC value \cite{2}, which is an asymptotically unbiased estimator for \( D(T \| A) \) (up to a fixed offset that is independent of the representative models). Since the fixed offset is independent of the models, and hence the choice of model classes, a comparison of the AIC values is a useful surrogate for the strategy of comparing the associated KL divergence of the different candidate model classes against \( T \).

As already mentioned explicitly, our model class comparison involves a comparison of their respective best fit representative models. Thus, in the rest of this thesis, when we are comparing model classes, we sometimes refer to this as comparing models.

3.2 Regression models

In regression modeling, we model a series of data points \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \) as the conditional probability density function \( f((y_1, y_2, \ldots, y_n) \mid \vec{\theta}, \{x_1, 2, \ldots, x_n\}) \) with the parameter vector \( \vec{\theta} \), which is a
$k$-vector that represents $k$ independent parameters. If we further assume that the data points are statistically independent, we can reexpress this probability density in the following form:

$$f \left( \{y_1, y_2, \ldots, y_n\} \mid \mathbf{x}, \{x_1, x_2, \ldots, x_n\} \right) = \prod_{i=1}^{n} f_i(y_i \mid \mathbf{x}, x_i), \quad (3.1)$$

where $f_i$ describes the conditional probability density of getting $y_i$ given $\mathbf{x}$ and point $x_i$.

LDT shows that the resulting error would approach a Gaussian, as the data is typically subjected to multiple independent noise sources. In our regression model, we will assume a Gaussian noise with $Y_i \sim N(g(x_i), \sigma_i^2)$, where $\mathbf{\theta}$ affects the fitted model (mean) $g$ and the variance $\sigma_i^2$. Using this added assumption of normality, the probability density further reduces to:

$$f \left( \{y_1, y_2, \ldots, y_n\} \mid \mathbf{x}, \{x_1, x_2, \ldots, x_n\} \right) = \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \right) \exp \left[ -\sum_{i}^{n} \frac{(y_i - g(x_i))^2}{2\sigma_i^2} \right]. \quad (3.2)$$

Given this probability distribution, the standard procedure is to apply MLE to solve for the parameters of our model. This is usually achieved by differentiation of the logarithm of the probability density function with respect to the parameters:

$$\frac{\partial \log f \left( \{y_1, y_2, \ldots, y_n\} \mid \mathbf{x}, \{x_1, x_2, \ldots, x_n\} \right)}{\partial \mathbf{\theta}_j} = 0. \quad (3.3)$$

We have $k$ such equations and solve this set of simultaneous equations. It should also be noted that the method of finding the $\chi^2$ fit is equivalent to MLE for the case of Gaussian noise.

### 3.3 Derivation of AIC from the KL divergence

The strategy of using the KL divergence in model selection has an information theoretic justification. However, it is impossible to directly evaluate the KL divergence of a best fit representative model $\hat{f}$ against $T$, since $T$ is unknown a priori. We can solve this problem by computing the AIC value.

In the derivation of AIC [2], we first isolate the model dependent term:

$$D(T \mid \hat{f}) = E_T[\log T] - E_T[\log \hat{f}], \quad (3.4)$$

The first term $E_T[\log T]$ only depends on $T$ so a comparison between the different KL divergences is essentially...
a comparison between the respective second terms $-\mathbb{E}[\log \hat{f}]$. AIC is essentially an unbiased estimate of $-\mathbb{E}[\log \hat{f}]$. Since $\hat{f}$ depends on the particular realization of the empirical data, finding the unbiased estimate of $\mathbb{E}[\log \hat{f}]$ involves evaluating $\mathbb{E}_X \mathbb{E}_Y [\log f(X \mid \hat{\theta}(Y))]$, where $X$ and $Y$ are sampled from $T$ and $\hat{\theta}(Y)$ is the MLE-derived parameters of $f$ using data $Y$.

To do so, we first define $\theta^*$ where it satisfies

$$\mathbb{E}_X [\log f(X \mid \theta^*)] = \max_{\theta} \mathbb{E}_X [\log f(X \mid \theta)].$$

Representing $T$ with the probability density $g(x)$ and using the Taylor expansion, we have:

$$\mathbb{E}_X [\log f(X \mid \hat{\theta}(Y))] = \int g(x) \log f(X \mid \hat{\theta}(Y))dx \approx \int g(x) \log f(X \mid \theta^*)dx + \left\{ \int g(x) \frac{\partial}{\partial \theta} \log f(X \mid \theta^*)dx \right\} (\hat{\theta}(Y) - \theta^*)$$

$$+ \frac{1}{2} (\hat{\theta}(Y) - \theta^*)^T \left\{ \int g(x) \frac{\partial^2}{\partial \theta \partial \theta'} \log f(X \mid \theta^*)dx \right\} (\hat{\theta}(Y) - \theta^*).$$

Since $\mathbb{E}_X [\log f(X \mid \theta)]$ is maximum at $\theta^*$, the second term vanishes. If we define the Hessian matrix $J(\theta^*)$ by

$$J(\theta^*) = -\mathbb{E}_X \left\{ \frac{\partial^2}{\partial \theta \partial \theta'} \log f(X \mid \theta^*) \right\},$$

we can rewrite the expansion as

$$\mathbb{E}_X [\log f(X \mid \hat{\theta}(Y))] \approx \int g(x) \log f(X \mid \theta^*)dx - \frac{1}{2} (\hat{\theta}(Y) - \theta^*)^T J(\theta^*) (\hat{\theta}(Y) - \theta^*),$$

where $\hat{\theta}(Y)$ is the MLE-derived parameters from data $Y$. Taking the asymptotic limit, 

$$(\hat{\theta}(Y) - \theta^*)^T J(\theta^*) (\hat{\theta}(Y) - \theta^*)$$

converges to a centrally distributed $\chi^2$ random variable with $k$ degrees of freedom, where $k$ is the dimension of the parameter $\theta$. Thus, we can show that

$$\mathbb{E}_Y \left\{ (\hat{\theta}(Y) - \theta^*)^T J(\theta^*) (\hat{\theta}(Y) - \theta^*) \right\} = k.$$

Thus,

$$\mathbb{E}_X \mathbb{E}_Y [\log f(X \mid \hat{\theta}(Y))] \approx \mathbb{E}_X \mathbb{E}_Y [\log f(X \mid \theta^*)] - \frac{k}{2}$$

$$= \mathbb{E}_X [\log f(X \mid \theta^*)] - \frac{k}{2}.$$
The next stage in the derivation involves evaluating the term $E_X \left[ \log f(X \mid \theta^*) \right]$. This is because we do not know $\theta^*$ and cannot evaluate it directly. To do so, we again take the Taylor expansion and set the first order expansion to zero:

$$
\log f(X \mid \theta^*) \approx \log f(X \mid \hat{\theta}(X)) + \frac{1}{2} \left( \theta^* - \hat{\theta}(X) \right)^T \frac{\partial^2 \log f(X \mid \hat{\theta}(X))}{\partial \theta \partial \theta} \left( \theta^* - \hat{\theta}(X) \right).
$$

(3.5)

Substituting this into $E_X \left[ \log f(X \mid \theta^*) \right]$, we obtain

$$
E_X \left[ \log f(X \mid \theta^*) \right] \approx -\frac{1}{2} E_X \left[ \left( \theta^* - \hat{\theta}(X) \right)^T J(\hat{\theta}(X)) \left( \theta^* - \hat{\theta}(X) \right) \right] + E_X \left[ \log f(X \mid \hat{\theta}(X)) \right].
$$

(3.6)

Taking the asymptotic limit, $J(\hat{\theta}) \to J(\theta^*)$, and recalling that $E_X \left[ \left( \hat{\theta}(X) - \theta^* \right)^T J(\theta^*) \left( \hat{\theta}(X) - \theta^* \right) \right]$ converges to a centrally distributed $\chi^2$ random variable with $k$ degrees of freedom, where $k$ is the dimension of the parameter $\theta$, we can show that

$$
E_Y E_X \left[ \log f(X \mid \theta^*) \right] \approx -\frac{k}{2} + E_Y \left[ \log f(Y \mid \hat{\theta}(Y)) \right].
$$

(3.7)

Thus, we can show that

$$
E_Y E_X \left[ \log f(X \mid \hat{\theta}(Y)) \right] \approx -k + E_Y \left[ \log f(Y \mid \hat{\theta}(Y)) \right].
$$

(3.8)

$-\log f(Y \mid \hat{\theta}(Y)) + k$ is thus an unbiased estimator (up to a fixed offset) of the KL divergence against $T$.

Thus, we can compute the AIC\(^1\) of the representative (best fit) models of various candidate model classes, which in the asymptotic limit is known \([2]\) to be

$$
\text{AIC} = 2k - 2 \log(L_{ML}),
$$

(3.9)

where $L_{ML}$ is the probability function using the MLE-derived parameters. This derivation was made under the assumptions that \([19]\):

1. “The reference model is contained in the candidate class under consideration.”
2. “The vector of MLE-derived estimators satisfies the conventional large-sample properties of MLE’s.” As the sample size increase, they asymptotically become normally-distributed unbiased minimum variance estimators.

\(^1\)It should be noted that as a matter of convention AIC is twice the unbiased estimator.
When we make a further assumption that the distribution follows the parametrization in (3.2),

\[
\text{AIC} = 2k + \frac{1}{2} \sum_{i=1}^{n} \log 2\pi\sigma^2_i + \chi^2_{ML},
\]

(3.10)

where \(\chi^2_{ML}\) is the \(\chi^2\) fit derived using the MLE parameters and \(n\) is the number of data points. If we ignore the common \(\frac{1}{2} \sum_{i=1}^{n} \log 2\pi\sigma^2_i\) term, this further reduces to

\[
\text{AIC} = 2k + \chi^2_{ML}.
\]

(3.11)

This is used when the error bars of the data are known. If the error bars are unknown, we would have to make the assumption that they have identical values of \(\sigma\) and are related to the residual sum of squares (RSS) by the relation: \(\sigma^2 = \text{RSS}/n\), where \(n\) is the number of data points. When we apply this assumption to (3.10) and ignore the model independent terms, (3.10) reduces to an alternative form of AIC [16]:

\[
\text{AIC} = 2k + n \ln \left( \frac{\text{RSS}}{n} \right).
\]

(3.12)

This reduces the AIC method to the simple act of calculating the AIC values for the different candidate model classes and ranking them according to their respective values, where a smaller AIC value indicates a better model. Its simplicity in calculating the value is part of the reason for its popularity.

It is worth noting that the AIC estimate (3.11) of the KL divergence assumes that the number of data points is sufficiently large. AIC in the form written above is an unbiased estimator for large datasets. For smaller datasets, the \(2k\) term can be corrected by an additional \(2k(2k+1)/(n-k-1)\) term to approximately correct for the bias due to finiteness of the dataset,\(^3\) where \(n\) is the number of data points in a single dataset. While further studies to obtain a more accurate expression for this term are possible, for the cases we shall consider, this correction is always less than 0.06 (which will be seen to be negligible for our purposes) and will only decrease in importance when more data is collected. We shall therefore ignore this correction altogether in this thesis.

\(^2\)This is only under our choice of models, which we have restricted to Gaussian models.

\(^3\)This was derived under Gaussian assumption for linear models [41, 43, 42].
Chapter 4

Confidence Levels from the AIC Differences

4.1 Introduction

An AIC value by itself has no meaning as it contains an arbitrary offset\(^1\) that is affected by the dataset. It only has meaning when it is compared against other AIC values derived from the same dataset. Originally, this comparison of AIC values involves an ordinal ranking of the AIC values, this original analysis was further extended by Akaike [3, 4, 16, 17] to take into account the interval between the AIC values.

It is intuitively obvious that the smaller the AIC difference between two models, the harder it becomes to judge which model is better; even if the AIC estimate of this difference in the KL divergence can be obtained without estimator error, the small difference would make it difficult to tell the two probabilistic models apart for a small number of observations. Hence, there is a need to associate a confidence level for distinguishing between a model \(A\) and another model \(B\) using the AIC difference between them.

4.2 The interpretation of AIC differences

We start with a setup of \(m\) candidate models and a set of corresponding AIC values \(\{\text{AIC}(1), \text{AIC}(2), \ldots, \text{AIC}(m)\}\). By first defining \(\text{AIC}_{\text{min}}\) as the smallest AIC value from this set, we can define the quantity \(\Delta_i\) as

\[
\Delta_i \equiv \text{AIC}(i) - \text{AIC}_{\text{min}}, \tag{4.1}
\]

where \(i \in \{1, \ldots, m\}\).

According to Burnham and Anderson [16, 17], the \(\Delta_i\) value allows for a quick ‘strength of evidence’ comparison and ranking of the candidate models. The rule of thumb suggested by them: models having \(\Delta_i \leq 2\) have substantial support (evidence), those where \(4 \leq \Delta_i \leq 7\) have considerably less support, while

\(^1\)This is the first term in (3.4).
models having $\Delta_i \geq 10$ have no or little support.

The AIC difference should be considered regardless of the size of the AIC values calculated. For example, if we compare two models with AIC values of 800,000,000 and 800,000,007, respectively. One might think a difference of 7 is trivial. However, we should remember that AIC is an estimate of the KL divergence up to a fixed incalculable offset that is independent of the candidate models considered. Hence, only differences in AIC are useful as ‘strengths of evidence’, since it is free of the effects of this offset.

The justification for using $\Delta_i$ was based on Akaike’s interpretation of AIC differences in a Bayesian framework [3, 4]. The simple transformation $\exp \left[ -\frac{\Delta_i}{2} \right]$ provides the likelihood of the model given the data: $L(\text{Model}_i \mid \text{data})$. This is derivation is shown in Appendix A. The relative likelihood of model $i$ over model $j$ is the ratio $L(\text{Model}_i \mid \text{data})/L(\text{Model}_j \mid \text{data})$ and is referred to as the ‘evidence ratio’ by Burnham and Anderson [16, 17].

### 4.3 Reinterpretation of the AIC differences

Akaike’s interpretation of the AIC differences (Appendix A) was done in the Bayesian framework and made assumptions about the structure of the model priors. In the Bayesian framework, this involves integrating conditional probabilities with model priors. Without indicating specific form for the priors, we are unable to do the integration in the Bayesian analysis.

We can interpret the AIC differences through another approach. From Section 2.6, the empirical distribution $f_N$ obtained from observing $N$ i.i.d. random variables, which are each generated according to distribution $f$, can mimic distribution $g$ according to the following probability:

$$P(f_N \sim g) \sim \exp \left[ -ND(g \parallel f) \right], \quad (4.2)$$

where $\sim$ implies the asymptotic relation for large $N$.

$P$ is the confusion probability we mentioned in Section 2.6. This relation can also be derived via Stein’s lemma (Appendix B). This second derivation is different from our earlier derivation as it is framed in the language of hypothesis testing.

In the context of regression models, we are dealing with datasets which consist of multiple data points. We should not confuse $N$ with the number of data points $n$. $N$ refers to the number of data samples instead.
If we make repeated observations to generate $N$ similar datasets, we can define the individual probabilities $P_i(C)$, where $P_i(C)$ is the probability that model $C$ is indistinguishable from the truth $T$ using dataset $i$. We can define an asymptotic geometric average as

$$P(C) = \lim_{N \to \infty} \left( \prod_{i=1}^{N} P_i(C) \right)^{1/N}.$$  

(4.3)

For the sake of brevity, we shall also call this the confusion probability.

To show its relation to the AIC difference, let us start with two models $A$ and $B$ with $f_A$ and $f_B$ as their respective probability distribution functions. Furthermore, their respective AIC values are $a$ and $b$, and there is a difference of $\Delta_{A,B} = a - b$ between their AIC values. The relative strength of $P(A)$ and $P(B)$ tells us which model is more faithful to the truth. Hence, it would be useful to study their ratio and relate it to $\Delta_{A,B}$.

$$\frac{P(A)}{P(B)} = \frac{\exp\left[ -D \left( \text{truth} \parallel f_A(X \mid \hat{\theta}_A) \right) \right]}{\exp\left[ -D \left( \text{truth} \parallel f_B(X \mid \hat{\theta}_B) \right) \right]} = \frac{\exp\left[ E_X[\theta_0] \log f_A(X \mid \hat{\theta}_A) \right]}{\exp\left[ E_X[\theta_0] \log f_B(X \mid \hat{\theta}_B) \right]} \approx \frac{\exp[-a/2]}{\exp[-b/2]} = \exp[-\Delta_{A,B}/2],$$

where $X$ represents data sampled from the truth $\theta_0$. $\hat{\theta}_A$ and $\hat{\theta}_B$ are the maximum likelihood parameters of model $A$ and $B$. We used the $\approx$ symbol to denote the relationship between a quantity and its estimator. Hence, we can show that the ratio of confusion probabilities $P(A)/P(B)$ leads to $\exp[-\Delta_{A,B}/2]$.

In this approach, we have used the frequentist framework and hence do not need to make assumptions about model priors, since they do not exist in such a framework. This is intuitively similar to interpreting $\exp[-\Delta_{A,B}/2]$ as the ratio of model likelihood $L(\text{Model}_A|\text{data})/L(\text{Model}_B|\text{data})$ in the Bayesian framework. There are subtle differences in both interpretations. The confusion probability interpretation assumes an (Frequentist) existence of some reference truth which is absent in the model likelihood (Bayesian) framework. Either way, we will refer to either the model likelihood and confusion probability as model confidence.

---

2They have the same set of explanatory variables and associated error bars. The set of response variables follow the same set of statistical distributions.
4.4 Akaike weights

Using either interpretation, we can show that for a pair of models $A$ and $B$,

$$\frac{\text{Model Confidence of } A}{\text{Model Confidence of } B} = \exp[-\Delta_{A,B}/2]. \quad (4.4)$$

(4.4) quantifies the intuitive idea that it is easier to select a model over another if the AIC difference is large.

It might be convenient to normalize $\exp[-\Delta_i/2]$ such that they sum to 1:

$$w_i = \frac{\exp[-\Delta_i/2]}{\sum_j \exp[-\Delta_j/2]}, \quad (4.5)$$

$w_i$ is known as the Akaike weights and has been suggested by Burnham and Anderson [17] as a useful way for quantifying the weight of evidence in favor of a particular model out of the list of candidate models. Using a Bayesian interpretation of AIC, they showed that $w_i$ can be interpreted as the probability that model $i$ is the KL best model for the data [17]. If we take the ratio $w_i/w_j$, we would get back the ratio of model confidence. $w_i$ is dependent on the number of candidate models and can be made absurdly small if the number of candidate models becomes too large. In contrast, $w_i/w_j$ does not suffer from such dependence and it is for this reason that we favor the use of probability ratios over Akaike weights.

4.5 Akaike threshold

One can use the idea of AIC differences, that was described in Section 4.2 and Section 4.3, to modify the AIC methodology described above and suppress the probability of obtaining incorrect results by introducing a threshold $\Delta_{\text{threshold}}$. Then, rather than ranking all models according to the smallness of their AIC values, one adopts the procedure where a model $A$ is ranked to be better than model $B$ if $\Delta_{A,B} < -\Delta_{\text{threshold}}$, while any two models with a AIC difference smaller than $\Delta_{\text{threshold}}$ are considered of equal rank. (4.4) shows that choosing a large enough value of the threshold $\Delta_{A,B}$ implies a high probability that the selected model is truly the better one. A large value of $\Delta_{\text{threshold}}$ however also increases the number of models where the AIC differences are in the range $-\Delta_{\text{threshold}} < \Delta_{A,B} < \Delta_{\text{threshold}}$. Since this procedure cannot discriminate between such models, we shall call such a model selection result indeterminate. In our convention, we also define the converse of the indeterminate case as the determinate case ($|\Delta_{A,B}| > \Delta_{\text{threshold}}$). Of course, for a pre-determined choice of $\Delta_{\text{threshold}}$ (corresponding to a predetermined confidence level), a better dataset
gives a smaller fraction of models which have indeterminate results. A universal value of the threshold \( \Delta_{\text{threshold}} = 5 \) without any regard to the properties of the models under comparison, as a rule of thumb, has been recommended by Liddle [48] as the minimum AIC difference between two models needed to make a ‘strong’ assertion that one model is better than the other.
Chapter 5

Cosmological Model Selection

This chapter includes previous work done by the author. The paper is titled ‘The reliability of the AIC method in Cosmological Model Selection’ and has been accepted for publication by the journal ‘Monthly Notices of the Royal Astronomical Society’. The paper is coauthored with Dr. Rahul Biswas.

5.1 Introduction

To illustrate the problems associated with using AIC, we need to introduce a practical example for which it can be applied. In this chapter, we will give some background information about the selection of late time expansion cosmological models from Type Ia SN data.

The late time acceleration of expansion of the universe has been firmly established [7, 14, 33, 35, 40, 44, 45, 56, 61, 63, 67, 79, 84], but there is no consensus on the physics behind this phenomenon. A number of possible explanations such as a small positive cosmological constant or vacuum energy, an otherwise unobserved dynamical field usually called dark energy [34, 60, 83, 86], or a modification of General Relativity [18, 26, 28] have been proposed as an explanation. With many models still consistent with the current data, it is clear that further progress of the field requires the collection of larger and complementary datasets and a definite framework for model selection. Several large new surveys such as DES\(^1\), BIGBOSS \(^2\) LSST \(^3\) EUCLID\(^4\) have been planned to study this late time acceleration by collecting more data [1, 53, 66]. Of course, even with accumulation of more quality data, the importance of analyzing the model selection process will not diminish, because reliable discriminating methods can always allow us to exploit the available data maximally. Hence, the refinement of statistical techniques would be particularly timely.

Statistical techniques addressing model selection have been applied to this context [10, 11, 12, 25, 47, 48, 49, 50, 51, 72, 76, 77, 81, 82]. Many of these discuss the use of information criteria like AIC and BIC which

---

\(^1\)http://www.darkenergysurvey.org/
\(^2\)http://bigboss.lbl.gov/
\(^3\)http://www.lsst.org/lsst/
\(^4\)http://sci.esa.int/euclid
Table 5.1: Different model classes considered in this thesis: We show the evolution of the Hubble function $H(z)$ with redshift $z$, the Hubble constant $H_0$ and other free parameters in the models. $k$ is the number of free parameters. Note that these models assume a flat universe.

<table>
<thead>
<tr>
<th>Model</th>
<th>$H(z)/H_0$</th>
<th>Free Parameters</th>
<th>$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda$CDM</td>
<td>$\sqrt{\Omega_m(1+z)^3 + (1-\Omega_m)(1+z)^3}$</td>
<td>$\Omega_m$</td>
<td>1</td>
</tr>
<tr>
<td>wCDM</td>
<td>$\sqrt{\Omega_m(1+z)^3 + (1-\Omega_m)(1+z)^3(1+w_0)}$</td>
<td>$\Omega_m, w_0$</td>
<td>2</td>
</tr>
<tr>
<td>CPL</td>
<td>$\sqrt{\Omega_m(1+z)^3 + (1-\Omega_m)(1+z)^3(1+w_0 + w_a + 1) \exp\left(-\frac{3w_a+2}{1+w_a}\right)}$</td>
<td>$\Omega_m, w_0, w_a$</td>
<td>3</td>
</tr>
<tr>
<td>DGP</td>
<td>$\sqrt{\Omega_m(1+z)^3 + \Omega_{rc} + \sqrt{\Omega_{rc}}}$</td>
<td>$\Omega_m$</td>
<td>1</td>
</tr>
</tbody>
</table>

Three of these models, $\Lambda$CDM, wCDM and CPL [23, 52], are dark energy models with different parametrization of the equations of state $w(z) = p(z)/\rho(z)$, where $p(z)$ and $\rho(z)$ are the pressure and density of dark energy, respectively. These models are nested: setting $w_a = 0$ in the CPL model, we obtain the wCDM model; setting $w_0 = -1$ in the latter gives the $\Lambda$CDM model. We also use the flat DGP model [28] which is a modified gravity model and cannot be nested in the previous classes of models.

\[5\] We have included $\Lambda$CDM in our set of models considered. This is currently considered the most viable model to describe the data.
Figure 5.1: Spectra of the major types of supernovae. Line (a) represents a Type Ia event. Line (b) represents a Type II event. Line (c) represents a Type Ic event. Line (d) represents a Type Ib event. Figure is taken from [32] and reproduced with the permission of Prof. Alexei Filippenko (author).

5.3 Type Ia SN data

5.3.1 Background

A supernova (SN) is an intensively luminous stellar explosion. Depending on the spectra observed, they are categorized into different types (Fig. 5.1). The different types are Type Ia, Type Ib, Type Ic and Type II. The data used is derived from Type Ia SN Events (or SNIa). A Type Ia event is identified by its unique Silicon II absorption feature at 6150 Å, near peak light. It is also distinguishable from Type II SN due to its lack of hydrogen lines, a feature present in Type II SN. For more information, we refer the interested reader to [32].
Figure 5.2: Type Ia SN Data. 371 data points were used from the Constitution Compilation [40]. The vertical axis indicates the distance moduli while the horizontal axis indicates the redshift. Vertical error bars are given for the distance moduli. Horizontal error bars are negligible and not indicated.
5.3.2 Obtaining observational data

SNIa events are characterized by fairly consistent luminosity, making them a good choice to serve as standard candles. The directly observed measure of brightness is known as the apparent magnitude \( m \). The distance moduli \( \mu \) is derived by comparing \( m \) to the absolute magnitude \( M \), which is the apparent magnitude for a similar event 10pc away. This relation is given by \( \mu = m - M \). \( \mu \) is important because it is related to the luminosity distance \( d_L \) by \( \mu = 5 \log_{10} \frac{d_L}{10pc} \), where \( d_L \) is a measure of distance between the SN event and the observer. Another important directly observed quantity is the red-shift \( z \). This measures the amount of expansion of space-time between the observed event and the observer. For this paper, we use data from the Constitution compilation [40] of the CFA3 sample [39] ESSENCE [54], SNLS [7] and ‘High-z’ samples [62].

5.3.3 Relating data to the model

Different late-time expansion cosmological models relate \( d_L \) to \( z \) by the equation [27]:

\[
d_L(z) = (1 + z) \frac{c}{H_0} \sqrt{|\Omega_k|} \left[ H_0 \sqrt{|\Omega_k|} \int_0^z \frac{dz'}{H(z')} \right],
\]

(5.1)

where \( \Omega_k = -\frac{k}{H_0^2} \). \( \Omega_k \) is a term that describes the space-time curvature of the model, where \( k > 0 \) describes a positive curvature, \( k < 0 \) describes a negative curvature and \( k = 0 \) describes a flat model. \( F(x) \) is defined by the following relations:

\[
F(x) = \sinh(x) \quad \text{for } k < 0,
\]

(5.2)

\[
F(x) = x \quad \text{for } k = 0,
\]

(5.3)

\[
F(x) = \sin(x) \quad \text{for } k > 0.
\]

(5.4)

In the equations above (5.1)-(5.4), \( c \) is the speed of light, \( H_0 \) is the Hubble constant and \( H(z) \) is the Hubble function that depends on the particular cosmological model used.

We fit the theoretical quantity of the distance moduli \( \mu(z_i) \) (\( z_i \) is the \( i \)th observed red-shift) against its observed value \( \mu_{\text{obs}} = m_i - M_i \), where \( m_i \) is the observed apparent magnitude and \( M_i \) is the absolute magnitude of the SN data. Note that the index \( i \) indicates the \( i \)th data point. The empirical plot is given in Fig. 5.2.
5.4 Getting a $\chi^2_{ML}$ fit from SNIa

We will assume that the universe is spatially flat, by setting the curvature term $\Omega_k$ to zero, and under this assumption

$$d_L(z) = (1 + z)c \int_0^z \frac{dz'}{H(z')},$$

(5.5)

where $c$ is the speed of light.

We start with the assumption that $\mu^{\text{obs}}_i$ has a Gaussian noise structure. We model it as

$$\frac{1}{\Pi_i \sqrt{2\pi\sigma^2_i}} \exp \left[-\sum_i \left( \frac{\mu^{\text{obs}}_i - \mu(z_i)}{2\sigma^2_i} \right)^2 \right],$$

(5.6)

where $\sigma_i$ are consistent with the error bars associated with $\mu^{\text{obs}}_i$ in the Constitution compilation [40]. Note that we ignored the uncertainty in $z_i$ since the error bars on $z_i$ are very small for spectroscopically determined redshifts. $\mu^{\text{obs}}$ was calculated from the apparent (observed) magnitude $m$ by assuming a fixed absolute magnitude $M$ value, which we are actually unsure about. We get around this problem by introducing a nuisance parameter $g$ and integrate it over a flat prior (Gaussian prior where the standard deviation $\to \infty$). To do this, we first integrate this over a Gaussian distribution of the nuisance parameter with standard deviation $\sigma^2_g$ to get

$$\int_{-\infty}^{\infty} \left( \frac{1}{\Pi_i \sqrt{2\pi\sigma^2_i}} \right) \exp \left[-\sum_i \left( \frac{\mu^{\text{obs}}_i - \mu(z_i) - g}{2\sigma^2_i} \right)^2 \right] \frac{1}{\sqrt{2\pi\sigma^2_g}} \exp \left[-\frac{g^2}{2\sigma^2_g} \right] dg.$$

We can rewrite this in matrix form:

$$\int_{-\infty}^{\infty} \left( \frac{1}{\Pi_i \sqrt{2\pi\sigma^2_i}} \right) \exp \left[-\frac{(X - gY)^T \Lambda(X - gY)}{2} \right] \frac{1}{\sqrt{2\pi\sigma^2_g}} \exp \left[-\frac{g^2}{2\sigma^2_g} \right] dg,$$

where $X$ is a vector with $n$ components, whose $i$th component is $\mu^{\text{obs}}_i - \mu(z_i)$. $Y$ is a vector with $n$ components where all the elements are ‘1s’ and $\Lambda$ is the inverse of the covariance matrix (which in this case is diagonal). The $^T$ symbol denotes the transposition. Performing the $g$ integral and setting $\sigma_g$ to a large value, the following marginalized function is obtained:

$$\frac{1}{\sqrt{\sigma^2_g Y^T \Lambda Y}} \left( \frac{1}{\Pi_i \sqrt{2\pi\sigma^2_i}} \right) \exp \left[-\frac{1}{2} X^T \left( \Lambda - \frac{\Lambda Y Y^T \Lambda}{Y^T \Lambda Y} \right) X \right].$$

(5.7)
5.5 Deriving the AIC results

This reduces the log likelihood to

\[ \frac{X^T C X}{2} - \frac{1}{2} \log(\sigma_g^2 Y^T \Lambda Y) - \sum_i \frac{1}{2} \log(2\pi \sigma_i^2), \tag{5.8} \]

where

\[ C = \Lambda - \frac{\Lambda Y Y^T \Lambda}{Y^T \Lambda Y}. \tag{5.9} \]

The second term suffers from a log divergence as \( \sigma_g \to \infty \). However, since AIC works by comparing relative log likelihood values, we can regularize this term away by setting it to zero. We can also ignore the third term since it is a fixed constant independent of the parameter choice. Thus, the only important term to consider is the minimized \( X^T C X \). Because of the marginalization against a flat prior, the rank of \( C \) is smaller than the rank of \( \Lambda \) by one and thus \( C \) cannot be inverted. The marginalization procedure also implies that the choice of the Hubble parameter \( H_0 \) and even the speed of light \( c \) is irrelevant to finding the maximum likelihood values of the other parameters.

This leads to the following relative AIC term:

\[ \text{AIC} = X^T(\hat{\theta}) C X(\hat{\theta}) + 2k, \tag{5.10} \]

where \( \hat{\theta} \) is the set of parameters that minimizes \( X^T C X \).

The first term corresponds to the modified maximum likelihood while the second term is the bias correction which is dependent on the number of free parameters. The maximum likelihood parameters can then be found using some standard minimization procedure (Appendix C). Specifically, these were found using the Gauss-Newton algorithm [13]. This allows us to calculate the AIC values for four candidate models. It should be noted that unlike other works [36, 76] that used AIC as a model selection tool, as described in the above, we marginalized away the \( H_0 \) term against a flat prior which reduces the number of free parameters by one. This technical difference alone should not affect our use of AIC. The data used consists of 371 SN events taken from the Constitution compilation (MLCS table) [40]. We computed the AIC values (Table. 5.2) for the different models and found that the DGP model has the smallest AIC value among the four models we considered in this thesis.

For the nested models, the parameter values that fit the data best turn out to be close to the \( \Lambda \)CDM model. Consequently, one does not gain much in terms of a lower \( \chi^2_{ML} \), while the extra free parameters are
Table 5.2: Different flat model classes considered in this thesis: We indicate the free parameters obtained from the Constitution compilation [40] using the maximum likelihood estimation. \( k \) is the number of free parameters. The respective AIC values were also obtained from the Constitution compilation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Free Parameters</th>
<th>( k )</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΛCDM</td>
<td>( \Omega_m = 0.325886 )</td>
<td>1</td>
<td>401.35</td>
</tr>
<tr>
<td>wCDM</td>
<td>( \Omega_m = 0.262752, w_0 = -0.84532 )</td>
<td>2</td>
<td>403.05</td>
</tr>
<tr>
<td>CPL</td>
<td>( \Omega_m = 0.397131, w_0 = -0.801712, w_a = -2.3239 )</td>
<td>3</td>
<td>404.66</td>
</tr>
<tr>
<td>DGP</td>
<td>( \Omega_m = 0.218302 )</td>
<td>1</td>
<td>401.13</td>
</tr>
</tbody>
</table>

penalized to give higher values of AIC for wCDM and CPL. The DGP model gives the best AIC value, which is only slightly better than the ΛCDM model. It is known that the simultaneous use of Cosmic Microwave Background (CMB) data from WMAP [25, 72] and Large Scale Structure (LSS) data [53] disfavors the DGP model compared to ΛCDM, since the parameter subspaces that provide the best fits for CMB, LSS and SNIa data do not overlap as much as in the ΛCDM model. We checked that our results are consistent with this, but will not discuss the CMB and LSS data to focus on methodology.
Chapter 6

Finite Datasets

This chapter includes previous work done by the author. The paper is titled ‘The reliability of the AIC method in Cosmological Model Selection’ and has been accepted for publication by the journal ‘Monthly Notices of the Royal Astronomical Society’. The paper is coauthored with Dr. Rahul Biswas.

6.1 Impact of AIC uncertainties in finite datasets

In the preceding chapter, we have discussed the AIC differences $\Delta_{A,B}$ without any regard for the fact that AIC is a statistical estimate. The associated uncertainty in the AIC estimate may not be negligible and may be dependent on the realization of noise in a particular dataset. Thus, there must be a statistical uncertainty in the value of $\Delta_{A,B}$\footnote{This is due to a statistical uncertainty in the AIC values coming from the variation of $\chi^2_{ML}$. However, the uncertainty in the AIC values of the models can be correlated, and turns out to be larger than the uncertainty in the AIC differences.} even when estimated from a dataset of similar quality.\footnote{In this paper, two SN datasets are said to have the same quality when they have the same number of data points, the same set of redshift $z$ values and the same set of error bars (standard deviation) that corresponds to the set of $z$ values.} The ensemble of such observations defines an empirical probability distribution of $\Delta_{A,B}$, and the particular value of $\Delta_{A,B}$ obtained from the current SN datasets is actually a sample value drawn from this probability distribution.

Ideally, we should be able to study the probability density distribution of $\Delta_{A,B}$ under repeated observations of results with sample size $n$: $P(\Delta_{A,B}|E_n)$, where $E_n$ denotes a collection of observation data which individually consists of $n$ observation points and are drawn from the underlying truth process. Because producing a large subset of $E_n$ is impossible, in this paper we instead use a bootstrap approach \cite{31} to generate ‘mock’ empirical datasets and estimate the probability distribution of $\Delta_{A,B}$.

6.2 Bootstrap

The bootstrap technique \cite{31} is a computer based method that was developed for estimating the accuracy of sample estimates (e.g. mean, variance). At the core of the technique is the use of the empirical data
to simulate the underlying probability distribution. It is very useful when the analytic expressions of the confidence intervals are too difficult or impossible to obtain.

### 6.2.1 Non-parametric bootstrap

A common example of this process is the non-parametric bootstrap. This is based on the plug-in principle [31] where the underlying probability distribution is simulated by resampling the empirical data with replacement. The assumption underlying this process is that each data point in the observed dataset is independent and identically distributed, and that the empirical distribution of the data is asymptotically close to the underlying probability distribution. Hence this underlying distribution can be simulated by resampling techniques. Starting with the original dataset, which we define as a set of \( n \) data pairs \( \{(x_i, y_i)\} \), we could resample with replacement this set of original data pairs \( n \) times to get a new surrogate dataset. If we repeat this process many times, we would get a distribution of datasets that reflects the underlying distribution of the original dataset. If we want to derive the parameter \( \theta \), we would ideally derive it from the underlying distribution \( F \) (i.e., \( \theta = t(F) \)), where \( t \) is some function to derive \( \theta \). However, \( F \) is unknown, so we use the empirical distribution \( \hat{F} \) to find its estimator instead \( \hat{\theta} \) (i.e., \( \hat{\theta} = t(\hat{F}) \)). This bootstrap technique also allows one to get confidence intervals for \( \hat{\theta} \).

### 6.2.2 Parametric bootstrap

In the preceding non-parametric bootstrap case, we made the assumption that each of the data pair \((x, y)\) follows the probability distribution \( Pr(X = x, Y = y) \). However for the case of regression models, we make the assumption that the data pairs are related by the conditional distribution \( Pr(Y = y|X = x) \). For example, if I have a data pair \((x, y)\) that has the form: \( y = f(x) + \epsilon \) where \( \epsilon \sim N(0, \sigma) \) and \( f(x) \) is a parametric curve that gives the underlying structure of the data, I can model the dataset with the conditional distribution: \( Pr(Y = y|X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-f(x))^2}{2\sigma^2}\right] \), where the distribution of \( y \) is conditioned on the value \( x \).

The approach to simulating a regression dataset is to sample the residuals of the regression. This involves fitting the data to a parametric model and sampling the residuals about the parametric mean. Since we do not know the true relation \( f \), we cannot obtain the purely empirical distribution of the residuals.\(^3\) Thus,

---

\(^3\)Note that these errors which include light curve fitting errors, intrinsic dispersion and peculiar velocity corrections are also required for calculating quantities like \( \chi^2 \) for most model selection schemes. Thus, an underlying assumption of the application of the AIC technique as in previous works is that these error estimates are correct. Since our focus is on the statistical uncertainties in AIC after following other underlying assumptions used in the literature, we also assume that these error estimates are correct.

---

37
for regression models, the standard bootstrap method always involves using a model-dependent probability distribution of residuals, which we refer to as the parametric bootstrap.\(^4\)

In order to proceed with the parametric bootstrap, we have to make certain assumptions first about the parametric relationship between the data input (explanatory variable) and the data output (response variable). We start with a set of \(n\) actual data points \(D = \{(x_1, y_1, \sigma_1), (x_2, y_2, \sigma_2), \ldots, (x_n, y_n, \sigma_n)\}\), where \(y_i\) is the response variable that is observed with the error bar \(\sigma_i\) and \(x_i\) the explanatory (input) variable. To produce a bootstrap sample set \(C_n\) consisting of \(n\) data points, we use a particular model \(C\): \(y = f(x)\) with the needed parameters chosen by maximum likelihood estimation. This model \(C\) is the parametric mean in which the residuals are distributed as above.

In the usual parametric bootstrap approach, the obtained set \(\{y_i - f(x_i)\}\) is regarded as the estimate of the distribution of residuals (noise distribution), but in our case, unfortunately, the variance of the residuals seem to depend on \(x\). We inferred this from the varying error bars in our data.

Therefore, in order to simulate the dataset, we make an additional assumption about the distribution of the residuals: they follow a Gaussian distribution centered about \(f(x)\) with a standard deviation equivalent to the corresponding error bar in the data.

### 6.3 Simulating data

We can apply this parametric bootstrap approach to the SNIa data. More specifically, the distance modulus \(\mu\) is related to the red-shift \(z\), so the regression model has the following structure \(\mu = f(z) + \epsilon\), where \(\epsilon\) is the residual (the error term). Davis et al. [25] used this method to compare between two regression models and find the standard deviation of their BIC differences; we will further extend this idea and show that studying the structure of the distribution of AIC differences is important.

When we wish to check the reliability of a particular model, we could use the probability distribution of residuals based on the model itself. Since we cannot have any model-free bootstrap data, we must choose a particular model to produce the residual distribution. We define this particular model as the reference model and treat it as the ‘truth’. Therefore, to estimate \(\Delta_{A,B}\), we need some model \(C\) as a reference probability model that is used to generate the parametric bootstrap data. Let us denote the estimate of \(\Delta_{A,B}\) based

---

\(^4\)Efron and Tibshirani [31] actually mentions two kinds of bootstraps for regression modeling. The first is similar to the non-parametric bootstrap and involves resampling the data pairs with replacement. The second, which we mentioned, involves sampling the residuals. Despite their declaration of the former as a form of bootstrap, we state that it is not strictly speaking a variant of bootstrap as it does not correspond to the plug-in principle.
on a dataset $d$ as $\Delta_{A,B}^d$. We wish to produce $\{\Delta_{A,B}^d | d \in C_n\}$, where $C_n$ denotes a collection of parametric bootstrap data generated by model $C$, which individually consists of $n$ observation points with the same data quality as our empirical data. Our approach may be outlined as follows.

Suppose we have a single dataset consisting of $n$ observed results $\{(z_1, \mu_1, \sigma_1), (z_2, \mu_2, \sigma_2), \cdots, (z_n, \mu_n, \sigma_n)\}$, where the $z_i$ values denote the $i$th observed redshift, $\mu_i$ the $i$th generated distance modulus, and $\sigma_i$ the observed error bars of $\mu_i$. We can create a bootstrap sample $C_n$ consisting of $n$ observation points based on model $C$. $C_n$ relates the same set of coordinates by the relation: $\mu_i = f(z_i) + \epsilon_i$, $i = 1, \cdots, n$, where $\epsilon_i$ is a stochastic term obeying a normal distribution of mean 0 and standard deviation $\sigma_i^2$: $N(0, \sigma_i^2)$, and $f(z)$ is the MLE of model $C$ ($\Lambda$CDM, DGP, etc.).

As mentioned above, we wish to simulate the distribution of $\{\Delta_{A,B}^d | d \in C_n\}$ as a proxy for $P(\Delta_{A,B}|E_n)$. To proceed, we choose a subset of models that have been often studied in the literature [76]. To make contact with observational data, we choose candidate models with parameter sets that are ‘best’ for the Constitution compilation [40] of SNIa data ($f(z)$ being the MLE-derived model), where we use the distance moduli and the error bars in data (Section 5.4). Since we need to find estimates of cosmological parameters for different models by maximizing likelihoods, we use the results from the more appropriate MLCS light curve fitter (for $R_V = 1.7$) in [40]. The details of finding the maximum likelihood and the corresponding AIC value is given in Section 5.4; 371 SNIa events were used.

In order to calculate the distribution of $\Delta_{A,B}^d$, we adopt the best fit model in a model class $C$ as a reference model to produce 5000 mock datasets of $C_{371}$ which is expected to be similar in quality to the Constitution compilation [40] of SN data. All the model classes in Table. 5.1 are successively chosen as the reference model $C$. Following our definition of ‘similar quality’, each simulated dataset has the same set of redshift values and error bars as the Constitution compilation [40], while their apparent magnitudes are those expected from a noisy realization of the reference model. The basic steps are:

1. We produce a mock dataset consisting of 5000 realizations of $d \in C_{371}$ for a reference model $C$ as outlined above.

2. Candidate models $A$ and $B$ are fitted to $d \in C_{371}$ by maximizing the likelihood and the AIC values of these models $A$ and $B$ are computed through (3.11).

3. Thus, for each element $d \in C_{371}$, we can make the AIC difference $\Delta_{A,B}^d$. We study $\{\Delta_{A,B}^d | d \in C_{371}\}$ for the reference model $C$ by plotting a histogram.

$^5$The error found in $z_i$ is not considered in the $\chi^2$ fit since the error bars on $z_i$ are very small for spectroscopically determined redshifts.
We should note that the probability distribution of $\Delta^d_{A,B}$ is due to errors introduced by the stochastic noise term, but ignores the effect of the uncertainties in the cosmological parameters of the reference model $C$. 
This chapter includes previous work done by the author. The paper is titled ‘The reliability of the AIC method in Cosmological Model Selection’ and has been accepted for publication by the journal ‘Monthly Notices of the Royal Astronomical Society’. The paper is coauthored with Dr. Rahul Biswas.

7.1 Introduction

Consistency is defined as follows: If the list of candidate models contains the reference model, and the probability of the model selection method picking the right model approaches 1 almost surely in the asymptotic limit, then we call such a method consistent. It has been shown that AIC is not an asymptotically consistent technique [69]. Despite that fact, AIC could still be a useful technique if it gives the correct answer a high percentage of the time. Thus, even if the technique does not return the true model 100% of the time, it would still be helpful to study the proportion of cases when it does. In this chapter, we use the bootstrap simulation of data to derive the proportion of such cases by simulating the reference model. We would thus investigate the related idea of statistical self-consistency. Bootstrap simulations are also used to estimate the error bounds of AIC differences.

7.2 Statistical self-consistency

The issue of statistical self-consistency is considered in the following sense: When the reference model is $C$, how often does the bootstrap AIC method outlined above choose model $C$ as a better model than the rest? The distribution of the values of $\Delta^d_{C,A}$ for $d \in C_n$ can tell us about the statistical self-consistency. We start with the case when $\Delta_{\text{threshold}} = 0$. If $P(A/C) \equiv P(\{d \in C_n | \Delta^d_{C,A} < 0\})$ is larger than some predefined proportion, which must not be less than $1/2$, we may say model $C$ is better than $A$ when $C$ is the reference.
Table 7.1: Percentage of cases where the AIC method with a threshold of 0 selects the correct model over other candidate models considered. We defined the difference and percentage using the following convention. If we define the bootstrap data as being produced by the reference model X (rows) and we are comparing it against model Y (columns), the difference is defined as the AIC value of X minus the AIC value of Y and is denoted by the symbol $\Delta^d_{X,Y}$. The table counts the percentage of negative $\Delta^d_{X,Y}$, $d \in X_n$. Note that a value greater than 50% indicates that the correct model is chosen a majority of the time. We use this cut-off of 50% or more as a simple definition for statistical self-consistency.

The results for this case is summarized in Table 7.1. The P-values\(^1\) for the table are extremely small.

From Table. 7.1, we obtained $\Delta^d_{DGP,CPL} < 0$, 89% of the time (i.e., $P(CPL/DGP) = 0.892$), when DGP model is the reference, and 84% of the time (i.e., $1 - P(DGP/CPL) = 0.836$), when CPL model is the reference. This means that, when we use $\Delta_{threshold} = 0$, the DGP model is significantly favored over the CPL model even if the CPL model is the reference. This means that the AIC method is statistically inconsistent when it is applied to the CPL model, and would be automatically disqualified under the current number of data points in the sample. In another example, we compared the DGP and $\Lambda$CDM models in the table and noticed $\Delta^d_{\Lambda CDM,DGP} < 0$, 59% of the time when $\Lambda$CDM model is the reference (i.e., $P(DGP/\Lambda CDM) = 0.594$), and 42% of the time when DGP model is the reference (i.e., $1 - P(\Lambda CDM/DGP) = 0.422$). This means that if the reference model was either $\Lambda$CDM or DGP and we apply AIC to compare between them using a zero threshold, AIC will only slightly favor the reference. The AIC technique is only statistically self-consistent (for a zero threshold) when applied in a comparison between DGP and $\Lambda$CDM, while we cannot use AIC to self-consistently study the other models under the current level of observation quality. However, a test that gives the right answer 3 out of 5 times is unreliable, since we can only do a single empirical test from our actual data. $\Lambda$CDM and DGP cannot be distinguished significantly using either reference models.

Thus, looking at both examples, we must conclude that there are insufficient data points to tell reliably the models apart using AIC when $\Delta_{threshold} = 0$. Another trend that results from the insufficiency of data

\(^1\)The highest P-value found was $3 \times 10^{-8}$ for $\Delta_{DGP,\Lambda CDM}$. Most of the P-values found were at the level of machine precision.
points is that the AIC procedure tends to favor models with a smaller number of free parameters. The trend persists even when we later increase $\Delta_{\text{threshold}}$. This seems to indicate that the addition of extra free parameters does not significantly improve the $\chi^2$ fit for the number of data points used.

### 7.3 Numerical study of the threshold

We next study the behavior when the threshold parameter is increased, $\Delta_{\text{threshold}} = 2$ and $5$, corresponding to choices made in the literature with moderate and strong evidence respectively. Unlike the previous $\Delta_{\text{threshold}} = 0$ case, we have to consider the effect of indeterminate cases, which we defined in Section 4.5 as the case when $|\Delta_{A,B}| < \Delta_{\text{threshold}}$, and setting $A$ to be the reference model.

In order to study the reliability of the AIC technique at different values of the threshold parameter $\Delta_{\text{threshold}} > 0$, we analyze the probability of the selected model being incorrect for different values of $\Delta_{\text{threshold}}$. To do so, we set $A$ to be the reference model and define the following:

\[ f_{\text{ind}} = \frac{\text{Number of samples with } |\Delta_{A,B}| < \Delta_{\text{threshold}}}{\text{Number of samples}} \]  
\[ f_{\text{all false}} = \frac{\text{Number of samples with } \Delta_{A,B} > \Delta_{\text{threshold}}}{\text{Number of samples}} \]  
\[ f_{\text{det false}} = \frac{\text{Number of samples with } \Delta_{A,B} > \Delta_{\text{threshold}}}{\text{Number of samples} \times (1 - f_{\text{ind}})} \]

$f_{\text{ind}}$ is the fraction of cases where the AIC procedure has an indeterminate result for a given value of the threshold parameter $\Delta_{\text{threshold}}$, so a high value of $f_{\text{ind}}$ reflects the inadequacy of AIC and the data to discriminate between the pair of models in question with a certain level of confidence for a relevant $\Delta_{\text{threshold}}$.

Using $A$ as the reference model, $f_{\text{all false}}$ is the fraction among all cases, where the AIC procedure results in an incorrect model selection. $f_{\text{det false}}$ is the fraction among determinate cases ($|\Delta_{A,B}| > \Delta_{\text{threshold}}$) where the AIC procedure results in an incorrect selection, and indirectly reflects the ratio of correct to incorrect model selections. Our results are summarized in Tables 7.2 and 7.3. In each table, the rows correspond to different reference models $X$, while the columns list candidate models $Y$ that were compared with the reference model.

For each pair of reference model $X$ and candidate model $Y$, we show the fractions $(f_{\text{ind}}, f_{\text{all false}}, f_{\text{det false}})$. First we reconsider the $\Delta_{\text{threshold}} = 0$ case in Table 7.1. By definition, the proportion of $\Delta_{\text{threshold}} = 0$ cases where the AIC method is not statistically self-consistent $(1 - P(A/B))$ is equivalent to both $f_{\text{all false}}$.
and $f_{false}^{det}$. We note that the values of $f_{false}^{det}$ are large, indicating a high failure rate and an unsatisfactory procedure.

We expect these failures to be suppressed when we choose higher values of the threshold $\Delta_{threshold}$. When we increase the threshold $\Delta_{threshold}$ to 2 and then to 5, we notice the expected suppression of $f_{false}^{det}$. However, $f_{false}^{det}$ does not decrease as dramatically as $f_{false}^{all}$.

We study the behavior of $f_{false}^{det}$, $f_{false}^{all}$ and $f_{ind}$ in Fig. 7.1 for different values of $\Delta_{threshold}$ as well as different choices of candidate and reference models; using bootstrap simulations. The $f_{false}^{det}$ values become dominated by noise as $f_{ind}$ increases, since the calculation is made from an ever decreasing number of determinate bootstrap cases. Hence, the values of $f_{false}^{det}$ at large values of $f_{ind}$ should be ignored. Nevertheless, we can still study the regime for smaller $f_{ind}$. It should also be noted that for the case of $\Delta_{threshold} = 5$, the proportion of indeterminate results was high (Table 7.3). For example, when CPL is the reference model, the proportion of $\Delta^d_{CPL,ΛCDM}$ and $\Delta^d_{CPL,DGP}$ between ±5 are both approximately 98%.

Needless to say, $f_{ind}$ increases asymptotically to one as we increase $\Delta_{threshold}$. Increasing $\Delta_{threshold}$ monotonically suppresses $f_{false}^{all}$, but not necessarily $f_{false}^{det}$. $f_{false}^{det}$ decreases with increasing $\Delta_{threshold}$ for the cases where the DGP model were wrongly picked over the reference ΛCDM model. However, for the other cases, when the candidate model has a smaller number of free parameters than the reference model, $f_{false}^{det}$ tends to gently increase before steeply decreasing at a certain value of $\Delta_{threshold}$. From Fig. 7.1, we can see that this sharp decline happens at roughly twice the difference in the number of free parameters between the candidate and reference models. We also study these quantities for the case where the reference model has less parameters than a candidate model in Fig. 7.2. In this case, $f_{false}^{det}$ actually increases rapidly at the point where $\Delta_{threshold}$ is equal to the difference in the number of free parameters in the two models. This implies that if the model underlying the empirical data was similar to the reference models studied, it is improbable that the data set would provide an AIC difference for the considered models greater than a large threshold $\Delta_{threshold}$ (eg., 5) as shown by $f_{false}^{det}$. However, if this dataset did yield an AIC difference larger than a predetermined $\Delta_{threshold}$, it does not necessarily mean that the AIC selected model has a high probability of being the true underlying model. This is because $f_{false}^{det}$ for a given $\Delta_{threshold}$ depends on the model pairs being considered, suggesting that even having AIC differences larger than a specified threshold $\Delta_{threshold}$ does not guarantee reliability of the AIC selection process.

In order to intuitively understand what leads to these examples, it is instructive to consider the shapes of the distribution of AIC differences when comparing the ΛCDM and CPL models or the ΛCDM and DGP models. We note that a comparison of nested models will always involve strongly asymmetric, exponential-
like distribution of AIC differences, while the comparison of non-nested models could result in almost symmetric distributions. This is because the $\chi^2_{ML}$ for the general case cannot be smaller than the more specific case. For example, $\Lambda$CDM is a special case of the more general $w$CDM and CPL models. Hence, when we fit these models against any data or simulated data (regardless of the reference model used to generate it), the $w$CDM and CPL $\chi^2_{ML}$ values cannot be larger than the $\Lambda$CDM $\chi^2_{ML}$ value and are usually smaller. This results in a sharp edge along a maximum value of $\Delta_{\Lambda CD M, CPL}$ and an exponential-like distribution of AIC differences between these models.

We show examples of such a Gaussian-like distribution in Fig. 7.3 and exponential-like distribution in Fig. 7.4. As an aside, we note that the different reference models considered do not make much difference in the shape of these distributions in the figures. Since $\text{AIC} = \chi^2_{ML} + 2k$, the edge of these exponential-like histograms is shifted by twice the difference between their number of parameters. Hence for the comparison between the CPL model and $\Lambda$CDM model in Fig. 7.1 and Fig. 7.2, where the difference between the $\chi^2_{ML}$ is almost zero, the distribution of AIC differences $\Delta_{\Lambda CD M, CPL}$ has a sharp edge at approximately $-4$. This implies that any AIC difference greater than a $\Delta_{\text{threshold}}$ value of approximately 4 will exclusively select the CPL model, irrespective of whether the reference model used was a CPL model or a $\Lambda$CDM model. At lower values of $\Delta_{\text{threshold}}$, the AIC selection procedure tends to select the model with the lower number of parameters since the $\chi^2_{ML}$ values are approximately the same. On the other hand, there is no similar constraint relating the $\chi^2_{ML}$ values of the $\Lambda$CDM and DGP (at the best discrepancy value); consequently the histogram turns out to be Gaussian-like. They have the same number of free parameters so this distribution is roughly centered about zero (the difference in $\chi^2_{ML}$ between these models). Obviously, if the quality of data was much better in terms of the error bars on each observation, or having a larger number of observations, the differences in the $\chi^2_{ML}$ terms would be much larger for the same choice of reference models used. In such cases, the model comparison would be purely data-driven, deriving its discriminatory power from the discrepancy term of AIC. This agrees with our intuitive idea that a better dataset should be able to resolve models better.

7.4 Error bars

The shapes of the distribution of AIC differences are also important when we study the spread of the statistical uncertainty of the distribution of the AIC differences since the statistical spread of the distribution cannot be specified unless we know the shapes beforehand. Due to the structural differences between the two kinds of distributions, we must define the ‘error’ bars according to the shape in order to make any useful
Table 7.2: Failure Rate for $\Delta_{threshold} = 2$: The $(f_{ind}, f_{false}^{all}, f_{false}^{det})$ values in the parentheses show $f_{ind}$, the percentage of total cases where the AIC procedure has an indeterminate result, $f_{false}^{all}$, the fraction of total cases where the AIC procedure results in an incorrect model selection; and $f_{false}^{det}$, the percentage of determinate cases where the AIC wrongly selects a candidate model $Y$ over the reference model $X$.

<table>
<thead>
<tr>
<th></th>
<th>DGP</th>
<th>$\Lambda$CDM</th>
<th>wCDM</th>
<th>CPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGP</td>
<td>$(97.1, 0.7, 25.5)%$</td>
<td>$(85.9, 4.6, 32.5)%$</td>
<td>$(25.0, 3.9, 5.3)%$</td>
<td></td>
</tr>
<tr>
<td>$\Lambda$CDM</td>
<td>$(97.2, 0.7, 24.1)%$</td>
<td>$-$</td>
<td>$(93.9, 4.0, 66.2)%$</td>
<td>$(24.2, 3.8, 5.0)%$</td>
</tr>
<tr>
<td>wCDM</td>
<td>$(86.8, 9.4, 71.3)%$</td>
<td>$(90.5, 1.0, 10.7)%$</td>
<td>$-$</td>
<td>$(96.0, 3.2, 79.0)%$</td>
</tr>
<tr>
<td>CPL</td>
<td>$(30.7, 63.7, 92.0)%$</td>
<td>$(32.0, 60.9, 89.6)%$</td>
<td>$(93.5, 0, 0)%$</td>
<td>$-$</td>
</tr>
</tbody>
</table>

Table 7.3: Failure Rate for $\Delta_{threshold} = 5$: The $(f_{ind}, f_{false}^{all}, f_{false}^{det})$ values in the parentheses show $f_{ind}$, the percentage of total cases where the AIC procedure has an indeterminate result, $f_{false}^{all}$, the fraction of total cases where the AIC procedure results in an incorrect model selection; and $f_{false}^{det}$, the percentage of determinate cases where the AIC wrongly selects a candidate model $Y$ over the reference model $X$. ‘-’ indicates an indeterminate result.

<table>
<thead>
<tr>
<th></th>
<th>DGP</th>
<th>$\Lambda$CDM</th>
<th>wCDM</th>
<th>CPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGP</td>
<td>$(100, 0, '-')%$</td>
<td>$(99.4, 0.6, 100)%$</td>
<td>$(99.2, 0.8, 100)%$</td>
<td></td>
</tr>
<tr>
<td>$\Lambda$CDM</td>
<td>$(100, 0, '-')%$</td>
<td>$-$</td>
<td>$(99.4, 0.6, 100)%$</td>
<td>$(99.4, 0.56, 100)%$</td>
</tr>
<tr>
<td>wCDM</td>
<td>$(99.5, 0, 0)%$</td>
<td>$(99.0, 0, 0)%$</td>
<td>$-$</td>
<td>$(99.3, 0.7, 0)%$</td>
</tr>
<tr>
<td>CPL</td>
<td>$(98.6, 0, 0)%$</td>
<td>$(98.1, 0, 0)%$</td>
<td>$(98.6, 0, 0)%$</td>
<td>$-$</td>
</tr>
</tbody>
</table>
Figure 7.1: Probabilities of different candidate models (Cand) being selected over different reference models (Ref) by using AIC for different values of $\Delta_{\text{threshold}}$ for the case of (a) (top left) the candidate model DGP being picked over the reference model $\Lambda_{\text{CDM}}$, (b) (top right) the candidate model $\Lambda_{\text{CDM}}$ being picked over the reference model $\text{wCDM}$, and (c) (bottom) the candidate model $\Lambda_{\text{CDM}}$ being picked over the reference model CPL. The solid blue curve shows the number of incorrect results as a fraction $f_{\text{false}}$ of the cases where the procedure returns a determinate result. The black dashed curve shows the number of incorrect results as a fraction of the total number of simulations. The fraction $f_{\text{false}}$ is extremely noisy and should be ignored when the fraction of indeterminate cases $f_{\text{ind}}$ (red, dotted) is large. The plots show that increasing $\Delta_{\text{threshold}}$ always decreases $f_{\text{false}}$, $f_{\text{false}}$ does not necessarily decrease. For a comparison of the models used, this shows that AIC tends to incorrectly select models with a lower number of parameters. In the comparison of nested models, $f_{\text{false}}$ drops sharply at twice the difference in the number of free parameters of the models compared.
Figure 7.2: Probability of the candidate model (Cand) CPL being selected over the reference model (Ref) ΛCDM for different values of the threshold. In this case, $f_{\text{false}}^{\text{det}}$ actually increases sharply at about twice the difference in free parameters between these models, showing that in this case AIC tends to incorrectly select the model with a higher number of parameters.

As an illustration, we look at statistical uncertainty of the distribution of $\Delta_{\Lambda_{\text{CDM}},\text{DGP}}^d$ for the two separate cases where the DGP and ΛCDM models are the references. The distribution of the differences in both cases are Gaussian-like. We notice that the standard deviation of $\Delta_{\Lambda_{\text{CDM}},\text{DGP}}^d$ is 0.89 with DGP model as the reference and 0.83 with ΛCDM model as the reference. The AIC difference between the DGP model and ΛCDM model $\Delta_{\Lambda_{\text{CDM}},\text{DGP}}$ in our original AIC analysis is 0.22 and smaller than the error bars. This result means that any subsequent analysis based on the value of $\Delta_{\Lambda_{\text{CDM}},\text{DGP}}$ observed is unreliable. The error bars could also be significant even if $\Delta_{\Lambda_{\text{CDM}},\text{DGP}}$ was larger than the error bars as we would have to modify any subsequent AIC difference analysis to include this uncertainty. When we look at $\Delta_{\Lambda_{\text{CDM}},\text{CPL}}^d$ (exponential-like distribution), the error bar region ranges from $-4.00$ to $-2.23$ for the case when the reference is the...
Figure 7.3: Probability distributions of the AIC differences between $\Lambda CDM$ and DGP ($\Delta_{\Lambda CDM,DGP}$) for different reference model X. X is used to generate the respective bootstrap samples and is written in the right upper corner of the figures. The horizontal axis indicates the $\Delta_{\Lambda CDM,DGP}$ value while the vertical axis indicates their relative frequency. If the process underlying our observations was really the best fit model of class X, then the values of the AIC differences under different realizations of noise would have the histogram distribution $\{\Delta_{\Lambda CDM,DGP} \mid d \in C_{371}\}$ shown. Vertical lines show the respective AIC differences that were derived from the observed data.
Figure 7.4: Probability distributions of the AIC differences between Λ CDM and CPL ($\Delta_{\Lambda CDM,CPL}^d$) for different reference model X. X is used to generate the respective bootstrap samples and is written in the right upper corner of the figures. The horizontal axis indicates the $\Delta_{\Lambda CDM,CPL}^d$ value while the vertical axis indicates their relative frequency. If the process underlying our observations was really the best fit model of class X, then the values of the AIC differences under different realizations of noise would have the histogram distribution $\{\Delta_{\Lambda CDM,CPL}^d | d \in C_{371}\}$ shown. The exponential-like distributions observed are due to the fact that ΛCDM and CPL are nested models. Vertical lines show the respective AIC differences that were derived from the observed data.
ACDM model and ranges from $-4.00$ to $-1.42$ when the reference is the CPL model. The $\Delta_{\Lambda\text{CDM},\text{CPL}}$ value calculated from the empirical data was found to be $-3.31$. However, the error bar region range of approximately 2 would make the value of $-3.31$ less certain. Instead of quantifying the odds ratio given by (4.4) as having a value of $0.19$, we now make a statement about its uncertainty by saying that the odds ratio can be a value between $0.14$ and $0.48$. These are just two examples in which one can carry out an analysis to determine the reliability of the model likelihood ratio $P(A)/P(B)$ that is calculated in (4.4).

We note that the statistical uncertainty of the AIC differences obtained above is smaller than the $\Delta_{\text{threshold}}$ value of 5 already mentioned above [25, 48, 72]. However, it is still significant enough and needs to be considered in our analysis.
Chapter 8

Further Criticism of AIC

8.1 Introduction

In the preceding chapters, we introduced AIC as

\[ AIC = -2 \log(L_{ML}) + 2k. \]

The first term in AIC favors a model that fits the data closely but a model that fits too closely tends to be more ‘complex’ (larger \( k \)), resulting in two competing terms. The popularity of AIC is due to the presence of these two competing terms. Prior to its discovery, there was no objective criterion to compare different models against data as the MLE approach favored overly complicated models. The \( 2k \) term seems to account for model complexity but this is actually an illusion.

8.2 \( 2k \) term as an inadequate measure of complexity

The \( 2k \) term alone seems to describe the complexity of the model by counting the number of free parameters. It is strictly speaking not a pure complexity term but a bias correction due to the finite size effect of the data. AIC, as a whole, is derived from complexity arguments. If we use a model \( q \) to describe the distribution \( p \) underlying the data, AIC is a direct asymptotic estimate of \( H[p, q] \), described in Section 2.3. If we associate this code length with the complexity of understanding a process generated by \( p \) using the wrong model \( q \), we can interpret the entire AIC criterion as a complexity quantity\(^1\) that needs to be minimized. The \( 2k \) penalty term only has meaning within this context as a bias correction and should not be misconstrued as a complexity term by itself.

In our derivation of the \( 2k \) term, we had assumed that the number of data points was large (asymptotic

\(^1\)This should not be confused with model complexity.
limit). It should be noted that, in this limit, its relative effect over the log likelihood term diminishes as well. In fact, due to the extensivity of $-2 \log(L_{\text{MLE}})$ and the non-extensivity of $2k$, the larger the number of data points, the smaller the finite size effect and the weaker the cost term becomes. In the limit of large number of data points, the bias term becomes negligible; causing AIC to breakdown and to behave like a pure MLE.

Furthermore, the $2k$ cost term only superficially mimics one aspect of complexity. As shown in (3.1), the statistical model that is typically used, consist of a product of independent probability distributions, where the number of independent distributions is equivalent to the number of data points. This means that the statistical model becomes more complicated as the number of data points increase. The $2k$ cost term does not account for this relationship.

Ignoring even the effect of the number of data points on the model complexity, the $2k$ term is still inadequate, because it can account only the number of free parameters as a ‘complexity’ measure. This is made worse by the fact that it treats all free parameters equally. The problem with this is that not all parameters are created equal. For example, this cost term does not make a distinction between the exponential curve $e^{ax+b}$ and the linear curve $ax+b$. Both models would correspond to a $k$ value of 2 since both have 2 free parameters each. Even worse, under this scheme, an approximation of an exponential curve via a polynomial leads to the absurd result that the approximation is more complex than the original function. Even when comparing between classes of polynomial, the AIC cost term cannot tell the difference between the curves $ax^2+bx$ and $ax+b$. A good cost scheme has to take into account the relative importance of the different free parameters and not just the mere number of terms.

### 8.3 Issues with the analytic derivation of the AIC cost term

AIC assumes the divergence between the proposed model and the true distribution is small. If the actual divergence is large, the Taylor expansion would blow up and it would be insufficient to just include the first order bias correction; in fact, one has to include an infinite number of corrections. The problem with an analytic correction is that its derivation is based on assumptions (such as a small divergence) that may not hold true.

---

We can treat the parameters the same if two conditions are satisfied. The first condition assumes that the Taylor expansion used in the AIC is valid. The second assumes that the number of data points is sufficiently large that we can make asymptotic assumptions. This may not necessarily be valid given the finite dataset used or the choice of candidate models.
8.4 Comparison of AIC to other model selection criterion

8.4.1 Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is “an asymptotic approximation to the transformation of the Bayesian posterior probability of a candidate model” [20]. In this asymptotic limit, the BIC favored model corresponds to the \textit{a posteriori} most probable model among the set of candidate models. This criterion has the following form:

\[
\text{BIC} = -2 \ln f(y|\hat{\theta}) + k \ln n, \tag{8.1}
\]

where \(n\) is the number of data points, \(f(y|\hat{\theta})\) denotes the likelihood evaluated at the set of MLE-derived model parameters \(\hat{\theta}\) and \(k\) is the number of free parameters in the model. This result is derived by assuming that the distribution underlying the set of data points is independent, identically distributed and belongs to the exponential family.

If we compare AIC (3.9) to BIC (8.1), we notice that the BIC cost term \(k \ln n\) starts to exceed the AIC cost term \(2k\) when \(n \geq 8\) (i.e., \(\ln 8 \approx 2.07944154\)). This means that BIC tends to favor smaller models compared to AIC. It also means that BIC seems to address (at least superficially) an earlier criticism we mentioned about AIC over its inability to account for the number of data points in its cost term.

In model selection, “an asymptotically efficient criterion will asymptotically select the fitted candidate model which minimizes the mean squared error of prediction” [20]. We also recall from Section 7.1 that a consistent criterion will asymptotically select the fitted candidate model having the correct structure with probability one. Taking into account both properties, AIC is asymptotically efficient yet not consistent while BIC is consistent yet not asymptotically efficient [20].

8.4.2 Minimum Description Length

Perhaps, the model selection criterion that is closest in principle to Occam’s Razor is the Minimum Description Length (MDL). It is based on the principle of choosing the model that gives the shortest description of the data and is related to the algorithmic complexity theories of Kolmogorov [46], Solomonoff [73, 74] and Chaitin [22]. It is similar to AIC in its basic philosophy of finding minimum code length descriptions. Unlike AIC, it does not assume the existence of some underlying distribution for the data and does not try to find the shortest code length to describe this estimated distribution. It tries to find the minimum code length description of the data instead of some assumed underlying distribution.
In one version of MDL [37], we model the data using a two-part code. The description (in bits) of the data using the model or hypothesis H can be heuristically represented by the formula

$$\text{Length} = L(\text{Data}|H) + L(H), \quad (8.2)$$

where $L(\text{Data}|H)$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis and $L(H)$ is the length, in bits, of the description of the hypothesis. The best model/hypothesis in describing the data is the model/hypothesis with the smallest length.

When this is applied to parametric modeling, and renaming the ‘Length’ as the ‘MDL’ criterion we get the following formula [64]:

$$\text{MDL} = -\ln f(y|\hat{\theta}) + \frac{k}{2} \ln \left( \frac{n}{2\pi} \right) + \ln \int d\theta \sqrt{\det I(\theta)}, \quad (8.3)$$

where $n$ is the number of data points, $k$ is the number of free parameters in the model, $\theta$ denotes the parameters of the model, $f(y|\hat{\theta})$ denotes the likelihood evaluated at the set of MLE-derived model parameters $\hat{\theta}$ and $I(\theta)$ is the Fisher Information matrix, defined as the expectation value $I_{ij}(\theta) = -E_{\theta} \left[ \frac{\partial^2 \log f(y|\theta)}{\partial \theta_i \partial \theta_j} \right]$.

It should be noted that when MDL was first derived, the higher order $\ln \int d\theta \sqrt{\det I(\theta)}$ was ignored, causing the earlier versions of the MDL formula to be superficially ‘equivalent’ to the BIC formula. This led to the mistaken belief that BIC and MDL were the same, even though they were derived from different principles.

For the interested reader, more information about MDL is available in [38].

### 8.4.3 Takeuchi Information Criterion

AIC is derived as an estimator of the expected KL discrepancy between the reference model and a fitted candidate model. As mentioned in Section 3.3, the asymptotic justification of the criterion requires two strong assumptions [19]:

1. “The reference model is contained in the candidate class under consideration.”
2. “The vector of MLE-derived estimators satisfies the conventional large-sample properties of MLE’s.”

Takeuchi Information Criterion (TIC) [78] is similar to AIC because it too tries to derive the expected KL divergence against the truth. However, it does so while relaxing assumption (1.). This results in the
criterion:
\[
\text{TIC} = -2 \ln f(y|\hat{\theta}) + 2 \left[ \text{trace} \left\{ J(\hat{\theta}) \left[ I(\hat{\theta}) \right]^{-1} \right\} \right],
\]
(8.4)
where \( I(\theta) \) is the Fisher matrix described in Section 8.4.2 and \( J(\theta) = \mathbb{E} \left\{ \frac{\partial \ln f(y|\theta)}{\partial \theta} \left\{ \frac{\partial \ln f(y|\theta)}{\partial \theta} \right\}^T \right\} \). When assumption (1.) is enforced, the cost term \( \left[ \text{trace} \left\{ J(\hat{\theta}) \left[ I(\hat{\theta}) \right]^{-1} \right\} \right] \to k \). Although TIC is a generalization of AIC, the difficulty in calculating the cost term made this approach a much less popular method.

8.4.4 Bootstrap estimate of AIC-like criteria

As shown with TIC, without any added assumptions, the analytical evaluation of the bias correction is hard to calculate. A method \([70]\) has been proposed to evaluate this term numerically via the bootstrap method. Generically, the bootstrap estimate takes the following form:

\[
\text{Criterion} = -2 \ln f(y|\hat{\theta}) + \text{Bias Term}.
\]
(8.5)

Many versions of the bootstrap bias term have been derived. Using the convention where \(*\) denotes a bootstrap sample, we have the following bootstrap evaluations:

\[
B_1 = \mathbb{E}_* \left[ \log \frac{f(y|\hat{\theta}(y^*))}{f(y^*|\hat{\theta}(y^*))} \right],
\]
(8.6)

\[
B_2 = 2\mathbb{E}_* \left[ \log \frac{f(y|\hat{\theta}(y^*))}{f(y^*|\hat{\theta}(y))} \right],
\]
(8.7)

\[
B_3 = 2\mathbb{E}_* \left[ \log \frac{f(y^*|\hat{\theta}(y))}{f(y^*|\hat{\theta}(y^*))} \right],
\]
(8.8)

\[
B_4 = 2\mathbb{E}_* \left[ \log \frac{f(y|\hat{\theta}(y^*))}{f(y^*|\hat{\theta}(y))} \right],
\]
(8.9)

\[
B_5 = 2\mathbb{E}_* \left[ \log \frac{f(y|\hat{\theta}(y))}{f(y^*|\hat{\theta}(y^*))} \right].
\]
(8.10)

The different forms were proposed independently from different sources.\(^3\) Shibata \([70]\) showed these

\[^3\text{Efron [29, 30] proposed }B_1, \text{ Cavanaugh and Shumway [21] proposed }B_2, \text{ and Shibata [70] proposed }B_3, B_4 \text{ and } B_5.\]
different forms to be asymptotically equivalent to one another. For the interested reader, more information about their derivation can be found in [21, 29, 30, 70]. What is interesting about the bootstrap estimate of the bias is that they seem to hint about a different kind of cost term. Just as the $2k$ cost term hints at the use of complexity as a cost term, but is not the same, the bootstrap estimate seems to hint at the use of stability as a cost term. This will be discussed in the next chapter.
Chapter 9

Stability as an Alternative Consideration to Complexity

A proposed consideration, as an alternative to complexity, would be the instability of the model. Stability has been proposed as a consideration in learning algorithm schemes [57, 59]. We will adapt this philosophy to parametric model selection by considering the instability of parametric models. Heuristically, we can define instability of a model class using the following procedure:

1. Find the MLE-derived model using the empirical data and model class. We will refer to this as the original model.
2. Perturb the data slightly. For example, we could do so by adding stochastic or bootstrap noise.
3. Find the MLE-derived model using the perturbed data and model class. We will refer to this as the perturbed model.
4. Quantify some mismatch between the perturbed and original models.

From (1.2), we can utilize this instability measure to define the heuristic form of a stability information criterion:

\[
\text{Criterion} = \text{Discrepancy} + \text{Instability},
\]

where the discrepancy term is identical to the discrepancy term in (1.2) and the instability term measures the model class’ susceptibility to data perturbations.

As shown in Section 1.1.1, an over-fitted model tends to fit not only the underlying structure of the data but its stochastic random noise as well. We can simulate the effects of random noise by perturbing the observed data and looking for patterns that are invariant against these perturbations. If the model class has a tendency to over-fit the noise structure, the MLE-derived models would be more susceptible to data perturbations. Thus, there would be a greater mismatch between the perturbed and original models. Due to the random nature of the noise, we could repeat the procedure many times and take a bootstrap average.
The issue of stability seems to be accounted for by $B_2$ (8.7). To see this, we reexpress the bias correction $B_2$ as $B_2 = 2E_* \left[ \log f \left( y|\hat{\theta}(y^*) \right) - \log f \left( y|\hat{\theta}(y) \right) \right]$. The more susceptible $f$ is to the bootstrap perturbation, the greater the difference between the two terms. This seems to indicate some kind of term that rewards stability.\footnote{The superficial similarity between bootstrap AIC and model stability was not mentioned by Efron, Cavanaugh or Shibata but noticed by the author of this thesis.} However such a term was not derived by stability considerations and only superficially mimics a term that rewards stability. The bias correction depends on the model’s sensitivity to bootstrap perturbations, but this sensitivity itself is not instability. Furthermore, this only superficially seems to consider only one type of data perturbation.

9.1 Non-bootstrap perturbation

To illustrate the case that the bootstrap bias correction appears to superficially model only a certain type of stability, we consider a type of data perturbation that is non-bootstrap.\footnote{Efron and Tibshirani [31] referred to this particular type of perturbation as bootstrap. However, we assert that this is not bootstrap as it does not correspond to the plug-in principle [31].} Fundamentally, the bootstrap principle involves simulating the empirical probability distribution using the empirical dataset. As mentioned in Section 3.2, for the case of regression models, this involves modeling the statistical distribution of the residuals. The form of stability suggested by the bootstrap bias correction involves a cost term that favors stability against such data perturbations.

However, if we sample the data with replacement, we do not always simulate the residual noise structure. As mentioned in (3.1) and (3.2), the model we used is the conditional probability density function $f \left( \{y_1, y_2, \ldots, y_n\} | \hat{\theta}, \{x_1, x_2, \ldots, x_n\} \right)$ where we assume some probability distribution for the residuals:

$$f \left( \{y_1, y_2, \ldots, y_n\} | \hat{\theta}, \{x_1, x_2, \ldots, x_n\} \right) = \prod_{i} f_i(y_i|\hat{\theta}, x_i) = \left( \prod_{i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \right) \exp \left[ -\sum_{i} \frac{(y_i - g(x_i))^2}{2\sigma_i^2} \right].$$

While the estimation of the probability distribution of residuals is important, the real focus in model selection is on the function $g(x)$ which contains the underlying structure of interest. The shape of this function should be stable against the choice of the conditional set of explanatory variables (i.e., $\{x_1, x_2, \ldots, x_n\}$). One way to see this is to create surrogate datasets by sampling with replacement and studying the instability of $g(x)$ with respect to this data perturbation. We should remind the reader that while this procedure is superficially similar to the bootstrap procedure for non regression models, it is \textit{not} a bootstrap procedure for regression models as we are not simulating the noise structure of the residuals.
For the sake of clarity, we will call a set of data points (used by a single regression model) a dataset and a set of dataset derived from any data perturbation procedure the perturbation set.

9.2 Quantifying model instability

The instability term has to quantify how easily or difficult the MLE chosen model (within a particular model class) changes under data perturbation. We will call the MLE chosen model that is derived from a perturbed dataset the perturbed model (PM) and the MLE chosen model that is derived from the original dataset the original model (OM). We can either perturb the data by sampling the original data points with replacement or sampling the estimated residuals (derived from the original data points) with replacement. For our thesis, we chose the former.

9.2.1 Using KL divergence

A natural language to quantify this difference is the KL divergence of PM from OM. Since there are many possible perturbed datasets that can be derived from this technique, we take a weighted average over all possible realizations of the generated perturbed datasets.

To simplify the derivation, we will assume the noise of the data is independently Gaussian distributed with identical variance:

\[
q^* \left( \{y_1, y_2, \ldots, y_n\} \mid \theta_q^*, \{x_{1.2}, \ldots, x_n\} \right) = \frac{1}{(2\pi\sigma^2_q)^{n/2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - g_q^*(x_i))^2}{\sigma^2_q} \right],
\]

\[
p \left( \{y_1, y_2, \ldots, y_n\} \mid \theta_p, \{x_{1.2}, \ldots, x_n\} \right) = \frac{1}{(2\pi\sigma^2_p)^{n/2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - g_p(x_i))^2}{\sigma^2_p} \right],
\]

where the \( p \) is the probability density of OM and \( q^* \) represents the probability density of a particular PM. \( \sigma_p \) and \( g_p \) indicates the standard deviation and model dependent mean of \( p \) respectively. \( \sigma^*_q \) and \( g^*_q \) indicates the standard deviation and model dependent mean of \( q^* \) respectively. We can derive a term \( \gamma_1 \) that describes the instability of the model by taking the Kullback-Leibler divergence and averaging it over many generated datasets \( E_* \).

\[
\gamma_1 = E_* \left[ \int d\vec{y} p \log \left( \frac{p}{q^*} \right) \right] = E_* \left[ n \log \frac{\sigma^*_p}{\sigma_p} + \frac{n(\sigma^2_p - \sigma^2_q)}{2\sigma^2_q} + \frac{1}{2\sigma^2_q} \sum_{i=1}^{n} (g_p(x_i) - g_q^*(x_i))^2 \right]
\]
Figure 9.1: $\gamma_1$ vs. model order. 50 data points were generated from the function $x^2 + 2x + 3 + w$, where $w \sim N(0, 1)$, at intervals of 0.1 from $x = -3.5$ to $x = 1.4$. We fitted several different order polynomial models to the data. For example, an order 2 model is a quadratic curve. We calculated the corresponding instability term $\gamma_1$ from (9.4) and plot the relationship.

where $\sigma$ and $g(x_i)$ represent the noise and mean of the model respectively.

It is understood that this quantity is not symmetric for large perturbations. But if we restrict ourselves to cases of small perturbation, $D(P \parallel Q) \approx D(Q \parallel P)$.

To highlight the contrast in behavior, against the AIC linear $k$ cost term, we study the behavior of $\gamma_1$ for different orders of polynomial fitting. Data was generated from the function $x^2 + 2x + 3 + \epsilon$, where $\epsilon \sim N(0, 1)$. From this function, 50 data points were generated at intervals of 0.1 from $x = -3.5$ to $x = 1.4$. This data was used to generate $\gamma_1$ for different polynomial model classes.

We first calculate $\gamma_1$ for the case for a zeroth order polynomial $f(x) = C$, where $C$ is a constant derived from the MLE of the data. We next calculate $\gamma_1$ for the case of a linear curve $f(x) = ax + b$, a quadratic curve and so on up till an order 9 polynomial. Using these values calculated, we plot the $\gamma_1$ values versus order to illustrate its behavior (Fig. 9.1).

We see that the instability cost term blows up quickly as one increases the order. This means that the cost term penalizes the higher order terms more strongly than the lower order terms. This makes sense since the higher order corrections over-fit the data more strongly. We do not want such a situation of over-fitting, so a higher order correction should be penalized more strongly.
9.2.2 Using an $l_p$-inspired Distance Measure

The problem with using a KL divergence is that it is not a distance measure. This brings up the question of whether we should use $D(q^* \parallel p)$, $D(p \parallel q^*)$ or even a symmetrized $\frac{1}{2}[D(p \parallel q^*) + D(q^* \parallel p)]$. Alternatively, we could use an $l_p$-inspired distance measure. For example, if we use an $l_2$-inspired distance measure, we can quantify the instability of the model class using the following procedure:

1. Generate the perturbation set using the original dataset.
2. Generate a set of functions (fitted curves) $g_i$ that is derived from the MLE fit to the different datasets in the perturbation set. The index $i$ corresponds to the $i$th dataset in the perturbation set.
3. Compute the $l_2$ related distance between two fitted curves $g_i$ and $g_j$, which we define as:
   \[ \delta_{ij} = \sum_{k=1}^{n} [g_i(x_k) - g_j(x_k)]^2, \]
   where $x_k$ belongs to the set of explanatory variables derived from the original dataset (i.e., \{x_1, x_2, \ldots, x_n\}).
4. Compute $\gamma_2$, which is the average of all these pairwise $\delta_{ij}$ using $\gamma_2 = \frac{1}{n \times n(n-1)} \sum_{i<j} \delta_{ij}$.

9.3 Demonstration of a stability information criterion (SIC)

Based on two proposed ways of quantifying model instability, we propose two different model selection criteria that account for model instability.

9.3.1 SIC based on KL divergences

The first step involves combining this instability cost term with a discrepancy term. Since, the instability term is based on the KL divergence, we base the discrepancy term on AIC. We want to minimize the information mismatch between the test model and the ‘truth’ up to some fixed offset. Defining $\eta = \text{RSS}/n$, where RSS is defined according to (1.1), we end up with the following information criterion (SIC):

\[ \text{SIC} = n \log \eta + 2k + \lambda \gamma_1, \]

where $\lambda$ describes the relative importance between $\gamma_1$ and AIC, which is described by $n \log \eta + 2k$.

---

3 Instead of computing $\left( \sum_{k=1}^{n} [g_i(x_k) - g_j(x_k)]^p \right)^{\frac{1}{p}}$ as our distance measure, we compute $\sum_{k=1}^{n} |g_i(x_k) - g_j(x_k)|^p$.

4 It should be noted that $g_i$ is the fitted curve and not a probability density function. It corresponds to function $g^*_q$ and $g^*_p$ in (9.2) and (9.3) respectively.
We test the usefulness of this criterion against AIC. We generate 1000 datasets. Each dataset is identically prepared from the quadratic curve $x^2 + 2x + 3 + w$ with $w \sim N(0, 1)$. For each dataset, 50 points were generated at intervals of 0.1 from $x = -3.5$ to $x = 1.4$.

We next apply AIC and SIC (with various values of $\lambda$) to the datasets and count the number of times the criterion picks the right model.

<table>
<thead>
<tr>
<th>Order</th>
<th>AIC</th>
<th>SIC $\lambda = 0.2$</th>
<th>SIC $\lambda = 0.5$</th>
<th>SIC $\lambda = 1$</th>
<th>SIC $\lambda = 2$</th>
<th>SIC $\lambda = 3$</th>
<th>SIC $\lambda = 10$</th>
<th>SIC $\lambda = 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>728</td>
<td>752</td>
<td>777</td>
<td>817</td>
<td>864</td>
<td>901</td>
<td>979</td>
<td>992</td>
</tr>
<tr>
<td>3</td>
<td>136</td>
<td>132</td>
<td>128</td>
<td>116</td>
<td>95</td>
<td>76</td>
<td>21</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
<td>76</td>
<td>65</td>
<td>48</td>
<td>30</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\geq 5$</td>
<td>49</td>
<td>40</td>
<td>30</td>
<td>19</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We see that SIC outperforms AIC for all cases.

### 9.3.2 SIC based on $l_2$ distance measure

In this version of SIC, the discrepancy term would be the lowest RSS fit for a given model class against the original dataset, where RSS is defined according to (1.1). We can thus define an SIC as:

$$
\text{SIC} = \text{RSS} + \lambda \gamma_2, \quad (9.6)
$$

where $\lambda$ describes the relative importance between $\gamma_2$ and the discrepancy RSS term.

In a test similar to the one performed above (9.3.1), we tested the same 1000 dataset on this criterion. Unlike the previous case, we tested it against AIC only for the case of $\lambda = 1$, counting the number of times the criterion picks the right model.
We see that the stability criteria also performs better against AIC for this example.

### 9.4 Summary

The purpose of this chapter is not to rigorously derive a new model selection criterion but to demonstrate the importance of stability as a consideration in model selection. We have used heuristic arguments to show why stability is an important consideration to guard against over-fitting in model selection. The issue of stability is related to the issue of complexity as more complicated models tend to be more unstable. Due to the influence of Occam’s Razor, complexity has been a popular consideration in model selection. We feel that this sole emphasis is misguided and hence model stability should also be considered.
Chapter 10

Conclusions

AIC has been widely used as a technique for model selection. Most commonly, this has been applied by computing the AIC values for each candidate model through (3.11) or (3.12), and selecting the model with the smallest AIC value as the best model. The issue of considering the magnitudes of AIC differences between the models to indicate the relative plausibility or confidence in the models has also been addressed by Akaike and elaborated by Burnham and Anderson through the use of Akaike weights (4.5). In the field of cosmology, AIC has been used in selecting models underlying the late time acceleration of the universe using data. There have also been suggestions of a rule of thumb that states an AIC difference of 5 or more would give strong evidence for the model with the smallest AIC value. This approximately corresponds to a ratio of model likelihoods of 12 or more.

10.1 Alternative interpretation of AIC difference

In this thesis, we propose a method for calculating the ratio of the model likelihood between models A and B based on their AIC differences $\Delta_{A,B}$ (Section 2.6). This is a method of arriving at the odds ratio of (4.4) assuming that the AIC difference between the candidate models is a perfect unbiased estimator of the difference between the models’ KL divergences with respect to the truth. Since it assumes the existence of some objective truth, this method is based on the frequentist interpretation of probability. This is different from the original interpretation of AIC differences which is based on Bayesian statistics.

10.2 Numerical studies of AIC in the context of cosmological model selection

The analysis of $\Delta_{A,B}$ was extended further by investigating the statistical uncertainty of this estimate. Our focus was not necessarily on the ‘best’ cosmological theory. Thus, we did not use the most exhaustive
datasets, nor explore in detail the systematics associated with the surveys considered. To this end, we studied the distribution of the differences of AIC estimates given a certain quality of data \( P(\Delta_{A,B}|E_n) \). Since we do not know the exact process underlying the empirical data \((E_n\text{ in Section 6.1})\), we approach this problem by studying surrogate processes, where the reference or generating model is assumed to be one of four (best fit) candidate models for the late time acceleration of the universe; following the approach used by Davis et al. [25]. These models were listed in Table 5.1 with the best fit free parameters equal to the maximum likelihood values of each of these models. This was obtained by fitting 371 SNIa extracted from the Constitution compilation [40].

Our simulations have demonstrated that, given the data used, there was insufficient data to reliably use AIC to tell all the models apart; in agreement with the general consensus [25]. To do so, we defined \( \Delta_{\text{threshold}} \) as the minimum AIC difference between two AIC values to reliably tell two models apart. For the case of \( \Delta_{\text{threshold}} = 0 \), the failure rate of the technique was shown to be particularly unsatisfactory. We also studied the statistical self-consistency of the AIC technique when \( \Delta_{\text{threshold}} \) is increased to 2 and 5. Increasing \( \Delta_{\text{threshold}} \) results in increasing the number of cases where we cannot make a conclusion based on the AIC procedure. This was demonstrated by \( f_{\text{ind}} \), which calculates the fraction of cases when the difference in AIC values between two models is less than \( \Delta_{\text{threshold}} \). We also studied \( f_{\text{false}}^{\text{det}} \), the proportion of cases where the AIC procedure using a threshold \( \Delta_{\text{threshold}} \) gives an incorrect result as a fraction of cases where we can make a conclusion (i.e., \(|\Delta_{A,B}| > \Delta_{\text{threshold}}\), where \( A \) is the reference model.). We showed that \( f_{\text{false}}^{\text{det}} \) does not necessarily decrease in the same universal way with an increasing \( \Delta_{\text{threshold}} \). Therefore, even when AIC chooses a model class (with a high level of \( \Delta_{\text{threshold}} \)), the result is unreliable for at least some models within that model class. The demonstrated examples would perhaps not arise if the data was good enough that the differences in \( \chi^2_{\text{ML}} \) was large. While AIC has been shown to be not consistent in the asymptotic limit, the technique is still be useful as it could quantify the likelihood of the AIC technique giving self-consistent results for finite size datasets. Even if it does not give a self-consistent result 100% of the time, a high likelihood (say 80% of the time) would mean that the technique can still be useful.

We also calculated the respective statistical uncertainty \((\sim 1 \sigma \text{ error bars})\) of \( \Delta_{A,B} \) and showed it was even larger than the observed \( \Delta_{A,B} \) between some of the models. This gives us a way to gage the adequacy in the number of data points since the statistical uncertainty would become smaller than the observed differences when there is a sufficient amount of data. As an important example, we considered the ΛCDM and DGP models since they were shown to have the two lowest AIC values in Table 5.1. It was shown that the statistical uncertainty of \( \Delta^d_{\Lambda\text{CDM,DGP}} \) was larger than the observed \( \Delta_{\Lambda\text{CDM,DGP}} \), making it difficult to determine the better model between the two. From our simulation, we also showed that the shapes of the
distribution of the AIC differences can be quite varied, ranging from a symmetric Gaussian-like distribution to an exponential-like distribution with a sharp edge and one sided tail. Thus, in order to use AIC reliably, one must pay proper attention to the statistical variation of $\Delta_{A,B}$.

In this thesis, we made a number of assumptions to study the AIC technique. All calculations in this thesis were only for an assumed reference. Since the empirical data does not give us $E_n$, there is no way to know the actual distribution of $\Delta_{A,B}$. However, we should note that AIC is a model comparison technique that assumes that one of the model classes contains the reference $C$. By restricting $C$ to the set of candidate models, we can at least look for statistical self-consistency in that assumption. It should be emphasized again that the reference models used were the best fit models and did not take into account the statistical uncertainty of the individual model parameters. That can be taken into account by sampling the distribution of parameters. As mentioned before, the whole point of this simulation is to highlight the statistical distribution of the AIC differences under different reference models and look for statistical inconsistencies under each of the assumptions. Another issue that should be noted is that the exact variance in the data is unknown and that the error bars in the data may not be reflective of the true error bounds. While we use these in our simulations, we note that the correctness of these error estimates was an assumption of previous AIC computations.

In summary, the reliability of the AIC estimator is an important issue that should be taken into consideration when using the AIC technique to select models. It should also be noted that such considerations are not just restricted to AIC but any technique that relies on the maximum likelihood estimators. This should be borne in mind when applying the techniques to any statistical analysis.

10.3 Stability as an alternative to complexity

Besides numerically studying the statistical reliability of AIC in the context of cosmological model selection, we also compared the AIC strategy against other model selection techniques. In our comparison with BIC and MDL, we showed that the AIC $2k$ cost term does not fully account for the model complexity. This cost term is a bias correction that is used to estimate the KL divergence of the model against the truth (up to a fixed offset). This quantity is better estimated using TIC. However, the cost term is difficult to calculate analytically. A practical strategy of computing the cost term (without resorting to the simplifying restrictions found in AIC) is via the use of the bootstrap technique to estimate the bias correction.
One variant of the bootstrap technique involves computing the bias correction:

$$\text{Bias} = 2E_x \left[ \log f \left( y|\hat{\theta}(y^*) \right) - \log f \left( y|\hat{\theta}(y) \right) \right],$$

where $^*$ indicates the bootstrap datasets. This seems to superficially favor a model that is more stable against bootstrap perturbations since a model that is stable against bootstrap perturbations would lead to parameter values $\hat{\theta}(y^*)$ that are only slightly different from $\hat{\theta}(y)$. However, the motivation for this term is to correct for the finite size effects of the dataset and not any stability considerations. We see an analogous relationship with the cost term in traditional AIC and complexity.

This hints at the use of model stability as an important consideration in model selection. Our focus was not necessarily on the most rigorous derivation of a model selection criterion but to demonstrate the principle of stability in model selection. Thus, we presented two plausible hypothetical model selection criteria that incorporated model stability as a consideration. These criteria included a cost term that measures the susceptibility (instability) of the model class to minor data perturbations. Two ways of measuring this susceptibility was proposed. One based on the KL divergence and the other inspired by the $l_2$ distance measure. We studied the performance of these criteria using the test example of polynomial model selection. From our analysis of the instability cost term, we see that it performs better than traditional AIC. To summarize, we feel that model stability should also be an important consideration in model selection.
Appendix A

Bayesian Interpretation of Akaike Differences

This appendix is a verbatim reproduction of the work by Akaike [4]. The motivation for this reproduction is to allow the reader to compare the existing Bayesian interpretation of AIC differences to our proposed frequentist interpretation. It is included for the convenience of the reader.

Consider a parametric family of data distributions \( \{ f(x|\theta) | \theta \in \Theta \} \) with an \( L \)-dimensional parameter \( \theta = (\theta_1, \theta_2, \ldots, \theta_L)^T \). We assume a parametrization such that the Fisher information matrix takes the form \( A = \frac{1}{\sigma^2} I \) where \( I \) is an \( L \times L \) identity matrix. Consider the selection of a model from a set of models where the \( k \)th model is defined with a parameter \( \theta_k \) which is constrained to lie in the subset of parameters with \( \theta_j = 0 \) for \( (j = k + 1, k + 2, \ldots, L) \).

To establish a characterization of AIC we consider a set of Bayesian models where the \( k \)th model is specified by the data distribution \( f(x|\theta) \) and a prior distribution \( \pi_k(\theta) \) given by

\[
\pi_k(\theta) = \left( \frac{1}{\sqrt{2\pi\delta_1}} \right)^k \exp \left[ -\frac{1}{2\delta_1^2} \sum_{i=1}^{k} \theta_i^2 \right] \left( \frac{1}{\sqrt{2\pi\delta_2}} \right)^{L-k} \exp \left[ -\frac{1}{2\delta_2^2} \sum_{j=k+1}^{L} \theta_j^2 \right], \tag{A.1}
\]

where it is assumed that \( \delta_1 > \sigma > \delta_2 > 0 \). This model represents our expectation of larger and smaller variations of the first \( k \) and the last \( L-k \) components of \( \theta \), respectively, compared with the expected sampling variability represented by \( \sigma \), of the corresponding components of the maximum likelihood estimate. The likelihood of the \( k \)th model is defined by

\[
l(k) = \int f(x|\theta)\pi_k(\theta)d\theta. \tag{A.2}
\]

To perform the integration we assume the representation

\[
f(x|\theta) = f(x|\theta^*) \exp \left[ -\frac{1}{2\sigma^2} (\theta - \theta^*)^T A (\theta - \theta^*) \right], \tag{A.3}
\]
where \( \theta^* \) denotes the maximum likelihood estimate. Except for the linear Gaussian model with homogeneous observation errors this representation will only be a reasonable approximation to the true likelihood function for \( \theta \) and \( \theta^* \) lying in a small neighborhood of the true parameter.

By adopting the above representation of \( f(x|\theta) \) we get

\[
-2 \log l(k) = k \log (\sigma^2 + \delta_1^2) + \frac{1}{\sigma^2 + \delta_1^2} \sum_{i=1}^{k} (\theta^*_i)^2 + (L-k) \log (\sigma^2 + \delta_2^2) + \frac{1}{\sigma^2 + \delta_2^2} \sum_{j=k+1}^{L} (\theta^*_j)^2 + c, \quad (A.4)
\]

where \( c \) is a constant independent of \( k \), and thus the limiting equation

\[
\lim_{\delta_1 \to \sigma} \lim_{\delta_2 \to \sigma} \left( \frac{4\sigma^2}{\delta_1^2 - \delta_2^2} \right) [-2 \log l(k) + 2 \log l(L)] = 2k + \frac{1}{\sigma^2} \sum_{j=k+1}^{L} (\theta^*_j)^2 - 2L. \quad (A.5)
\]

Since AIC of the original \( k \)th model is given by \( \text{AIC}(k) = -2 \log f(x|\theta^*_k) + 2k \), and since from the above representation we have

\[
f(x|\theta^*_k) = f(x|\theta^*) \exp \left[ -\frac{1}{2\sigma^2} \sum_{j=k+1}^{L} (\theta^*_j)^2 \right], \quad (A.6)
\]

the right hand side of the above equation is identical to \( \text{AIC}(k) - \text{AIC}(L) \), i.e., the AIC difference is obtained as the differential log likelihood ratio of the Bayesian models. Thus the difference of the two AIC’s tend to be proportional to the log likelihood ratio of the corresponding Bayesian models where both \( \delta_1 \) and \( \delta_2 \) approach to \( \sigma \).
Appendix B
Confusion Probability

This appendix contains verbatim reproduction of part of the work done by Balasubramanian [8]. It also contains a verbatim reproduction of Stein’s lemma taken from Chapter 12.8 in a book by Cover and Thomas [24]. The motivation for these reproductions is to provide a starting point to our proposed frequentist interpretation of AIC differences. They are included for the convenience of the reader. It should also be noted that this appendix is also present in a paper co-written by the author. The paper is titled ‘The reliability of the AIC method in Cosmological Model Selection’ and has been accepted for publication by the journal ‘Monthly Notices of the Royal Astronomical Society’. That paper is coauthored with Dr. Rahul Biswas.

We start with an almost identical repeat of Balasubramanian’s explanation of error probabilities [8], which is framed in the language of hypothesis testing. Suppose \( \{x_1, x_2, \ldots, x_N\} \in \mathcal{X}^N \) are drawn independent and identically distributed (i.i.d.) variables from one of \( f_1 \) and \( f_2 \) with \( D(f_1 \parallel f_2) < \infty \). Let \( A_N \subseteq \mathcal{X}^N \) be the acceptance region for the hypothesis that the distribution is \( f_1 \) and define the type I and type II error probabilities as \( \alpha_N = \int_{A_C N}^f f_1^n(A_N) \alpha \) and \( \beta_N = \int_{A_N}^f f_2^n(A_N) \), respectively. \( A_C N \) is the complement of \( A_N \) in \( \mathcal{X}^N \), and \( f^n \) denotes the product distribution on \( \mathcal{X}^N \) describing \( N \) i.i.d. outcomes drawn from \( f \). In this definition \( \alpha_N \) is the probability that \( f_1 \) is mistaken for \( f_2 \), and \( \beta_N \) is the probability of the opposite error. Stein’s lemma tells us how low we can make \( \beta_N \) given a particular value of \( \alpha_N \). Indeed, let us define \( \beta_N^\epsilon = \min_{\alpha_N \leq \epsilon} \beta_N \) for a positive \( \epsilon \). Then Stein’s lemma tells us

\[
\lim_{\epsilon \to 0} \lim_{N \to \infty} \frac{1}{N} \ln \beta_N^\epsilon = -D(f_1 \parallel f_2).
\]

To prove Stein’s lemma, we refer to the proof by Cover and Thomas [24], which is provided almost verbatim here for convenience sake. Defining \( \delta \in \mathbb{R}^+ \), we first state \( A_N \) more explicitly as:
\[
A_N = \left\{ x \in \mathcal{X}^N : \exp[N (D(f_1 \parallel f_2) - \delta)] \leq \frac{f_1(x)}{f_2(x)} \leq \exp[N (D(f_1 \parallel f_2) + \delta)] \right\}
\]
Then, we have the following properties:
1. $f_1^N(A_N) \to 1$.

Proof:

$$f_1^N(A_N) = f_1^N \left( N \sum_{i=1}^{N} \log \frac{f_2(x_i)}{f_1(x_i)} \in (D(f_1 \parallel f_2) - \delta, D(f_1 \parallel f_2) + \delta) \right) \to 1 \text{ by the law of large numbers, since } D(f_1 \parallel f_2) = \mathbb{E}_{f_1} \left( \log \frac{f_1(x)}{f_2(x)} \right).$$

Therefore, for any positive $\epsilon$, $\alpha_N < \epsilon$ for large $N$.

2. $f_2^N(A_N) \leq \exp \left[ -N (D(f_1 \parallel f_2) - \delta) \right]$.

Proof:

$$f_2^N(A_N) = \sum_{A_N} f_2(x),$$

$$\leq \sum_{A_N} f_1(x) \exp \left[ -N (D(f_1 \parallel f_2) - \delta) \right],$$

$$= \exp \left[ -N (D(f_1 \parallel f_2) - \delta) \right] \sum_{A_N} f_1(x),$$

$$= \exp \left[ -N (D(f_1 \parallel f_2) - \delta) \right] (1 - \alpha_N).$$

3. $f_2^N(A_N) \geq \exp \left[ -N (D(f_1 \parallel f_2) + \delta) \right]$.

Proof:

$$f_2^N(A_N) = \sum_{A_N} f_2(x),$$

$$\geq \sum_{A_N} f_1(x) \exp \left[ -N (D(f_1 \parallel f_2) + \delta) \right],$$

$$= \exp \left[ -N (D(f_1 \parallel f_2) + \delta) \right] \sum_{A_N} f_1(x),$$

$$= \exp \left[ -N (D(f_1 \parallel f_2) + \delta) \right] (1 - \alpha_N).$$

4. $\lim_{N \to \infty} \frac{1}{N} \log \beta_N = -D(f_1 \parallel f_2)$.

Proof:

From 2. and 3. we know:

$$\frac{1}{N} \log \beta_N \leq -D(f_1 \parallel f_2) + \delta + \frac{\log(1 - \alpha_N)}{N},$$

$$\frac{1}{N} \log \beta_N \geq -D(f_1 \parallel f_2) - \delta + \frac{\log(1 - \alpha_N)}{N}.$$
5. No other sequence of acceptance regions does better.

Proof: Let \( B_N \subseteq \mathcal{X}^N \) be any other sequence region with \( \alpha_{N,B_N} = f_1^N(B_N^c) < \epsilon \). Let \( \beta_{N,B_N} = f_2^N(B_N) \)

\[
\begin{align*}
\beta_{N,B_N} &= f_2^N(B_N), \\
&\geq f_2^N(A_N \cap B_N), \\
&= \sum_{A_N \cap B_N} f_2(x), \\
&\geq \sum_{A_N \cap B_N} f_1(x) \exp[-N(D(f_1 \parallel f_2) + \delta)], \\
&= \exp[-N(D(f_1 \parallel f_2) + \delta)] \sum_{A_N \cap B_N} f_1(x), \\
&\geq (1 - \alpha_N - \alpha_{N,B_N}) \exp[-N(D(f_1 \parallel f_2) + \delta)],
\end{align*}
\]

where the last inequality is due to the following:

\[
\sum_{A_N \cup B_N} f_1(x) = f_1(A_N \cap B_N),
\]

\[
= 1 - f_1(A_N^c \cup B_N^c),
\]

\[
\geq 1 - f_1(A_N^c) - f_1(B_N^c),
\]

\[
= 1 - \alpha_N - \alpha_{N,B_N}.
\]

Hence, \( \frac{1}{N} \log \beta_{N,B_N} \geq -D(f_1 \parallel f_2) - \delta - \frac{\log(1 - \alpha_N - \alpha_{N,B_N})}{N} \), and since \( \delta > 0 \), \( \lim_{N \to \infty} \frac{1}{N} \log \beta_{N,B_N} \geq -D(f_1 \parallel f_2) \). Thus no sequence of sets \( B_N \) has an exponent better than \( D(f_1 \parallel f_2) \).

In summary, property 1. shows that \( A_N \) is the sequence that is generated by \( f_1 \) in the asymptotic limit. Properties 2., 3. and 4. derive the error probability of Stein’s lemma and property 5. shows that \( A_N \) is asymptotically optimal and the best error exponent is \( D(f_1 \parallel f_2) \).

Thus, we can interpret \( \exp[-D \text{ (truth \parallel model)}] \) as the probability of confusing the model with the truth or model probability, using the work of Balasubramanian [8].
Appendix C

Minimization Procedures

This is a summary of the minimization techniques that were used to compute the $\chi^2$ fit for the Type Ia SN data.

C.1 Gauss-Newton algorithm

The Gauss-Newton algorithm [13] is adapted from Newton’s method [58] to solve the special case of non-linear least square minimization ($\chi^2$ fitting). However, unlike Newton's method, the Gauss-Newton algorithm does not require a second order derivative, only a first order derivative, making it significantly easier to compute.

To derive the algorithm, let us first define the following. Let $\beta$ denote a vector of $n$ model parameters (i.e., $\beta = \{\beta_1, \ldots, \beta_n\}$). Let there be $m$ functions of $r_i$, $\{r_1, \ldots, r_m\}$.

The least square problem involves minimizing the function

$$S(\beta) = \sum_{i} r_i^2(\beta).$$  \hspace{1cm} (C.1)

The optimal $\beta$ (minimizes $S(\beta)$) is solved by an iterative process. Defining $\beta^{(k)}$ as the $k$th iteration, we start by guessing $\beta^{(0)}$ and using the updating step:

$$\beta^{(k+1)} = \beta^{(k)} + \Delta,$$  \hspace{1cm} (C.2)

where $\Delta$ is a small step.

To find $\Delta$, we first use the Taylor Expansion to get

$$S(\beta + \Delta) \approx S(\beta) + \frac{\partial S}{\partial \beta} \cdot \Delta + \frac{1}{2} \Delta^T \left[ \frac{\partial^2 S(\beta)}{\partial \beta_i \partial \beta_j} \right] \Delta + o(\Delta^3).$$  \hspace{1cm} (C.3)

74
If we define the Jacobian matrix, \( J_r(\beta) = \frac{\partial r_i}{\partial \beta_j} \), we can replace \( \frac{\partial S}{\partial \beta_i} \) with \( J_r^T r \), and the Hessian matrix \( \frac{\partial^2 S(\beta)}{\partial \beta_i \partial \beta_j} \) can be approximated by \( J_r^T J_r \) (assuming small residuals).

This leads to
\[
S(\beta + \Delta) \approx S(\beta) + J_r^T r \Delta + \frac{1}{2} \Delta^T J_r^T J_r \Delta. \tag{C.4}
\]

Taking the derivative with respect to \( \Delta \) and setting it to zero, we get
\[
S'(\beta + \Delta) \approx J_r^T r + J_r^T J_r \Delta = 0. \tag{C.5}
\]

This is the zero gradient condition that minimizes \( S(\beta) \) and leads to the normal equation
\[
\left( J_r^T J_r \right) \Delta = -J_r^T r. \tag{C.6}
\]

The \( \Delta \) solution of this normal equation is used in the updating step shown above.

## C.2 Correlated sum of squares

The sum of squares for \( n \) data points can be reexpressed as \( X^T X \) where \( X \) is an \( n \times 1 \) vector. This sum of squares can be easily and quickly minimized by the Gauss-Newton algorithm. However the marginalization procedure in our cosmological example results in a correlated sum of squares. We would have to minimize \( X^T \Lambda X \) where \( \Lambda \) is a symmetric \( n \times n \) matrix.

Since \( \Lambda \) is positive definite, real and symmetric, we can apply the Cholesky decomposition: \( \Lambda = C^T \times C \), where \( C \) is the Cholesky decomposition. Using the transform \( X \to X' = C \times X \), we can apply the Gauss-Newton algorithm to minimize \( X'^T X' \) and hence minimize the correlated sum of squares.

## C.3 Hybrid grid-gradient search

When we first applied the Gauss-Newton technique to the CPL and wCDM models, two problems were encountered.

1. The procedure would occasionally get unphysical \( \Omega_m \) values (i.e., \( \Omega_m < 0 \)).
2. The program would halt because it hits a singularity. This is due to the fact that it involves the numerical integration \( \int_0^z \frac{dz'}{H(z')} \).

These problems can be solved by taking a grid search method. However, such an approach would be too time consuming, since we need to do the minimization procedure over multiple bootstrap samples. One way to get around the problem would be to implement a hybrid method. This first involves temporarily fixing \( \Omega_m \) and using a gradient search to find the optimal values for the other parameters given some fixed \( \Omega_m \). We repeat this search for a series of \( \Omega_m \) values, starting at \( \Omega_m = 0 \) and ending at \( \Omega_m = 0.6 \) at small regular intervals.

C.4 Integration

The numerical integration was done using a one dimensional quadrature. The algorithm was copied from numerical recipes. For information of the technique, we refer the reader to [58].

C.5 Implementation

The minimization code was written in Visual C++ and ran on a home PC.
References


80
