H 24/3598

MONASH UNIVERSITY
THESIS ACCEPTED IN SATISFACTION OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
ON..... 24 February 2004 .....

.....[redacted].....

Sec. Research Graduate School Committee

Any other publications:

Grace W. Rumantir and Chris S Wallace, Minimum Message Length Criterion for Second-Order Polynomial Model Selection Applied to Tropical Cyclone Intensity Forecasting, in M.R. Berthold, et al. (eds.), Advances in Intelligent Data Analysis V, LNCS 2810, pp. 486-496, Springer-Verlag Berlin Heidelberg 2003.

G.W. Rumantir, Frequent Flyer Points Calculators: More Than a Table Lookup, in A. Abraham, J. Ruiz-del-Solar, M. Koppen (eds.), Soft Computing Systems: Design, Management and Applications, Frontiers in Artificial Intelligence and Applications vol. 87, pp. 871-880, IOS Press, Amsterdam, 2002

Grace W. Rumantir and Chris S Wallace, Sampling of Highly Correlated Data for Polynomial Regression and Model Discovery, in F. Hoffman, et al. (eds.), Advances in Intelligent Data Analysis IDA 2001, LNCS 2189, pp. 370-377, Springer-Verlag Berlin Heidelberg 2001.

Grace W. Rumantir, Tropical Cyclone Intensity Forecasting Model: Balancing Complexity and Goodness of Fit, in R. Mizoguchi and J. Slaney (eds.), PRICAI 2000 Topics in Artificial Intelligence, LNAI 1886, pp.230-240, Springer-Verlag Belin Heidelberg 2000.

Grace W. Rumantir, Minimum Message Length Criterion for Second-order Polynomial Model Discovery, in T. Terano, H. Liu and A.L.P. Chen (eds.), Knowledge Discovery and Data Mining: Current Issues and New Applications, PAKDD 2000, LNAI 1805, pp.40-48, Springer-Verlag Berlin Heidelberg 2000.

[Title of publication, publisher and date of publication]

4. Declaration by candidate

Candidate's signature: [redacted] Date: 14/10/03

5. Ratification by academic unit

This is to ascertain that the Department / School / Centre / Institute has no objection to the candidate's options regarding access to the Library thesis copy. If so, please sign below and return the completed form to: **Monash Graduate School, Building 3D, Clayton Campus.**

Supervisor's signature: [redacted] Date: 14/10/03

(please print name) [redacted]

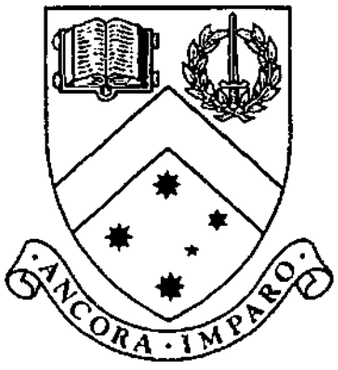
Errata

p.77, Sample Size 2000, Method MML, column ModelErr: "0.0022" for "0.0007"
p.84 point (4): "our" for "out"
p.87 para 2, first sentence: "called," for "called"
p.88 5^{th} line: "substracting" for "subtracting"
p.93 2^{nd} para: "Osciallation" for "Oscillation"
p.105 para 1, last sentence: "page 4.1" for "page 96"
p.108 para 2: "on test data set" for "on the test data set"
p.133 last sentence: "a set variables" for "a set of variables"

Minimum Message Length Criterion for Second-order Polynomial Model Selection Applied to Tropical Cyclone Intensity Forecasting

Grace Widjaja Rumantir

B.Civ.Eng, GradDipComp, MAppSc(IT), MBA



A thesis submitted for the degree of

Doctor of Philosophy

in the School of Computer Science and Software Engineering

Monash University

Monash University

October 2003

To Mama

Contents

List of Tables

List of Figures

List of Procedures

List of Symbols

β_k the coefficient of the regressor/independent variable x_k

$\underset{N\times K}{\mathbf{X}}^{\mathrm{T}}$ is a brief form of $(\underset{N\times K}{\mathbf{X}})^{\mathrm{T}}$; the transpose of the N × K matrix of independent variables. N is the number of rows and K is the number of column of the matrix.

$L(\theta)$ the cost of encoding model θ. This is represented as $L(\theta) = -\log P(\theta)$ with $P(\theta)$ as the probability density of the model θ

$L(D)$ the cost of encoding data set D. This is the code length representation of the marginal probability density of D, $L(D) = L(\theta) + L(D|\theta)$

$L(D|\theta)$ is the length of the encoding of the probability of observing data set D given the current model θ. This is represented as $L(D|\theta) = -\log P(D|\theta)$

r_n the error of the model of the n^{th} data item.
$r_n = y_n - \sum_{k=1}^{K} \beta_k x_{nk}$ with the assumption that $r_n \sim NID(0, \sigma^2)$

x_{nk} the n^{th} data item of the regressor/independent variable x_k

y_n the n^{th} data item of the target variable y.
$y_n \sim N(\sum_{k=1}^{K} \beta_k x_{nk}, \sigma^2)$

K the number of independent variables/regressors

N the number of data items in the sample set

List of Publications

The following is the list of publications arising from this thesis.

Referred Journal Papers (C1 publications)

G. W. Rumantir and C. S. Wallace, Minimum Message Length Criterion for Second-order Polynomial Model Selection Applied to Tropical Cyclone Intensity Forecasting, In M.R. Berthold, et al., editors, *Advances in Intelligent Data Analysis V, LNCS2810*, pp. 486–496, Springer–Verlag, Berlin Heidelberg, 2003.

G. W. Rumantir and C. S. Wallace, Sampling of Highly Correlated Data for Polynomial Regression and Model Discovery, In F. Hoffmann, et al., editors, *Advances in Intelligent Data Analysis, LNCS 2189*, pp. 370–377, Springer–Verlag, Berlin Heidelberg, 2001.

G. W. Rumantir, Tropical Cyclone Intensity Forecasting Model: Balancing Complexity and Goodness of Fit, In R. Mizoguchi and J. Slaney, editors, *PRICAI 2000 Topics in Artificial Intelligence, LNAI 1886*, pp. 230–240, Springer–Verlag, Berlin Heidelberg, 2000.

G. W. Rumantir, Minimum Message Length Criterion for Second-order Polynomial Model Discovery, In T. Terano, H. Liu, A.L.P. Chen, editors, *Knowledge Discovery and Data Mining: Current Issues and New Applications, PAKDD 2000, LNAI 1805*, pp. 40–48, Springer–Verlag, Berlin Heidelberg, 2000.

Abstract

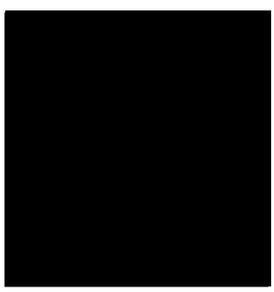
The main contributions of this thesis are four-fold. First, a new criterion based on the Minimum Message Length principle specifically formulated for the task of polynomial model selection up to the second order is presented. A structured description of most of the criteria commonly used for polynomial model selection is also presented. Second, a programmed optimisation search algorithm for second-order polynomial models that can be used in conjunction with any model selection criterion is developed. Third, critical examinations of the differences in performance of the various criteria when applied to artificial *vis-a-vis* to real tropical cyclone data are conducted. Three different data sets are used: data generated from known artificial models, real climatological and atmospheric data from the Atlantic tropical cyclone basin and artificial data generated based on the covariance matrix of the real data from the Atlantic basin. Fourth, a novel strategy which uses a synergy between the new criterion built based on the Minimum Message Length principle and other model selection criteria namely, Minimum Description Length, Corrected Akaike's Information Criterion and Structured Risk Minimization is proposed. With this combinatorial strategy, polynomial models with increasing levels of complexity can be made available. This enables human experts to study the level of contribution of each individual variable and make an informed decision on which model to choose. Thus, whilst the combinatorial strategy converges to one best model in an

automated manner, it still allows human experts to choose models with higher/lower levels of complexity than the proposed best model.

A concise description of the full application of this novel strategy for building a tropical cyclone intensity change forecasting model from the collection of variables involved and data pre-processing to the model selection procedure is outlined. The forecasting model developed using this new automated strategy has better performance than the benchmark models SHIFOR (Statistical HurrIcane FORcasting) and SHIFOR94 which are being used in operation in the Atlantic basin. Unlike the benchmark models and all the other models in operational use, the new model uses seasonal variables which have already been proven by atmospheric scientists to have strong influence on the activity of cyclones in the Atlantic basin. This new strategy can be used for any domain data which can be represented as second-order polynomial models.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other institution. To the best of my knowledge, this thesis contains no material proviously published or written by another person, except where due reference is made in the text of the thesis. Where the work in this thesis is based upon joint research, this thesis discloses the relative contributions of the respective authors.



Grace Widjaja Rumantir

Melbourne, October 2003

Acknowledgments

To me, this research has been a great adventure of setting up and carrying out a research agenda at an advanced level in a cross-cutting highly interdisciplinary research between Computer Science and Meteorology. In this adventure, I have been enriched by the ideas and goodwill of the people that I have had the privilege to be associated with.

I have received continual valuable guidance and challenges from my principal supervisor, Emeritus Prof. Chris Wallace. My associate supervisor, Dr. Peter Tischer, with his constructive proofreading and suggestions has been instrumental in the production of this thesis. My external supervisor, Dr. Greg Holland, formerly of the Bureau of Meteorology, currently of the Aerosonde Ltd., has been responsible for introducing me to tropical cyclone research and to the key people in the area. Dr. Chris Landsea of the Hurricane Research Division – National Oceanic and Atmospheric Administration (HRD/NOAA), with his generosity in giving me data and programs on the Atlantic tropical cyclone basin and answers to my questions on the data, has played a pivotal role in the tropical cyclone part of this thesis. I have learned a great deal from these great scientists, only a part of which is written in this thesis.

The support of Dr. Robert Abbey, Jr. of the US Office of Naval Research has been instrumental in enabling me to start pursuing a joint collaborative research between the Bureau of Meteorology and the School of Computer Science and Software Engineering at Monash University. Prof. Ann Henderson-Sellers, formerly of the Climatic Impact Centre - Macquarie University, currently of the Australian Nuclear Science and Technology Organisation, has been responsible for opening the door to research in atmospheric science to me. Mr. Rex Falls of the Bureau of Meteorology - Queensland Regional Office has been generous in making himself available as the industrial partner of this research project. Mr. Milton Richardson and A/Prof. Trevor Dix have been very helpful in sorting out the administration of the funding for this project.

I am grateful for the love, support and confidence that my family has always offerred me. I am especially thankful to my mother for wanting this thesis finished, may be more than I do. I am thankful to my younger sister, Magdalena Rumantir, MD, PhD who, after looking up to me all her life, dared to say that I would be an embarrassment to her if I did not get my PhD completed.

I was supported for this research by an Australian Postgraduate Award (Industry) under Project Grant No. AP9700180 with the Brisbane Tropical Cyclone Warning Centre, Bureau of Metecrology - Queensland Regional Office as the industrial collaborator through the Tropical Cyclone Coastal Impacts Project (TCCIP). I acknowledge a one year postgraduate scholarship from the Climatic Impacts Centre of Macquarie University prior to my transfer to Monash University.

Chapter 1

Introduction

This chapter provides a broad overview of polynomial model selection, criteria of a good model, methods for selecting models, Minimum Message Length principle for model selection and some background information of tropical cyclone intensity forecasting research. This chapter also gives an outline of this thesis highlighting the contributions made to polynomial model selection research and tropical cyclone forecasting research[1]. It provides the setting and motivations for the results presented in the subsequent chapters of this thesis.

1.1 Polynomial Model Selection for Forecasting

Model selection is about finding structure from data. The goal of model selection is to use a limited amount of sample data to find a model or a set of models that best explain the structure of the data population. To achieve this goal, two main tasks need to be carried out. The first is in selecting the variables that are useful to be included in the model and the second is in finding the structure of the model.

[1] An earlier version of parts of this chapter has been published in [45]

Finding the structure of a model involves two tasks. The first is in determining the form in which the variables collaborate, e.g. Bayesian networks, neural networks, decision tree, polynomials, etc. The second is in determining the degree of influence each variable has in the model.

One use of a model is for the purpose of forecasting future values of a particular variable. For a model that takes the form of a polynomial, for example, the values of the variables of the model measured at one time are used to predict the value of one other variable which is seen to be dependent on them. The dependent variable is thus formulated as a polynomial function of the selected set of independent variables.

This thesis is about finding a model selection strategy which includes an optimisation search algorithm to select variables and model selection criteria to evaluate the ability of a model to predict future values. The form of the model is fixed to be a second-order polynomial which may use the squares of variables and products of two variables. For a given form, once the variables are selected, the coefficients of the variables are determined by computing a least squares solution to the data. Hence, the way in which models differ is in the variables they use.

The form of the forecasting model was fixed because models based on the real data used to test the model selection strategy proposed in this thesis already exist and are in operational use in the Atlantic basin. Hence, the model to be built using the new model selection strategy can be benchmarked against these models.

A model can be used to obtain a point estimate of the dependent variable, i.e. given the values of the selected variables, a single value is obtained for the dependent variable. A model can also be used to predict a probability distribution for the dependent variable. For example, rather than predicting the intensity of a particular cyclone 24 hours into the future, the model would assign probabilities that the

cyclone will intensify by up to certain intensity measurements. This kind of forecasting requires an assumption of the type of probability distribution of the dependent variable. This thesis will focus mainly on the point estimation problem.

In [94], Chris Wallace, the inventor of the Minimum Message Length principle, is cited to have said that "measurement + modelling = science". Along a similar line, Peter Tischer said that "measurement + mathematics = science". The task of building a model often involves data gathered from observations/measurements and data of derived variables which are calculated based on some deterministic mathematical modelling techniques. Both types of data are then used as input to some stochastic modelling techniques. For example, for the task of building a tropical cyclone forecasting model, data from field measurements, e.g. wind speed, humidity, as well as derived variables such as potential intensity, prediction of cyclone direction, are used as input to build a model. So in this case, the second component of the above formulas includes not only the task of finding a model given the variables, but also the task of generating derived variables from existing measurements.

In this thesis, the pool of variables from which the tropical cyclone intensity forecasting models will be selected consists of the original set of measured variables and derived variables calculated from them. Since the form of the models is fixed to be second-order polynomial as explained earlier in this Section, such variables include the squares of variables and the product of two variables. Chapter 4 will discuss the measured variables and derived variables resulting from thermodynamic computations of the variables.

Apart from for the purpose of forecasting, a model can be selected for the following other purposes:

- To use variables that can be measured more cheaply to estimate the value of a variable which happens to be expensive to measure
- To provide an explanation of the effects of the selected variables on the dependent variable

The particular task of model selection explained so far is in selecting a model which, based on the values of the selected variables, is able to determine the value of the dependent variable. The variables concerned are usually continuous. A different task of model selection is that of classification. In classification, each value of the dependent variable represents a category which can be represented as a discrete variable. A category represents a certain combination of values of the selected variables. The task of finding a classification model is to find the selected variables that can be used to determine the category of a given set of their values.

1.2 The Number of Variables in Model

A good forecasting model is one which yields predictions on future data with

- minimum bias, i.e. the sampling distribution of the prediction errors centers on zero, and
- minimum variance, i.e. the spread of the sampling distribution of the prediction errors is small

Given sample sets of data of the dependent and potential independent variables, it is known that the more independent variables used in the forecasting model, the better the fit to the sample set of the dependent variable can be. The aim is not

to find a model that best fits the sample set it is built from, but for the model to represent the general features of the function of the dependent variable so that it can make good predictions on different sample sets. This means the model should include enough of the independent variables to capture the general features of the dependent variable and exclude those that only capture the features specific to the particular sample set of the dependent variable from which the model is built.

1.3 The Correct Model

When dealing with a real-world application domain where observational data are used to build a model, the true model that represents the dependent variable is not known. Hence, no model built using the sample data sets available can be said to be correct. What can be shown is that some models are either consistent or inconsistent with the data based on certain performance criteria.

In the face of the many forms a model can take, it is a common practice in model selection research to make an assumption of the form the model should take. For example, in this thesis, the form that the model is assumed to have is a second-order polynomial. The model selection strategy will then find the set of independent variables that constitute the polynomial model and the values of their individual parameters. The parameters represents the degree of influence an independent variable has on the dependent variable. If a variable has no influence at all then the value of its parameter is zero, i.e. the variable is not selected in the model. If, in reality, the dependent variable is actually not representable in the form that the model is assumed to have, then the predictive performance of the model may not be very good.

Poor predictive power can also be caused by poor quality of the data sets. The existence of noise and/or erroneous measurements may lead to the wrong model being selected. The detection and correction of outliers and the choice of a model selection method that are robust in the presence of noisy data can help.

The absence of some influential independent variables in the pool of variables to choose from may cause poor predictive power of the forecasting model selected unless these missing variables can be substituted by a combination of other variables in the model.

It is intrinsically difficult to build a forecasting model for some problem domains. This is because the problem domain is chaotic. A slight change in the value of a particular variable can significantly change the balance of the domain. In such a domain, the forecasting model built may have limited predictive power but will still be useful in assisting decision makers in making informed decisions. Examples of this kind of domain are predicting the outcome of a lotto draw, stock market predictions and tropical cyclone intensity forecasting. A tropical cyclone can strengthen or weaken depending on the terrain, climatological, synoptic/environmental and seasonal factors in the vicinity of the cyclone and the areas it is predicted to be heading to. The current and future states of these factors in these areas are difficult to know for certain the further into the future the forecasting is to be done.

The starting point of the model selection research done in this thesis is therefore the assumption that the dependent variable takes the form of a second-order polynomial. The unknown quality and limited quantity of data sets make for a difficult problem domain. The goal of the research is to design the best model selection strategy to enable the best second-order polynomial forecasting model to be built in an automated manner.

1.4 Methods for Model Selection

Model selection is about comparing one model with another. In building a model, a good model selection criterion should serve as a stopping rule of the selection process when the best model has been found amongst all of the models considered.

For the task of automated model selection, the first test of the robustness of a criterion is in its ability to recover the true model that has generated sets of artificial data which have additional noise and irrelevant variables included in them. This test is done in Chapter 3 using a set of true models for the model selection criteria considered in this thesis.

If a model selection criterion manages to recover true models of artificially generated data, it has the potential to find a model that is consistent with the real data presented to it within the assumptions and constraints of the model selection strategy in place. This test is done in Chapter 5 in building tropical cyclone intensity forecasting model for the Atlantic basin using the atmospheric and climatological data sets seen to be influential to the tropical cyclone intensity data.

There are at least two categories of model selection criteria:

1. Criteria which require a separate test data set to decide on a model. These methods require the data set to be divided into 2 parts. The first set of data, the training data set, is used to calculate the parameters of the variables of a model. The second set of data, the test data set, is used to compare and make a choice between two models.

2. Complexity-penalised criteria. These methods require only one data set in deciding on a model. They penalise more complex models, hence seek to find the balance between the complexity of the model and the fit of the model to

the data. When comparing two models with different complexities (number of variables, order of polynomial, etc), these methods will choose the more complex model only when the improvement in the fit to the data outweighs the penalty attributed to the increase in model complexity.

In Chapter 2 of this thesis, model selection methods from both categories are outlined and a new complexity-penalised criterion based on the Minimum Message Length principle [19] for the task of second-order polynomial model selection is presented. The performance of the methods are tested using artificial data in Chapter 3.

1.4.1 Minimum Message Length Principle

Minimum Message Length principle takes the form of a two-part message. The first part represents the encoding of the model and the second part represents the encoding of the residual between the original data and the data calculated from the model. When comparing two models, the more complex model will have a longer first-part message, but may have a shorter second-part message, due to the improvement of its fit to the data. The criterion of a better model in MML principle is one which yields a shorter total two-part message. An MML model selection criterion avoids overfitting the data at hand by choosing the model with the minimum level of complexity and the maximum fit. This ensures that the fit to the data is due to the general patterns captured in the model, not the noise pertaining to the particular data set at hand. This way, the model chosen should perform as well in other sample data sets taken from the same population. In this sense, MML model selection criterion follows Occam's Razor principle, "It is vain to do with more when one could do with fewer".

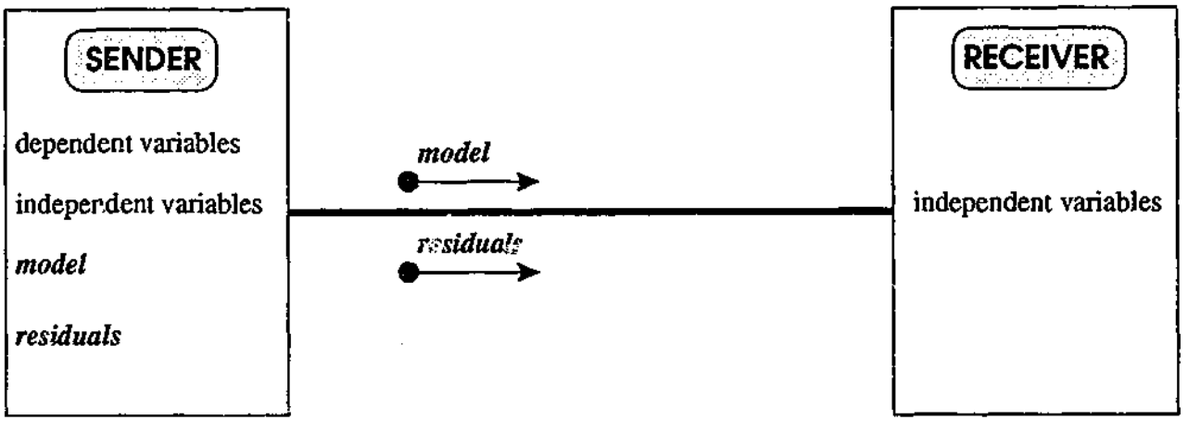


Figure 1.1: MML metaphor of sending data over a communication line for the problem of polynomial model selection. The task is to find an optimum way to send data of the dependent variables from the sender to the receiver. MML principle states that the cheapest way is by sending the optimum encoding of the model and the residual data. The model is developed from the independent variables and the residual data set is the difference between the model estimates and the real data of the dependent variable

From information theory perspective, Minimum Message Length (MML) principle takes the metaphor of sending data over a communication line between a sender and a receiver. For the specific problem of polynomial model selection, this metaphor can be explained with the assumptions that the sender has a data set of a dependent variable and sets of data of independent variables which are known to be influential to the dependent variable and the receiver has the data sets of the independent variables. So, the independent variables are the fixed common knowledge owned by both the sender and the receiver. The task is to find an optimum way to send the data of the dependent variables from the sender to the receiver. This metaphor is illustrated in Figure 1.1.

The most expensive way to get the data set to the receiver, in terms of the number of bits required, would be to send the encoding of every single data point verbatim. This is equivalent to what is known as a table lookup. The cheapest way would be by first developing a model from the independent variables on the sender side and

then sending the optimal encoding of the model and residual data (i.e. the difference between the estimates of the dependent variables calculated from the model and the real values of the dependent variables). The receiver can then calculate the true values of the dependent variables from the model and residual data sent.

1.4.1.1 Applications of The Minimum Message Length Principle

As outlined in [94], research in the Minimum Message Length principle has always been motivated by the need to solve real world problems. The article recorded that MML research started in 1962 when Chris Wallace wrote a computer program to classify grit particles found in cores drawn from oil wells as requested by an Australian geologist. The challenge to build a generic automated classification program led to the proposal of a two-part message as a measure of the suitability of a classification and hence the Minimum Message Length principle [29, 30]. To date, SNOB, the program built in 1966 [18], has evolved into an automated computer program which can quickly classify a large amount of data from a large number of discrete or continuous variables. SNOB is publicly available and widely used in modelling research.

Some real-world applications of the MML principle include learning the shape patterns of megalithic stone circles [56, 60], finding models for DNA sequences [31, 76, 75, 77], predictions of bushfire activity [26], models for footy tipping competition [1], learning simple grammars [88], and learning prolog programs [107]. A summary of existing applications of the MML principle can be found in [74].

An application of the MML principle to linear regression polynomial model selection problems has been discussed in [97]. The tasks in that thesis are in finding the

polynomial-order of the variables from artificial data generated from known models. In contrast, this thesis deals with the problem of finding variables and their combinations up to the second-order. To the best of my knowledge, this is the first time the MML principle is used to find polynomial models from real tropical cyclone data.

1.5 Optimisation Search Algorithms

The model selection criteria discussed in Section 2 are used to compare models. In the selection process, the models to be compared are to be found using an optimisation search algorithm. A model selection criterion is used as the objective function of the search algorithm. The task of an optimisation search algorithm is therefore to find, amongst all the possible models considered in the search space, a model for which the objective function yields the best/smallest value. This model is called the global minimum of the search space relative to the objective function used. Exhaustive search algorithms which consider every single model available in the search space will guarantee that the global minimum will be found. However, if the search space is large, implementations of these algorithms might not be feasible due to the time and the amount of computation resources required.

In the face of these constraints, research in this area focuses in finding algorithms which do not consider all the possible models and yet manage to find a global optimum, i.e. a model which is good enough for the problem domain being studied. Sometimes the speed in which this global optimum is found is also taken into consideration. One major criterion of a good optimisation search algorithm is in its ability to escape or jump away from local minima to converge to a global optimum.

The combination of a good optimation search algorithm and objective function is imperative in model selection tasks. A programmed optimisation search algorithm used for all of the objective functions outlined in Chapter 2 is given as a part of Chapter 3.

1.6 Tropical Cyclones

Tropical cyclones[2] are low pressure weather systems in the tropical seas, in which the atmospheric pressure decreases to a minimum value at the centre ("the eye"), with the winds blowing in a spiral inward toward this centre [8, p. 255]. This spiral rotation is sustained by the *coriolis force* [64, p. 368] which is caused by the earth's rotation. When viewed from space, the earth rotates clockwise in the southern hemisphere and counter-clockwise in the northern hemisphere. Therefore tropical cyclones have a well-defined clockwise wind rotations in the southern hemisphere and counter-clockwise wind rotations in the northern hemisphere. Because the coriolis force is zero at the equator, tropical cyclones only form beyond 5 degrees of latitude from the equator [9, p. 170].

The World Meteorological Organization (WMO) based in Geneva, Switzerland, divides the coastal areas affected by tropical cyclone around the globe into seven tropical cyclone basins. Table 1.1 lists the basins together with the areas covered by each basin and the organisation responsible for the collection of the cyclone data in each basin. Data for this thesis comes from the North Atlantic Basin therefore the forecasting model built is for the Atlantic basin. However, the model selection

[2]Tropical cyclones have different names: Hurricane in the United States, Typhoon in the Western Pacific north of the equator, Baquio in the Philippines, and Kamikaze in Japan

Table 1.1: The seven tropical cyclone basins and the areas covered by each basin and the organisation responsible for the collection of the cyclone data in each basin. *Source:* [39, p. 1.18]

Basin No.	Basin Name	Areal Extent	Principal Data Source
1	North Atlantic	North Atlantic Ocean, Caribbean Sea and Gulf of Mexico	National Hurricane Centre (NHC), Florida
2	Eastern North Pacific	North America to 180°E	National Hurricane Centre (NHC), Florida
3	Western North Pacific	West of 180°E, including South China Sea	Joint Typhoon Warning Centre (JTWC), Guam
4	North Indian	Bay of Bengal and Arabian Sea	Indian Meterological Department
5	Southwest Indian	South Indian Ocean west of 100°E	Meterological Service, Reunion
6	Southeast Indian/Australia	Southern Hemisphere 100 – 142°E	Bureau of Meteorology (BOM), Australia
7	Southwest Pacific/Australia	Southern Hemisphere	Regional Specialized Meteorological Centre (RSMC), Fiji

strategy proposed in this thesis can be used to build forecasting models for any of the basins.

1.6.1 Tropical Cyclone Intensity

Tropical cyclones are characterized by a strong thermally directed circulation with the rising of warm air near the center and the sinking of cooler air outside. The low-level inflow is directed down the radial pressure gradient, from higher towards lower pressure, which the outflow takes place at much higher level where the radial pressure gradient is very weak [64, p. 430]. Tropical cyclones breed over the warm tropical seas and do not form unless the sea-surface temperature is above 26.5°C [12, p. 2].

Tropical cyclones depend on a regular supply of abundant water vapour [9, p. 170]. The warm core of the tropical cyclone serves as a reservoir of potential energy which is continuously being converted into kinetic energy by the thermally directed

Table 1.2: Saffir-Simpson Scales [103]. Note: The storm surge height is the height above normal predicted tide level at the time the tropical cyclone crosses the coast [112]

Magnitude	Saffir-Simpson Scale	Central Pressure (hPa)	Maximum Wind Gust		Surge Height (m)
			(m/s)	(km/h)	
mild	1	> 990	20–30	75–110	0–1
moderate	2	970–985	35–40	125–160	1.5–2.5
severe	3	950–965	50–60	180–215	3–4
very severe	4	930–945	65–75	235–270	4.5–5.5
catastrophic	5	< 920	80–90	290–325	6–7

circulation [64, p. 430]. Unless disrupted by their surroundings, tropical cyclones will intensify until they reach a Maximum Potential Intensity (MPI). The variables seen to positively or negatively influence tropical cyclone intensity will be discussed in Chapter 4 in this thesis.

There are two ways of measuring tropical cyclone intensity. The first is by taking the average maximum sustained surface winds over 1 minute period and the second over 10 minute period. The longer the averaging period used, the lower the maximum wind speed for a cyclone of a given intensity [39]. Following Simiu and Scalon [34], the WMO applies a multiplication factor of 0.871 to convert between 1-min and 10-min winds.

Tropical cyclone intensity can be categorised using Saffir-Simpson Scale [103] shown in Table 1.2. The scale is based primarily on the central pressure to which the maximum wind speed and maximum storm surge height are approximately related.

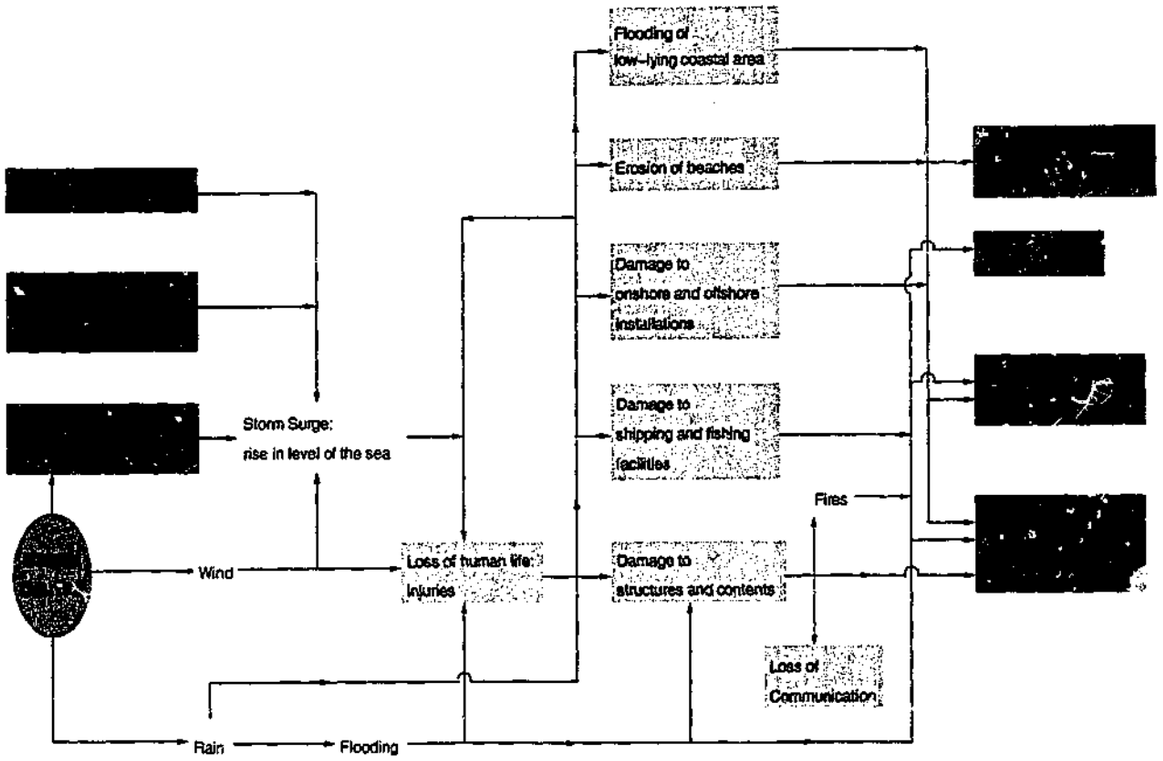


Figure 1.2: Potential damage caused by tropical cyclones. *Source:* Adapted from [8, p. 256]

1.6.2 Damage Caused by Tropical Cyclones

Tropical cyclones cause annual disasters resulting in the loss of lives and property particularly to communities over the coastal areas around the globe due to strong winds, flooding, and storm surges. Figure 1.2 illustrates the potential damage from tropical cyclones.

Effective mitigation measures to reduce the vulnerability to tropical cyclones can only be achieved by accurate and comprehensive assessment of the hazard levels and the vulnerability of the coastal communities [37]. Forecastings of the track and intensity of an incoming cyclone are the first most important steps of the assessment of the hazard levels. The assessment of the vulnerability of the communities at risk

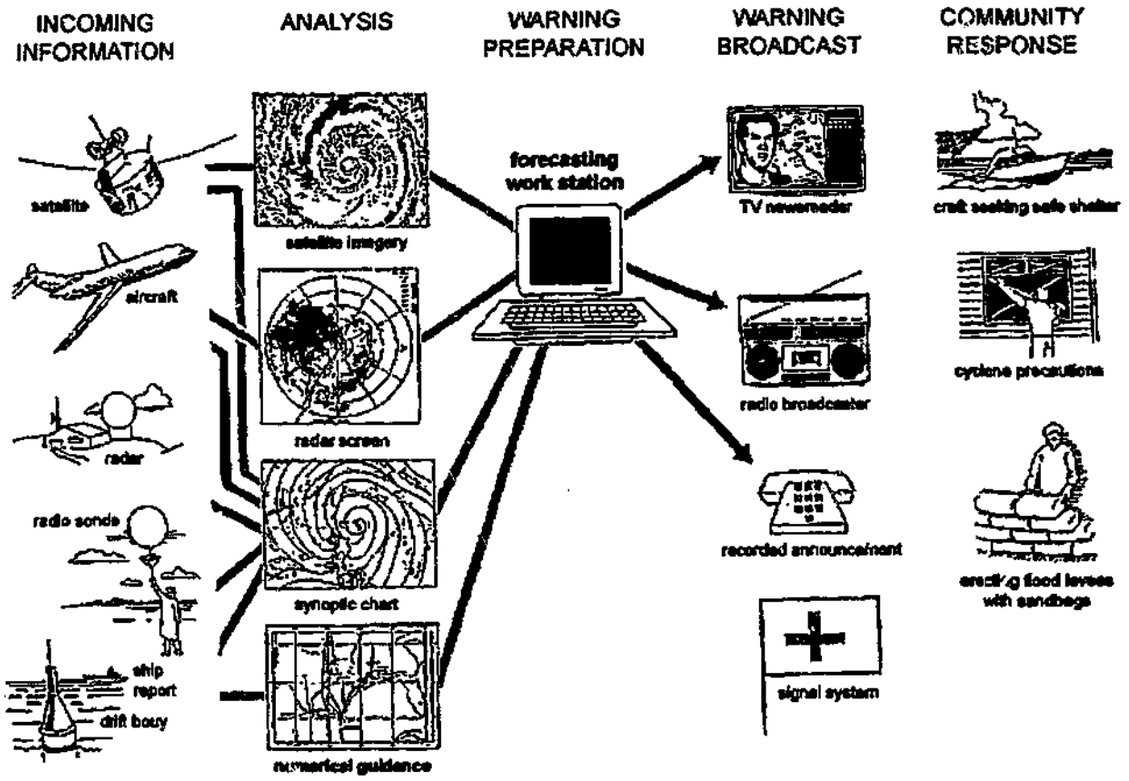


Figure 1.3: Schematic view of the tropical cyclone warning process. *Source:* [39, p. 6.4]

is done through up-to-date data recording of the population, properties, community infrastructure and facilities.

The loss due to tropical cyclones can be minimised with the implementation of effective warning system. Figure 1.3 gives a schematic view of the tropical cyclone warning process with the forecasting work station in the centre. False negative in the form of underforecasting results in the unpreparedness of the population, late evacuations and maximum loss of lives and property in the area affected. False positive in the form of overforecasting can be equally damaging due to the economic interruptions especially to shipping and fishing industries, coastal tourist areas, and off-shore oil and gas development plants.

1.6.3 Tropical Cyclone Intensity Forecasting

There are four stages in the life cyclone of a tropical cyclone [12, p. 2], namely the formative, immature, mature and decaying stages. The average life cyclone lasts about 9 days, some cyclones may last 20 days or more, and some may intensify explosively to the mature stage in less than 48 hours.

Despite its evident importance, tropical cyclone intensity forecasting is one of the least understood and the least researched areas in atmospheric science. This is due to the high degree of uncertainty in the physics behind tropical cyclone intensifications and the fact that most atmospheric scientists are not familiar with the various model selection methods studied by computer scientists.

At present, intensity forecasts are commonly made using manual pattern recognition method using satellite data, multiple linear regression models and subjective assessment of satellite imagery and environmental conditions. Whilst efforts have been made to incorporate more variables that might be influential to tropical cyclone intensity in building statistical forecasting models, the method used to build the models tends to remain the same, i.e. the least square multiple linear regression method using significance tests.

Forecasting models created for cyclones in the North Atlantic basin using the multiple linear regression method are SHIFOR (Statistical HurrIcane FORcasting) [10] , SHIFOR94 [24] and SHIPS (Statistical Hurricane Intensity Prediction Scheme) [82]. The multiple linear regression method used to build these models will be discussed in Chapter 2 in this thesis and the data sets used to build them will be discussed in Chapter 4.

1.7 Thesis Outline

This thesis proposes a new model selection criterion based on the Minimum Message Length principle, a programmed optimisation search algorithm and a new strategy to build a tropical cyclone intensity forecasting model using the combination of the best state-of-the-art model selection criteria. The data used to build SHIFOR and SHIFOR94, the forecasting models being used for the North Atlantic tropical cyclone basin, is used in this thesis, therefore the new models built are benchmarked against these two models.

Chapter 2 gives a structured overview of a wide range of criteria commonly used in the literature. In this chapter, the model selection criteria are divided into two categories, i.e. the criteria which need test data sets to decide on a model, and the complexity-penalised criteria. A novel complexity-penalised criterion based on the Minimum Message Length (MML) principle is proposed. All of the statistical tropical cyclone forecasting models available to-date have been built using the first category of criteria. This thesis explores the feasibility of using the complexity-penalised criteria, the second category of criteria to build tropical cyclone forecasting models. The performance of the new MML criterion is compared against most of the other complexity-penalised criteria and benchmarked against the existing forecasting models.

As explained in Section 1.5 above, the best way to find variables to be included in a model would be to use exhaustive search algorithms which examine every possible combination of variables. However, these algorithms are often prohibitive in terms of the demand in time and computing resources. A programmed optimisation non-backtracking search algorithm used in this thesis which has been designed to be less resource-intensive but cover more areas in the search space than a simple greedy

search algorithm is outlined as a part of Chapter 3. This enables the search to converge to a better result.

The experiments in Chapter 3 serve as a screening procedure to test the robustness of each of the complexity-penalised polynomial selection criteria outlined in Chapter 2. This chapter uses these criteria and the optimisation search algorithm outlined in the previous paragraph for the task of recovering true models from the artificially generated data in an automated manner. The experiments also test if the programmed optimisation search algorithm proposed does cover enough area in the search space to recover the true models from which artificial data sets have been generated.

The criteria that have passed the test using artificially generated data in Chapter 3 would need to be further tested in their ability to select forecasting models using real data. The forecasting models to be built in this thesis are for the purpose of tropical cyclone intensity forecasting. Chapter 4 gives an overview of tropical cyclones and research in tropical cyclone intensity forecasting. The kinds of data involved in the research and the data pre-processing methods employed to come up with potential regressors are outlined. This chapter serves as an overview of the variables used to build the tropical cyclone intensity forecasting models in Chapter 5.

Chapter 5 proposes a novel strategy in finding forecasting models using a combination of the polynomial model selection criteria that have passed the tests both in Chapter 3 and in Chapter 5. These criteria are used as the cost functions of the optimisation non-backtracking search algorithm which has been proposed and tested in Chapter 3. The strategy using a combination of these criteria has the ability to rank the performances of good models with increasing levels of complexity. It is found that whilst the strategy has the ability to select the best model from the data

at hand in an automated manner, it still provides an option for the experts to make the final informed decision on which of the good models will ultimately be used.

This chapter comes up with new tropical cyclone intensity forecasting models for cyclones in the Atlantic basin which are better in terms of performance criteria of parsimony and predictive ability than the benchmark models.

Results shown in Chapter 5 suggest that the behaviour of cyclones in the Atlantic basin has changed significantly between the periods observed. This means that the practice of partitioning of the data as training and test data sets based on a chronological order that has always been done in building tropical cyclone models is flawed since the training and test data sets do not capture the same features. This chapter shows a better way of partitioning the data which contributes to the discovery of better forecasting models than the benchmark models.

Chapter 6 further tests the new strategy proposed in Chapter 5 using sets of artificial data generated from the covariance matrix of the data used to build the tropical cyclone forecasting models discussed in that chapter. This test shows the minimum number of data needed for the model selection strategy to converge to the model from which the data is generated and the maximum level of noise it can sustain while still converging to the model. The experiments shows that if the model to be discovered indeed is the model that has generated the data, the new model selection strategy is able to recover the model from a very limited amount of data which contains high level of noise.

Chapter 8 gives the summary of the main contributions of this thesis to polynomial model selection research and tropical cyclone intensity forecasting research. The new method based on the Minimum Message Length principle and the new optimization search algorithm are summarized. Remarks on the results of the critical

examinations of the differences in performance of the various criteria when applied to artificial *vis-a-vis* to real tropical cyclone data are given. The novel model selection strategy using the combination of the four criteria proven to be the most robust of all criteria tested in Chapter 3 and the search strategy used is summarised. Final comments on the successful application of the strategy for building tropical cyclone intensity change forecasting model are given.

The new model selection strategy proposed in this thesis can be used for any domain data which can be represented as second-order polynomial models. Areas for future research in tropical cyclone intensity forecasting in terms of new types of data and new tropical cyclone basins to be explored is outlined in Chapter 7.

Chapter 2

Methods For Polynomial Model Selection

2.1 Introduction

This chapter describes the form of the polynomial models considered in this thesis and the various model selection methods employed in the experiments done in this thesis[1]. A new model selection method based on the Minimum Message Length principle is explained.

Section 2.2 gives the form of the standardized second-order polynomial regression models considered in this thesis. We restrict ourselves to linear second-order polynomials because the practical aim of this thesis is to find a method that will yield a better tropical intensity forecasting model than the existing models used in operation in the Atlantic Basin. Since the existing models are all in the linear

[1] An earlier version of parts of this chapter has been published in [42]

second-order polynomial form, we choose this form for ease of comparison. This does not mean that the methods used cannot be used for other forms of polynomials.

Section 2.3 provides descriptions of the methods considered in this thesis. Those methods belonging to the family of criteria that require division of the data sets into training and test data sets to decide on a model are outlined in Section 2.3.2. In particular, the least squares method used to build the existing tropical cyclone intensity forecasting models used in operation in the Atlantic basin is outlined. Those methods belonging to the family of complexity-penalised criteria are given in Section 2.3.3. All of the experiments done in this thesis are done using methods belonging to this family. The new method built based on the Minimum Message Length principle is given in Section 2.3.3.1 on page 32.

2.2 Second-order Polynomial Models

Polynomial regression concerns with the task of estimating the value of a target variable from a number of regressors/independent variables. The standardized second-order polynomial regression models considered in this thesis typically take the form

$$y_n = \sum_{p=1}^{P} \zeta_p u_{np} + \sum_{p=1}^{P} \sum_{q \geq p}^{P} \zeta_{pq} u_{np} u_{nq} + r_n \Leftrightarrow y_n = \sum_{k=1}^{K} \beta_k x_{nk} + r_n \qquad (2.1)$$

where for each data item n:

y_n : the n^{th} value of the target variable y

u_{np} : the n^{th} value of the single regressor/independent variable u_p

ζ_p : the coefficient of single regressor u_p

ζ_{pq} : the coefficient of compound regressor $u_p \times u_q$

x_{nk} : the n^{th} value of the regressor x_k; $x_{nk} = u_{np}$ or $x_{nk} = u_{np}u_{nq}$; $q \geq p$

β_k : the coefficient of regressor x_k; $K = 2P + P!/2!(P-2)! = (P^2 + 3P)/2$

r_n : the n^{th} value of the noise/residual/error term r

The values of the error term r is assumed to be uncorrelated, normally and independently distributed $r_n \sim NID(0, \sigma^2)$. For a given set of x_k, this assumption causes y to have the same normal distribution form and variance as r, that is, $y_n \sim N(\sum_{k=1}^{K} \beta_k x_{nk}, \sigma^2)$.

The standardized values of each dependent variable are calculated using the following formula

$$w_n = \frac{y_n - \bar{y}}{s_y} \tag{2.2}$$

where:

w_n : a standardized value of the n^{th} value of the dependent variable y

y_n : the n^{th} value of the dependent variable y

$\bar{y}$: the sample mean of the dependent variable y

s_y : the sample standard deviation of the dependent variable y

The standardized values of each regressor/independent variable are calculated using a similar formula

$$u_{np} = \frac{z_{np} - \bar{z}_p}{s_p} \tag{2.3}$$

where:

u_{np} : a standardized value of regressor p

z_{np} : the original value of regressor p

$\bar{z}_p$: the sample mean of regressor p

s_p : the sample standard deviation of regressor p

The standardized variance of the product of variables is also kept to unity by standardizing the product of the standardized single variables in the same way it is done for the single variables.

The subsets of potential regressors that may form the second-order polynomial models are chosen from the model space using the optimisation search method outlined in Section 3.2.2 on page 57. The coefficients for the regressors of each model can be calculated using using Equation 2.15 on page 29 for ordinary least squares method and Equation 2.40 on page 42 for a modified least squares method, which still yields a set of coefficients which represents uncorrelated regressors when the covariance matrix of the regressors is ill-conditioned (i.e. having small eigenvalues). For this reason, Equation 2.40 will be used in this thesis for all of the model selection criteria outlined in Section 2.3.

2.3 Model Selection Criteria

The sample data set available to be used for building models that can explain the structure of the data population is usually just a small proportion of the population. Hence, the model to be built should capture only the general properties found in

the sample data set and not the specific properties or noise pertinent only to the particular sample data set at hand.

For data sets generated from a multi-variate second-order polynomial, for example, the structure of the data is expressed in some combination of variables included in the model. If the data contains some degree of noise and exogenous variables, the model selection process should ensure that these are not included in the model. This way, the model can be used to explain the structure of any other data set sampled from the same population. In other words, the model should generalise well on unseen data.

Two polynomial models can be compared using various methods which can be categorised into two main categories: criteria that require a test data set to decide on a model and complexity-penalised criteria. Methods in the first category decide between models based on model performance for test data which was not used to create the model. Methods in the second category decide on models based on their ability to describe the data that was used to derive the model. The two categories will be discussed in the following sections.

2.3.1 Probability and Code Length

The task of a model selection method is to select the model with the highest posterior probability given a data set. Following Shannon's information theory [14], a probability can be conveniently represented as a *code length* by taking the negative logarithm of the probability. This way, the range of the probability values from 0 up to 1 is replaced with the range of the code length from infinity down to 0. This

means, maximising probability corresponds to minimising code length.

$$L = -\log P(D) \qquad \text{where} \quad L\epsilon(0,\infty)$$
$$\text{for} \quad P(D)\epsilon(0,1)$$

Code lengths are measured in *bits* when the logarithm is taken to base 2 and in *nits* when natural logarithms are used. For convenience, some method comparisons done in this thesis are done in the code length representations of the measurements.

2.3.1.1 Maximum Likelihood Method

Maximum likelihood is a method to find the best fit of a model to a data set. This is represented in the probability of the data given the model. For the polynomial model represented in Equation 2.1, the model consists of K variable coefficients β_k and the data consists of N target data points. The maximum likelihood estimate of the model is then calculated from the residuals of the fitted model to the data, i.e. $r_n = y_n - \sum_{k=1}^{K} \beta_k x_{nk}$. With the assumption that $r_n \sim NID(0, \sigma^2)$ the maximum likelihood estimate of the model becomes its least squares estimate.

Since the residuals are assumed to be uncorrelated, normally and independently distributed as explained in Section 2.2 above, the likelihood is equal to the joint probability of the residuals given the model which is simply the product of the probability of the residuals of each data point given in the following equation. For encoding purposes, as outlined in [16, 98], each data point in the sample is to be measured to any accuracy $\pm\delta/2$[2] where $\delta << \sigma$. The area occupied by δ under the probability density distribution curve of the residuals is then approximated by mid-point rule as a rectangular area with width δ and length $\frac{1}{\sigma\sqrt{2\pi}} e^{-(y_n - \sum_{k=1}^{K} \beta_k x_{nk})^2/2\sigma^2}$

[2]The accuracy term $\pm\delta/2$ is discussed in more detail in Section 2.7 of [63]

$$P(data|model) = P(y|\sigma, \{\beta_k\}) \tag{2.4}$$

$$= \prod_{n=1}^{N} \frac{\delta}{\sigma\sqrt{2\pi}} e^{-(y_n - \sum_{k=1}^{K} \beta_k x_{nk})^2/2\sigma^2} \tag{2.5}$$

$$\tag{2.6}$$

Since δ is constant and hence has no effect in the comparisons of the code lengths of models, it is usually dropped from the code length representation of Equation 2.5 given below.

$$\begin{aligned} L(D|\theta) &= -\log P(data|model) && (2.7)\\ &= -\log P(y|\sigma, \{\beta_k\}) && (2.8)\\ &= -\log\left[\prod_{n=1}^{N} \frac{\delta}{\sigma\sqrt{2\pi}} e^{-(y_n - \sum_{k=1}^{K} \beta_k x_{nk})^2/2\sigma^2}\right] && (2.9)\\ &= -N\log\delta + \frac{N}{2}\log 2\pi + N\log\sigma + \sum_{n=1}^{N} \frac{r_n^2}{2\sigma^2} && (2.10) \end{aligned}$$

where:

r_n : the error of the model of the n^{th} data item

$r_n = y_n - \sum_{k=1}^{K} \beta_k x_{nk}$

The first two terms in Equation 2.10 are constant over the values of the data we consider and play no role in the construction of the model, hence can be omitted.

The variable coefficients of the model are the values that minimise Equation 2.10 with respect to β_k. They are calculated by taking the partial derivative of Equation 2.10 with respect to β_k:

$$\frac{\partial(L(D|\theta))}{\partial\beta_k} = 0 \tag{2.11}$$

$$\frac{1}{2\sigma^2}\left(2\sum_{n=1}^{N} r_n(-x_{nk})\right) = 0 \tag{2.12}$$

$$\frac{1}{\sigma^2}\left(-\sum_{n=1}^{N} r_n x_{nk}\right) = 0 \tag{2.13}$$

Resubstituting $r_n = y_n - \sum_{k=1}^{K} \beta_k x_{nk}$ into Equation 2.13 we get

$$\begin{aligned} \sum_{n=1}^{N} x_{nk}\left(y_n - \sum_{k=1}^{K}\beta_k x_{nk}\right) &= 0 \\ \Leftrightarrow \sum_{n=1}^{N}\sum_{k=1}^{K} x_{nk}x_{nk}\beta_k &= \sum_{n=1}^{N}\sum_{k=1}^{K} x_{nk}y_n \end{aligned} \tag{2.14}$$

To calculate the estimates of the coefficients, $\hat{\beta}_1, \ldots, \hat{\beta}_K$, we represent Equation 2.15 in a matrix format, where for brevity, $(\underset{N\times K}{\mathbf{X}})^{\mathbf{T}}$ is written as $\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}$.

$$\begin{aligned} (\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}})\ \underset{K\times 1}{\hat{\beta}} &= \underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times 1}{\mathbf{y}} \\ \Leftrightarrow \underset{K\times 1}{\hat{\beta}} &= (\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}})^{-1}\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times 1}{\mathbf{y}} \end{aligned} \tag{2.15}$$

The spread of the distribution of the data y represented by σ is calculated by taking the partial derivative of Equation 2.10 with respect to σ:

$$\frac{\partial(L(D|\theta))}{\partial\sigma} = 0 \tag{2.16}$$

$$\frac{N}{\sigma} - \frac{1}{\sigma^3}\sum_{n=1}^{N} r_n^2 = 0 \tag{2.17}$$

Hence, $\hat{\sigma}$, the estimate of the spread, σ, is

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} r_n^2 \tag{2.18}$$

2.3.2 Criteria Which Require a Test Data Set to Decide on a Model

These criteria choose a model which best fits the data at hand. For each potential combination of variables considered for the model given in Equation 2.1 on page 23, the variable coefficients can be calculated from a sample data set, called the training data, using the maximum likelihood approach outlined in Section 2.3.1.1 which, if the residuals are assumed to be normally distributed, is the same as the least squares method. Comparison between models is then done based on the performance of each model on test data, i.e. a separate data set that is not used to calculate the coefficients. One performance criterion used is the average squared error of the model on the test data, $(\sum_{n=1}^{N} r_n^2)/N = \sum_{n=1}^{N}(y_n - \sum_{k=1}^{K}\beta_k x_{nk})^2/N$. Other methods belonging to this category, among others, are Predictive sum of squares (PRESS) [28], Adjusted Coefficient of Determination (R_{adj}) [84], F-ratio Statistics [87], and Mallows' C_p [13].

All of the statistical tropical cyclone intensity forecasting models in operational use for the Atlantic basin have been built using the least squares method outlined in Section 2.3.1.1 on test data. Two of the models, namely Statistical Hurricane Intensity FORcasting (SHIFOR) [10], SHIFOR94 (a modification of SHIFOR) [24] are used as benchmark models for the proposed forecasting model found in the experiments done in Chapter 5 using some of the complexity-penalised criteria outlined in the next section.

2.3.3 Complexity-Penalised Criteria

It is understood that the more complex the combination of variables included in the model, the more the model fits the particular sample data set at hand. So, to ensure that the model does not overfit the sample data set to the point that it performs poorly on unseen data, the complexity-penalised criteria seek to balance between model complexity and quality of fit. These criteria typically manifest in cost functions that sum up the reward for fitting the data and the penalty for increased model complexity.

Because of this balancing mechanism, it is possible to compare two models with different complexities using only the training data, the data used to derive the model. A more complex model is preferred only when the fit to the data outweighs the penalty for increased complexity. It is because of the existence of the penalty term for model complexity that these criteria are categorised as complexity-penalised criteria.

Some of the most commonly used complexity-penalised criteria listed below are discussed in the subsequent sections. A new model selection criterion based on

the Minimum Message Length (MML) principle applicable to finding second-order polynomial models is proposed in Section 2.3.3.1.

- Minimum Message Length (MML) Criterion [19]
- Predictive Minimum Message Length (PMDL) Criterion [59]
- Minimum Description Length (MDL) Criterion [59]
- Akaike's Information Criterion (AIC) [46]
- Corrected Akaike's Information Criterion (CAICF) [50]
- Bayesian Information Criterion (BIC) [38]
- Structural Risk Minimization (SRM) [109]

2.3.3.1 A New Minimum Message Length (MML) Criterion

From a Bayesian perspective, the MML principle states that amongst all of the competing models under consideration, the best model is one that yields the highest posterior probability by maximizing the product of the prior probability of the model and the probability of the data given the model. Since in this problem, a model has real-valued coefficients, a conventional Bayesian approach can ascribe only a prior probability density to a specific model. MML instead assigns a finite prior probability to a model by considering the code length needed to decribe it.

As explained in Section 1.4.1 on page 8 and Section 2.3.1 on page 26, following information theory, the best model according to the MML principle is one that gives the shortest two-part message $L(D)$. The first part of the message $L(\theta)$ shows the cost of encoding the model θ and the second part $L(D|\theta)$ shows the cost of

encoding the sample data D given the model θ. This generic MML principle can be represented in the following equation.

$$L(D) = L(\theta) + L(D|\theta) \tag{2.19}$$

$$\Leftrightarrow -\log P(D) = -\log P(\theta) - \log P(D|\theta) \tag{2.20}$$

In the search for the best model, the MML principle dictates that a new more complex model will be seen as a better model than the model at hand if it results in a shorter two-part message $L(D)$. This can only be achieved when the increase in model complexity reflected in the increase in the message length that encodes the model $L(\theta)$ is less than the decrease in the message length that encodes the data given the model $L(D|\theta)$.

The first part of the message $L(\theta)$, the cost of encoding the model θ, is composed of two parts. The first is the cost of encoding the model structure (i.e. which combination of variables makes the model) L_s and the second is the cost of encoding the model parameters L_p. Hence the total message length shown in Equation 2.20 becomes

$$L(D) = L_s + L_p + L(D|\theta) \tag{2.21}$$

This thesis proposes a new formula for estimating the cost of encoding the model $L(\theta)$. This new formula is specific to the particular model selection problem considered in this thesis, i.e. the second-order polynomial models, where the models may be formed using single and products of independent variables.

Taking the metaphor of sending data over a communication line between a sender and a receiver (illustrated in Figure 1.1 on page 9), this thesis proposes that the cost of sending the model structure L_s comprises three parts listed below. The new formula has a prior expectation that it is more expensive to send compound variables than it is to send single variables.

1. the cost of sending the single variables
2. the cost of sending the compound variables (i.e. products of variables)
3. the cost of sending the combination of single and compound variables

The cost of sending the model structure L_s is thus:

$$L_s = -\log h(\nu, j) - \log h(\xi, l) - \log \frac{1}{{}_JC_j \; {}_LC_l} \tag{2.22}$$

where:

L_s : the cost of encoding the model structure

$h(\nu, j)$: a suitable prior probability function on the integer range $0 \ldots J$ for choosing j number of single variables

J : the maximum number of single variables

$h(\xi, l)$: a suitable prior probability function on the integer range $0 \ldots L$ for choosing l number of compound variables;

L : the maximum number of compound variables

${}_JC_j$: the number of possible combinations of j single variables taken out of J total number of single variables

${}_LC_l$: the number of possible combinations of l compound variables taken out of L total number of compound variables

It is then proposed that for the costs of sending the number of single and compound variables, $-\log h(\nu, j)$ and $-\log h(\xi, l)$ respectively, the prior probability functions $h(\nu, j)$ and $h(\xi, l)$ should follow the geometric series. It is assumed that a single variable is more likely to be chosen than a compound variable, hence it is cheaper to send a single variable than a compound variable. For this reason, the term ν is given a bigger value than ξ in the experiments done in this thesis.

$$h(\nu, j) = \frac{\nu^j(1-\nu)}{1-\nu^{J+1}} \tag{2.23}$$

$$h(\xi, l) = \frac{\xi^l(1-\xi)}{1-\xi^{L+1}} \tag{2.24}$$

where: $\nu > \xi$

For the cost of sending the combination of single and compound variables, Peter Tischer (personal communication) has pointed out that having calculated the costs of sending the numbers of single and compound variables in Equations 2.23 and 2.24, we need to use the information of the numbers of available single and compound variables in the formula given below

$${}_JC_j \; {}_LC_l = \frac{J!}{j!(J-j)!} \; \frac{L!}{l!(L-l)!} \tag{2.25}$$

where:

${}_JC_j$: the number of possible combinations of j single variables
taken out of J total number of single variables

${}_LC_l$: the number of possible combinations of l compound variables
taken out of L total number of compound variables

Equation 2.25 was however not used in the experiments in this thesis because it was discovered in hindsight. Instead, the experiments reported in Chapters 3, 5 and 5 were done using the following equation

$$_KC_k = \frac{K!}{k!(K-k)!} \tag{2.26}$$

where:

$_KC_k$: the number of possible combinations of k variables
taken out of K total variables where: $K = J + L$ and $k = j + l$

The full derivation of Equation 2.22 is given in Appendix A on page 145.

Whilst Equation 2.26 is certainly a better formula than Equation 2.25, Equation 2.25 is a good enough encoding to have performed well in the experiments in the model selection tasks outlined in Chapter 3 (i.e. using artificial data) and in Chapter 6 (i.e. using artificial data generated from the covariance matrix of a set of real data).

The experiments using the real tropical cyclone data outlined in Chapter 5 show that the MML formula which incorporates Equation 2.25 may have caused the search strategy (outlined in Chapter 3), when using MML as the cost function, to stop at a less complex model with longer message length than the better but more complex model model. Incorporation of Equation 2.26 instead of Equation 2.25 in the MML formula shown in Equation 2.21 will result in a shorter message length for a model, enabling the MML method to choose more complex models which fit the data better.

To calculate the cost of sending the model parameters, L_p, we follow Wallace and Freeman [19]. The formulae to calculate L_p outlined below has also been used in [20], [21] and [22] for causal models.

It is assumed that the sender and receiver of the message have some prior knowledge/expectation about the possible models with real-valued parameters, $\theta = (\beta_k, \sigma)$. Taking the assumption that the parameters are normal distributed, $N(0, \alpha^2\sigma^2)$, and a prior for σ is proportional to $\frac{1}{\sigma}$, the prior probability density $h(\theta)$ on the space of possible models is

$$\begin{aligned} h(\theta) &= prior(\sigma, \{\beta_k\}) \\ &= prior(\sigma) prior(\{\beta_k|\sigma\}) \\ &= \frac{1}{\sigma} \prod_{k=1}^{K} \frac{1}{\alpha\sigma\sqrt{2\pi}} e^{\frac{-\beta_k^2}{2\alpha^2\sigma^2}} \end{aligned} \tag{2.27}$$

where α is a hyper-parameter reflecting the *a priori* expected strength of causal effects relative to unexplained variation [21]. The significance of α is explained later on page 42.

The adoption of a discrete message/code string of length L for some model θ is equivalent to regarding θ as having a prior probability of 2^{-L}. Prior probability 2^{-L} is a discrete probability, not the probability density function $h(\theta)$. For this reason, the MML principle assigns to θ a prior probability $h(\theta) \times v(\theta)$ where $v(\theta)$ is the volume of a region of the search space which includes θ and other models so close to θ that the data cannot be expected to distinguish among them.

As shown in [19] the whole message length shown in Equation 2.20 is minimised when $v(\theta)$ is chosen to be proportional to

$$\frac{1}{\sqrt{\mathbf{I}(\theta)}} \tag{2.28}$$

where $\mathbf{I}(\theta)$ is the Fisher information associated with the real-valued parameters of the model θ [19]. The Fisher information is the determinant of the expected second

derivative of the negative log likelihood function, $-\log P(D|\theta)$, given in Equation 2.10 on page 28. After the differentiations done in Appendix B on page 148, the Fisher information for the second order polynomial models under consideration in this thesis takes the form

$$\mathbf{I}(\theta) = \mathbf{I}(\sigma, \{\beta_k\}) = 2N\sigma^{-2(K+1)} |\underset{N\times K}{\mathbf{X}^{\mathrm{T}}} \underset{N\times K}{\mathbf{X}}| \tag{2.29}$$

where:

$|\underset{N\times K}{\mathbf{X}^{\mathrm{T}}} \underset{N\times K}{\mathbf{X}}|$: the determinant of the covariance matrix of the independent variables x_k for $k = 1, \ldots, K$

N : the number of data items

K : the number of independent variables

The effect of quantization of $v(\theta)$ (i.e. the discretization of θ) in forming the optimum code [62, pp.59–61] is given in the last three geometric terms of L_p in Equation 2.30. These terms result in an increase in the message length.

$$\begin{aligned} L_p &= -\log \frac{h(\theta)}{\sqrt{\mathbf{I}(\theta)}} - \frac{1}{2}(K+1)\log 2\pi + \frac{1}{2}log(K+1)\pi - 1 \\ &= \frac{1}{2}\log \mathbf{I}(\theta) - \log h(\theta) - \frac{1}{2}(K+1)\log 2\pi + \frac{1}{2}log(K+1)\pi - 1 \end{aligned} \tag{2.30}$$

Substituting Equations 2.27 and 2.29 into Equation 2.30 gives

$$\begin{aligned}
L_p &= \frac{1}{2}\log(2N\sigma^{-2(K+1)}\,|\mathbf{X}^{\mathrm{T}}\mathbf{X}|) - \log\left(\frac{1}{\sigma}\prod_{k=1}^{K}\frac{1}{\alpha\sigma\sqrt{2\pi}}e^{\frac{-\beta_k^2}{2\alpha^2\sigma^2}}\right) \\
&\quad \frac{1}{2}(K+1)\log 2\pi + \frac{1}{2}log(K+1)\pi - 1 \\
&= \frac{1}{2}\left(\log 2N - 2(K+1)\log\sigma + \log|\mathbf{X}^{\mathrm{T}}\mathbf{X}|\right) - \log\frac{1}{\sigma} - \sum_{k=1}^{K}\log\left(\frac{1}{\alpha\sigma\sqrt{2\pi}}e^{\frac{-\beta_k^2}{2\alpha^2\sigma^2}}\right) \\
&\quad -\frac{1}{2}(K+1)\log 2\pi + \frac{1}{2}log(K+1)\pi - 1 \\
&= \frac{1}{2}\log 2N - (K+1)\log\sigma + \frac{1}{2}\log|\mathbf{X}^{\mathrm{T}}\mathbf{X}| \\
&\quad + \log\sigma + \frac{K}{2}\log 2\pi + K\log\alpha + K\log\sigma + \sum_{k=1}^{K}\frac{\beta_k^2}{2\alpha^2\sigma^2} \\
&\quad -\frac{1}{2}(K+1)\log 2\pi + \frac{1}{2}log(K+1)\pi - 1 \\
&= \frac{1}{2}\log 2N + \frac{1}{2}\log|\mathbf{X}^{\mathrm{T}}\mathbf{X}| + \frac{K}{2}\log 2\pi + K\log\alpha + \frac{1}{2\alpha^2\sigma^2}\sum_{k=1}^{K}\beta_k^2 \\
&\quad -\frac{1}{2}(K+1)\log 2\pi + \frac{1}{2}log(K+1)\pi - 1 \qquad (2.31)
\end{aligned}$$

The second part of the message length, the cost of encoding the data given the model $L(D|\theta)$ is simply the likelihood function given in Equation 2.10 on page 28.

Thus by substituting Equations 2.22, 2.31 and 2.10 into Equation 2.21, the total message length becomes

$$\begin{aligned}
L(D) &= L_s + L_p + L(D|\theta) \\
\Leftrightarrow L(D) &= -\log h(\nu, j) - \log h(\xi, l) - \log \frac{1}{{}_K C_k} \\
&\quad + \frac{1}{2}\log 2N + \frac{1}{2}\log|\mathbf{X}^{\mathrm{T}}\mathbf{X}| + \frac{K}{2}\log 2\pi + K\log\alpha + \frac{1}{2\alpha^2\sigma^2}\sum_{k=1}^{K}\beta_k^2 \\
&\quad - \frac{1}{2}(K+1)\log 2\pi + \frac{1}{2}log(K+1)\pi - 1 \\
&\quad + \frac{N}{2}\log 2\pi + N\log\sigma + \sum_{n=1}^{N}\frac{r_n^2}{2\sigma^2} \\
&= -\log h(\nu, j) - \log h(\xi, l) - \log \frac{1}{{}_K C_k} \\
&\quad \frac{1}{2}\log 2N + \frac{1}{2}\log|\mathbf{X}^{\mathrm{T}}\mathbf{X}| + \frac{K+N}{2}\log 2\pi + K\log\alpha + N\log\sigma \\
&\quad + \frac{1}{2\sigma^2}\left(\sum_{n=1}^{N} r_n^2 + \sum_{k=1}^{K}\frac{\beta_k^2}{\alpha^2}\right) - \frac{1}{2}(K+1)\log 2\pi + \frac{1}{2}log(K+1)\pi - 1
\end{aligned} \tag{2.32}$$

where: $h(\nu, j)$, $h(\xi, l)$ and $\frac{1}{{}_K C_k}$ are respectively given in Equations 2.23, 2.24 and 2.26. As given in Equation 2.10, $r_n = y_n - \sum_{k=1}^{K}\beta_k x_{nk}$

We examine the partial derivatives of Equation 2.32 with respect to σ and β_k to find the values of these parameters that will minimise the total message length.

The partial derivative of Equation 2.32 with respect to σ is

$$\begin{aligned}
\frac{\partial L(D)}{\partial \sigma} &= 0 \\
\Leftrightarrow \frac{N}{\sigma} - \frac{1}{\sigma^3}\left(\sum_{n=1}^{N} r_n^2 + \sum_{k=1}^{K}\frac{\beta_k^2}{\alpha^2}\right) &= 0
\end{aligned} \tag{2.33}$$

We therefore obtain $\hat{\sigma}$, the estimate of σ

$$\hat{\sigma}^2 = \frac{1}{N}\left(\sum_{n=1}^{N} r_n^2 + \sum_{k=1}^{K} \frac{\beta_k^2}{\alpha^2}\right) \tag{2.34}$$

The partial derivative of Equation 2.32 with respect to β_k is

$$\begin{aligned} \frac{\partial L(D)}{\partial \beta_k} &= 0 && (2.35)\\ \Leftrightarrow \quad \frac{1}{2\sigma^2}\left(2\sum_{n=1}^{N} r_n(-x_{nk}) + 2\frac{\beta_k}{\alpha^2}\right) &= 0 \\ \Leftrightarrow \quad \frac{1}{\sigma^2}\left(-\sum_{n=1}^{N} r_n x_{nk} + \frac{\beta_k}{\alpha^2}\right) &= 0 \\ \Leftrightarrow \quad \frac{\beta_k}{\alpha^2} - \sum_{n=1}^{N} r_n x_{nk} &= 0 && (2.36) \end{aligned}$$

Substituting r_n with $y_n - \sum_{k=1}^{K} \beta_k x_{nk}$, Equation 2.36 becomes

$$\frac{\beta_k}{\alpha^2} - \sum_{n=1}^{N}\left(y_n - \sum_{k=1}^{K} \beta_k x_{nk}\right) x_{nk} = 0 \tag{2.37}$$

Representing Equation 2.37 in matrix format, we get

$$\begin{aligned} \frac{1}{\alpha^2}\underset{K\times 1}{\hat{\beta}} - \underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times 1}{\mathbf{y}} + \left(\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}}\right)\underset{K\times 1}{\hat{\beta}} &= 0 \\ \Leftrightarrow \quad \left(\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}} + \frac{1}{\alpha^2}\underset{K\times K}{\mathbf{I}}\right)\underset{K\times 1}{\hat{\beta}} &= \underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times 1}{\mathbf{y}} \\ \Leftrightarrow \quad \underset{K\times 1}{\hat{\beta}} &= \left(\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}} + \frac{1}{\alpha^2}\underset{K\times K}{\mathbf{I}}\right)^{-1}\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times 1}{\mathbf{y}} \end{aligned} \tag{2.38}$$

where:

$\underset{K\times 1}{\hat{\beta}}$: the row matrix of K parameters β_k

$\underset{N\times K}{\mathbf{X}}$: the matrix of K independent variables for N data items

$(\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}})$: the covariance matrix of K independent variables

α : a hyper-parameter reflecting the *a priori* expected strength of causal effects relative to unexplained variation [21]

$\underset{K\times K}{\mathbf{I}}$: the identity matrix K × K

$\underset{N\times 1}{\mathbf{y}}$: the row matrix of the dependent variable y for N data items

Hence from Equations 2.34 and 2.38, the parameter estimates are:

$$\hat{\sigma}^2 = \frac{1}{N}\left(\sum_{n=1}^{N} r_n^2 + \sum_{k=1}^{K}\frac{\beta_k^2}{\alpha^2}\right) \tag{2.39}$$

$$\underset{K\times 1}{\hat{\beta}} = (\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}} + \frac{1}{\alpha^2}\ \underset{K\times K}{\mathbf{I}})^{-1}\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times 1}{\mathbf{y}} \tag{2.40}$$

where $r_n = y_n - \sum_{k=1}^{K}\beta_k x_{nk}$.

As in [21], the experiments in this thesis uses $\alpha^2 = 1$. Since no one knows how much unexplained variation play a role in the data at hand (unless the data is generated from a true model with no noise), the value $\alpha^2 = 1$ is seen as a reasonably compromise. It says that it is assumed that the strength of the causal effect (of the known variables) is approximately the same as the strength of the effect of the unexplained variation (of the unknown variables).

With the use of the term $\frac{1}{\alpha^2}$ in Equation 2.40 some people may think that we have been using ridge regression [3, 5, 4] shown in the equation below

$$\underset{K\times 1}{\hat{\beta}} = (\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}} + k\ \underset{K\times K}{\mathbf{I}})^{-1}\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times 1}{\mathbf{y}} \tag{2.41}$$

It turns out to be the case, with a few comments. The origin of the term α in Equation 2.40 is different from that of the term k in Equation 2.41. The term α in Equation 2.40 is directly derived from the use of a prior for the parameters shown in Equation 2.27 on page 37. The term k in Equation 2.41, on the other hand, is introduced as a way to eliminate the effect of collinearity/correlations among the independent variables in solving a regression problem. The way to eliminate the effect of collinearity among the independent variables used in this thesis is by using orthogonal transformation of independent variables outlined in Section 3.3.1 on page 60.

The MML principle gives the best explanation of why one should use the additional term k in ridge regression, that is because we need to encode the parameters necessitating the use of a prior for them, hence the need to use the additional term $k = \frac{1}{\alpha^2}$. Another logical explanation for the advantage of using the additional term $k = \frac{1}{\alpha^2}$ in Equation 2.40 is that the equation will yield a set of coefficients $\underset{K\times 1}{\hat{\beta}}$ which represents a set of uncorrelated regressors even if the matrix $\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}}$ is ill-conditioned, (i.e. having small eigenvalues). Because of these observed advantages, Equation 2.40 are used for all of the model selection criteria used in the experiments in these thesis. This is despite the fact that all of the other model selection criteria outlined in this thesis originally use the Maximum Likelihood method to calculate the parameter estimates of a model.

2.3.3.2 Predictive Minimum Description Length (PMDL) Criterion

Rissanen [57, 59] introduced the concept of *Stochastic Complexity* to measure the amount of uncertainty in data. Stochastic Complexity is a the generic term used for Predictive Minimum Description Length (PMDL) and Minimum Description Length

(MDL) criteria. PMDL is given in this section and MDL is given in the next section. Interestingly enough, the term Minimum Description Length (MDL) is more widely used to refer to Stochastic Complexity (SC) theory.

Minimum Message Length and Stochastic Complexity are the two prominent learning theories which use the minimum encoding principle from Information theory. The similarities and differences of MML and SC are discussed in [99] and [97].

The two basic concepts in the Stochastic Complexity theory [59] are a parametric class of probabilistic models where the number of parameters may range over all natural numbers and a utility function whose minimised value is its stochastic complexity. Rissanen gives three different interpretations of stochastic complexity:

1. Stochastic complexity as the greatest lower bound for the description length which can be taken as a formal measure of the amount of randomness in the data, defined relative to the selected class of models. This randomness comes both from the sampling uncertainty and the uncertainty due to the distribution of the data

2. If stochastic complexity, $I(x)$, denotes the infinum of the code lengths from the data x, relative to a class of models, then $P(x) = 2^{-I(x)}$ gives the probability distribution that is the most likely explanation of the data that can be obtained with the same class.

3. In forecasting tasks, stochastic complexity is the shortest code length defined using the accumulated prediction errors.

Rissanen [59, page 226] defines the stochastic complexity of the data x of length n, relative to a class of distributions, as follows: let $\{f(x|k,\theta)|\theta = (\theta_0,\ldots,\theta_k), k = 0,1,\ldots\}$ denote a parametric class of distributions represented by densities, such

that for each $f(\cdot|k,\theta)$ the marginality conditions required for a random process are satisfied. For each k, let $\pi(\theta|k)$ be a strictly positive distribution in the k-dimensional parameter space. Then the stochastic complexity can be defined as

$$I(x) = -\log \sum_{k=0}^{R} Q(k) \int f(x|k,\theta) d\pi(\theta|k) \tag{2.42}$$

where $Q(k) = 1/(R+1)$, and $R \leq n$ is the range of the number of parameters. The stochastic complexity $I(x)$ is defined only relative to a class of models consisting of $f(x|k,\theta)$ and $\pi(\theta|k)$.

Rissanen then gives several model selection criteria, each of which gives a code length as an upper bound approximation of the abstract quantity yielded by the stochastic complexity Equation 2.42. Two of these criteria are used in this thesis. The first is Predictive Minimum Description Length (PMDL) [58, 59] shown below in Equation 2.43. This criterion is used when there exists a sequential ordering in the data. The second is the more general Minimum Message Length shown in the next section.

$$I(x) = \frac{N}{2} \log \sum_{n=1}^{N} r_n^2 + \frac{1}{2} \log |\underset{N\times K}{\mathbf{X}^{\mathrm{T}}} \underset{N\times K}{\mathbf{X}}| \tag{2.43}$$

where:

r_n : the error of the estimate of the n^{th} data item of the dependent/target variable y
$r_n = y_n - \sum_{k=1}^{K} \beta_k x_{nk}$

N : the number of sample data items

$|\underset{N\times K}{\mathbf{X}^{\mathrm{T}}} \underset{N\times K}{\mathbf{X}}|$: the determinant of the covariance matrix $\underset{N\times K}{\mathbf{X}^{\mathrm{T}}} \underset{N\times K}{\mathbf{X}}$ of

K independent variables for N data items

2.3.3.3 Minimum Description Length (MDL) Criterion

When the data are modelled as independent without any proper ordering imposed, then the upper bound of the stochastic complexity measure shown in Equation 2.42 is given below [57, 59]

$$CodeLength_{MDL78} = -\log f(D|\theta) + \frac{K}{2}\log N + (\frac{K}{2}+1)\log(K+2) \qquad (2.44)$$

where:

$f(D|\theta)$: the likelihood function given in Equation 2.5 on page 28.
The result of $-\log f(D|\theta)$ is given in Equation 2.10 on page 28

K : the number of independent variables in the model

N : the number of sample data items

2.3.3.4 Akaike's Information Criterion (AIC)

Like all the other approaches in the category of complexity-penalised criteria, Akaike's Information Criterion (AIC) [46, 47, 48, 49] takes the form of penalised likelihood functions. The AIC Equation 2.45 used in this thesis is taken from [46].

$$AIC = -2(\log f(D|\theta) - K) \qquad (2.45)$$

where:

$f(D|\theta)$: the likelihood function given in Equation 2.5 on page 28.
The result of $-\log f(D|\theta)$ is given in Equation 2.10 on page 28

K : the number of independent variables in the model

The first term in Equation 2.45 measures the fit of the model to the data and the second term penalises complex models. The goal is to minimise the Kullback-Leiber distance of the selected density from the true density. Despite being a pioneer in the learning theory research, AIC is criticised for its expectation that the true distribution is non-existent (see [59]). AIC is shown in [73] and [101] to be asymptotically optimal if the true distribution was not in the finite dimensional family of models being considered.

2.3.3.5 Corrected Akaike's Information Criterion (CAICF)

Bozdogan, a student of Akaike, extends AIC criterion to the criterion called Corrected Akaike's Information Criterion (CAICF) [50, page 23]. CAICF penalises complex models more severely than AIC by using the Fisher Information.

$$CAICF = -\log f(D|\theta) + \frac{1}{2} log I(\theta) + K + \frac{1}{K}\log N \tag{2.46}$$

where:

$f(D|\theta)$: the likelihood function given in Equation 2.5 on page 28.
The result of $-\log f(D|\theta)$ is given in Equation 2.10 on page 28

$I(\theta)$: the Fisher information given in Equation 2.29 on page 38

K : the number of independent variables in the model

N : the number of sample data items

2.3.3.6 Bayesian Information Criterion (BIC)

Bayesian principle says that models should be compared using their posterior probability distributions (see [35]). Schwarz [38] assumes that the prior probabilities of all models were equal and derived Bayesian Information Criterion (BIC) as an asymptotic expression of the likelihood of a model

$$BIC = -2(\log f(D|\theta) - \frac{1}{2}K \log N) \tag{2.47}$$

where:

$f(D|\theta)$: the likelihood function given in Equation 2.5 on page 28.
The result of $-\log f(D|\theta)$ is given in Equation 2.10 on page 28

K : the number of independent variables in the model

N : the number of sample data items

2.3.3.7 Structural Risk Minimization (SRM)

Vladimir Vapnik [109] describes the general model of learning from examples through three components as illustrated in Figure 2.1:

Generator which produces random vectors $x \in R^n$ drawn independently from a fixed but unknown probability distribution function $F(x)$

Supervisor which returns an output value y to every input vector x according to a conditional distribution function[3] $F(y|x)$, also fixed but unknown

[3]This is the general case which includes the case where the Supervisor uses a function $y = f(x)$ [109, page 15]

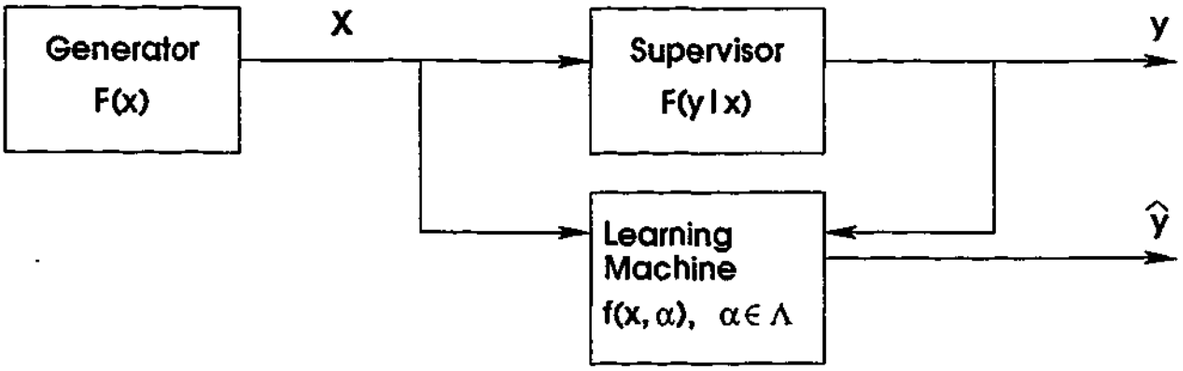


Figure 2.1: Vapnik's model of learning from examples. During the learning process, the Learning Machine observes the pairs (x, y) (the training set). After training, the machine must return a value $\hat{y}$ on any given x. The goal is to return a value $\hat{y}$ which is close the Supervisor's response y. *Source:* modified from [109, page 16].

Learning Machine which is capable of implementing a set of functions $f(x, \beta)$, $\beta \in \Lambda$, where Λ is a set of parameters.

The problem of learning is that of choosing from the given set of functions $f(x, \beta)$, $\beta \in \Lambda$, the one which best approximates the Supervisor's response. The selection of the desired function is based on a training set of N independent and identically distributed (i.i.d) observations drawn according to $F(x, y) = F(x)F(y|x)$:

$$(x_1, y_1), \ldots, (x_N, y_N) \tag{2.48}$$

To solve this problem, one measures the discrepancy/loss $L(y, f(x, \beta))$ between the response y of the Supervisor to a given input x and the response $f(x, \beta)$ provided by the Learning Machine. The expected value of the loss, is given by the risk function:

$$R(\beta) = \int L(y, f(x, \beta)) dF(x, y) \tag{2.49}$$

The goal is to find the function $f(x, \beta_0)$ which minimises the risk function $R(\beta)$ over the class of function $f(x, \beta), \beta \in \Lambda$, in the situation where the joint probability distribution $F(x, y)$ is unknown and the only available information is contained in the training set shown in Equation 2.48.

For the problem of regression estimation, the Supervisor's answer y is taken as a real value and $f(x, \beta), \beta \in \Lambda$ as a set of real functions which contains the regression function

$$f(x, \beta_0) = \int y \, dF(y|x) \tag{2.50}$$

This regression function is known to be the function which minimises the risk function shown in Equation 2.49 with the loss function:

$$L(y, f(x, \beta)) = (y - f(x, \beta))^2 \tag{2.51}$$

Thus, according to Vapnik, the problem of regression estimation is the problem of minimising the risk function (Equation 2.49) in the situation where the probability measure $F(x, y)$ is unknown but the training data (Equation 2.48) are given.

The risk function (Equation 2.49) is then replaced by the empirical risk function which is constructed on the basis of the training set

$$R_{emp}(\beta) = \frac{1}{N} \sum_{n=1}^{N} L(y, f(x, \beta)) \tag{2.52}$$

Vapnik [109, pages 18–19] says that the Empirical Risk Minimization inductive principle (ERM principle) states that the function $L(y, f(x, \beta_0))$ which minimises the risk function $R(\beta)$ shown in Equation 2.49 is approximated by the function

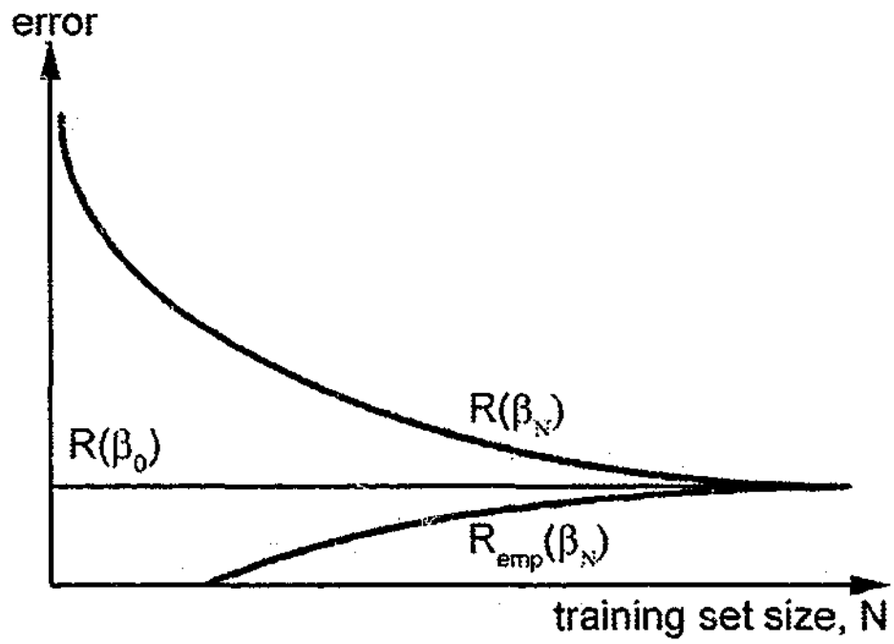


Figure 2.2: Both the value of empirical risks $R_{emp}(\beta_N)$ and the values of risk for functions that minimise the expected risk $R(\beta_N)$ converge to minimal possible risk $R(\beta_0)$. *Source:* modified from [110, page 8].

$L(y, f(x, \beta_N))$ which minimises the empirical risk function $R_{emp}(\beta)$ shown in Equation 2.52.

For $A \leq L(y, f(x, \beta)) \leq B$, $\beta \in \Lambda$ a set of bounded loss functions, it is necessary and sufficient that the empirical risk $R_{emp}(\beta)$ converges uniformly and rapidly to the actual risk $R(\beta)$ over the set $L(y, f(x, \beta))$, $\beta \in \Lambda$ for the ERM principle to be consistent for any probability measure $P(x)$ of the data x. This condition is illustrated in Figure 2.2.

The necessary and sufficient condition for consistency of ERM principle and fast convergence is described in the following equation [110, page 11]

$$\lim_{N \to \infty} \frac{G^{\Lambda}(N)}{N} = 0 \tag{2.53}$$

where $G^{\Lambda}(N)$ is the growth function.

Equation 2.53 is satisfied if the growth function $G^{\Lambda}(N)$ for the set of functions $L(y, f(x, \beta)),\ \beta \epsilon \Lambda$ is bounded by a logarithmic function with coefficient h shown in the following equation [110, page 12].

$$G^{\Lambda}(N) \ < \ h(\ln \frac{N}{h} + 1) \tag{2.54}$$

where h is an integer for which

$$G^{\Lambda}(h) \ = h \ln 2,$$
$$G^{\Lambda}(h+1) \ \neq \ (h+1) \ln 2$$

The term h in Equation 2.54 is called the VC-dimension (i.e. Vapnik-Chervonenkis dimension). Vapnik (see [109, page 75] or [110, page 12]) asserted that the finiteness of the VC-dimension of the set of functions implemented by the Learning Machine forms the necessary and sufficient condition for consistency of the ERM principle and fast convergence.

The VC-dimension of the set of linear functions (see [109, page 78] or [110, page 13])

$$L(y, f(x, \beta)) = \sum_{k=1}^{K} \beta_k x_k + \beta_0,\ \beta_0, \ldots, \beta_K \ \epsilon \ (-\infty, \infty) \tag{2.55}$$

in K-dimensional coordinate space $X = (x_1, \ldots, x_K)$ is equal to $h = K + 1$, because the VC-dimension of corresponding linear functions is equal to $K + 1$. For the set of linear functions, the VC-dimension equals the number of free parameters $\beta_0, \ldots, \beta_K$.

Vapnik [109, page 90] then says that the ERM principle is however intended for dealing with large sample sizes. When the sample size is small, a small $R_{emp}(\beta_N)$ does not guarantee a small value of the actual risk. For this reason, Vapnik proposes a new principle, called Structural Risk Minimization (SRM) inductive principle. SRM principle is intended to minimise the risk function with respect to both empirical risk and VC-dimension of the set of functions. For this reason, it is clear that the SRM principle defines a trade-off between the quality of the approximation of the given data and the complexity of the approximating function.

Using the VC-dimension for the set of linear function shown above, the risk function for the linear second-order polynomial models evaluated in this thesis is shown in the equation below [110, page 17]

$$\Phi(\beta) = \frac{R_{emp}(\beta)}{1-\sqrt{\frac{(K+1)(\ln \frac{N}{(K+1)}+1)-\ln \eta}{N}}} \tag{2.56}$$

$$\Leftrightarrow \quad \Phi(\beta) = \frac{\frac{1}{N}\sum_{n=1}^{N} r_n^2}{1-\sqrt{\frac{(K+1)(\ln \frac{N}{K+1}+1)-\ln \eta}{N}}} \tag{2.57}$$

where:

$\Phi(\beta)$: the estimate of the prediction risk for a linear polynomial with coefficients β

r_n : the error of the estimate of the n^{th} data item

$r_n = y_n - \sum_{k=1}^{K} \beta_k x_{nk}$

$(K+1)$: the VC-dimension of polynomials with K number of parameters

η : a probability constant $\eta = 0.125$. Vapnik asserted that the inequality $R(\beta) \leq \Phi(\beta)$ holds true with probability at least $1-\eta$.

In this thesis, we use $\eta = 0.125$ following [110]

(see [109, pages 79–82, 85–87])

K : the number of independent variables in the model

N : the number of sample data items

2.4 Conclusion

This chapter presents the model selection methods most commonly used in the literature. Based on the existence of the term that quantifies model complexity, the criteria can be divided into two categories: those which require a test data set to decide on a model and complexity-penalised criteria. The first category calculate the model parameters using the training data set and makes the decision whether or not the model will be selected based on its performance on a separate test data set. The second category claims to have a term to quantify model complexity, hence is able to make the decision solely based on the performance of the model on training data. This chapter proposes a new model selection criterion based on the Minimum Message Length (MML) principle which includes the cost of using the combination of single and compound variables in a model in the quantification of model complexity.

Given an optimisation search algorithm which finds models with increasing degree of complexity from the search space, the goal of automated model selection is to have an objective function which serves as a stopping rule of the search when the global optimum has been reached, i.e. the search can stop because a model with the right combination of variables and degree of complexity has been found. Chapter 3 tests the robustness of each of the model selection criteria discussed in this chapter when used as a stopping rule in an automated model selection process. The optimisation search algorithm used for the process is also discussed in the chapter in Section 3.2.2 on page 57.

Chapter 3

Automated Second-order Polynomial Model Discovery

3.1 Introduction

The aim of this chapter is to test the ability of the complexity-penalised model selection methods outlined in Chapter 2.3.3, to discover, in an automated manner, a model that has generated the data under consideration[1]. Of particular interest is the ability of the methods to discover second-order independent variables, independent variables with weak causal relationships with the target variable given a small sample size, and independent variables with weak links to the target variable but strong links from other variables which are not directly linked with the target variable.

What is involved in the model selection task is outlined in Section 3.2. A summary of the model selection criteria tested is given in Section 3.2.1. Section 3.2.2 outlines the common non-backtracking search strategy that has been programmed

[1] An earlier version of this chapter has been published in [42]

for this thesis and is used with all of the model selection criteria. Another search algorithm, namely simulated annealing[89], has actually been programmed and used to do the same model selection task. However, we have been unable to find the right initial temperature to enable a search process to converge to the true model. This is consistent with what has been reported in the literature, that is the implementation of simulated annealing is more of an art than a science. Avoidance of entrapment in local minima is dependent on the "annealing schedule", the choice of initial temperature, how many iterations are performed at each temperature, and how much the temperature is decremented at each step as cooling proceeds [92]. The difficulties in automating the setting of the above variables in simulated annealing prompted us to disregard the search strategy for the second-order polynomial model selection task at hand.

3.2 Model Selection: What It Involves

The task of model selection involves three key features. The first key feature is in deciding the form of the models to be considered. The second is measuring the cost of selecting a particular model. The third is the search for a model through the space of all possible models.

The models considered in this thesis are of the form of the linear second-order polynomials shown in Equation 2.1 in Section 2.2 on page 23. The cost function can take the form of any of the model selection methods discussed in Section 2.3. The search strategy is given in Section 3.2.2.

Table 3.1: Model selection methods tested in this chapter for the task of discovering the true model that has generated a set of artificial data

Method		Reference	Equation	Page No.
Complexity-penalised Methods				
Minimum Message Length	MML	[19]	2.32	40
Predictive Minimum Description Length	PMDL	[59]	2.43	45
Minimum Description Length	MDL	[59]	2.44	46
Akaike's Information Criterion	AIC	[46]	2.45	46
Corrected AIC	CAICF	[50]	2.46	47
Bayesian Information Criterion	BIC	[38]	2.47	48
Structured Risk Minimisation	SRM	[109]	2.57	53

3.2.1 Model Selection Methods Tested

In this chapter, all of the model selection methods discussed in Section 2.3 are tested for the task of discovering the true model that has generated a set of artificial data. The summary of the methods tested is given in Table 3.1.

3.2.2 Optimisation Search Algorithm

A non-backtracking search algorithm has been developed to be used as a common search engine for the different stopping criteria. This algorithm starts with an empty model. The variable which gives the lowest model cost amongst all of the potential variables in the search space as reflected by the model selection criterion used, will be chosen as the first variable for the model.

At any stage thereafter, consider all models formed by adding one new variable. Choose the variable which leads to the best new model. Now consider deleting a variable from the model. Examine all models which are formed by deleting one variable. Find the best such model which is better than the current best model. If we can improve the current model by removing that variable, then remove it from

the current best model. The search terminates when we can no longer add a variable to get a better model.

Appendix C on page 152 gives a pseudocode of the search algorithm. In case a model selection method overfits the data, a limit in the maximum number of variables that a model can have is imposed to enable the search to terminate in a reasonable amount of time. To automate termination of a model selection search that is seen to have overfitted the data, in this thesis, a model is set to have a maximum of 70 variables. An example of the trace results of an experiment of running the search algorithm using MML as the cost function is given in Appendix D on page 157.

3.3 Experimental Design

Three true models as shown in Figure 3.1, 3.2 and 3.3 have been designed for the experiments. Each true model consists of a target variable and a set of single and compound independent variables. Not all of the variables are necessarily directly or at all connected to the target variable. Each value of an independent variable is chosen randomly from a normal distribution $N(0,1)$[2]. For each model, 6 training and test data sets comprising 500, 1000, 2000, 4000, 6000 and 10000 instances respectively are generated.

The product of two independent variables is calculated from the standardized values of each variable. Each value of the target variable is calculated from the values of all of the independent variables directly linked to it multiplied by the respective link weights plus a unit normal noise value, $\epsilon \sim NID(0,1)$.

[2]The unit normal random data sets $u_1,\ldots,u_K$ used in this thesis are generated using the random number generator program `FastNorm2.c` written by Chris Wallace [17].

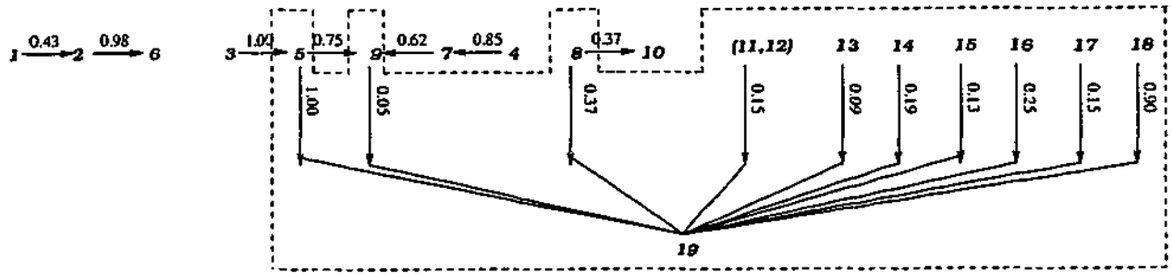


Figure 3.1: Model 1. The independent variables and the link coefficients to be discovered are in the dashed-line box, i.e. only variables with direct links to the target variable. Hence the polynomial model to estimate variable 19 is $y = x_5 + 0.05x_9 + 0.37x_8 + 0.15x_{(11,12)} + 0.09x_{13} + 0.19x_{14} + 0.13x_{15} + 0.25x_{16} + 0.15x_{17} + 0.90x_{18} + \varepsilon$, with $\varepsilon \in N(0.1)$. The link between two independent variables indicates the correlation coefficient between the variables. For example, $1 \xrightarrow{0.43} 2$ indicates the correlation coefficient between variable 1 and variable 2 is 0.43.

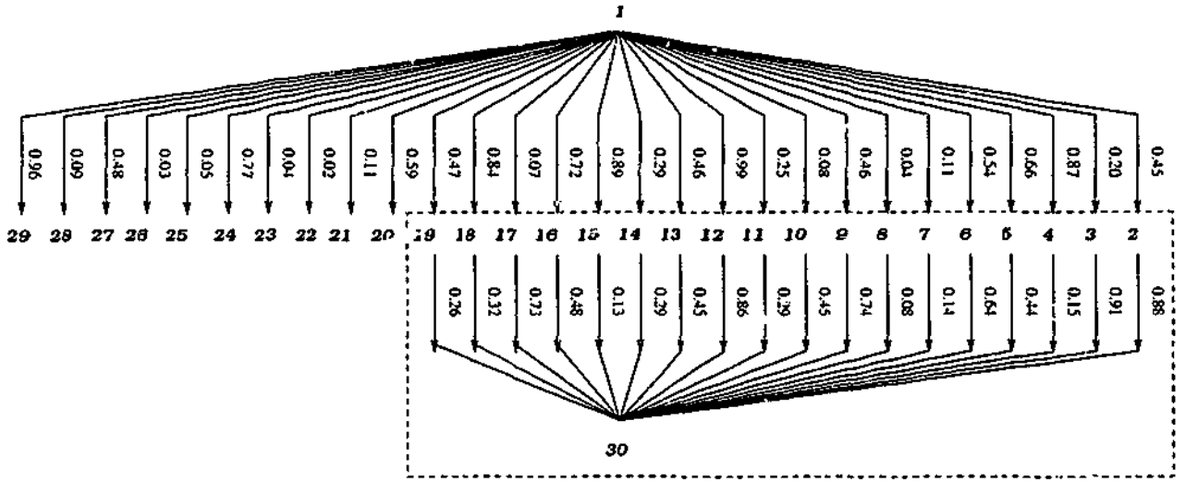


Figure 3.2: Model 2. Variable 1 is directly linked to all of the variables with direct links (some of which are very weak) to the target variable. Hence, the polynomial model to estimate variable 30 is $y = 0.26x_{19} + 0.32x_{18} + 0.73x_{17} + 0.48x_{16} + 0.13x_{13} + 0.29x_{14} + 0.45x_{13} + 0.86x_{12} + 0.29x_{11} + 0.45x_{10} + 0.74x_9 + 0.08x_8 + 0.14x_7 + 0.64x_6 + 0.44x_5 + 0.15x_4 + 0.91x_3 + 0.88x_2 + \varepsilon$, with $\varepsilon \in N(0.1)$. Large link coefficients are deliberately placed between variable 1 and these variables to see if this will cause variable 1 also to be chosen.

The search engine is presented with the data of the target variable and all of the available independent variables and the possible products of the single variables.

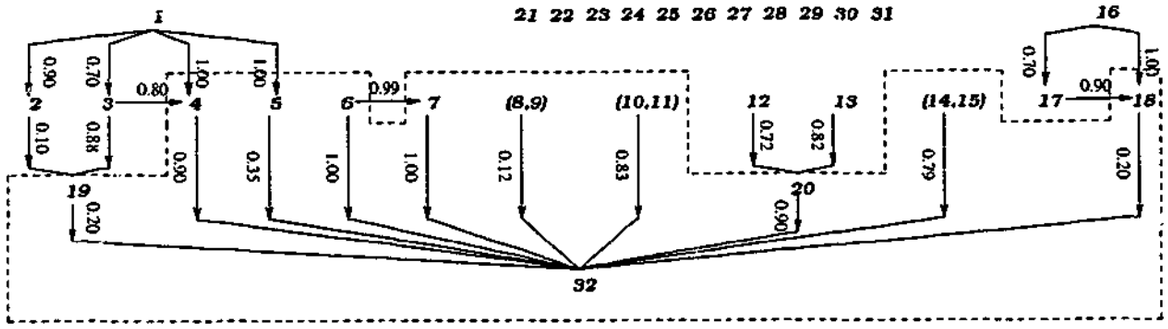


Figure 3.3: Model 3. The polynomial model to estimate variable 32 is $y = 0.20x_{19} + 0.90x_4 + 0.35x_5 + 1.00x_6 + 1.00x_7 + 0.12x_{(8,9)} + 0.83x_{(10,11)} + 0.90x_{20} + 0.79x_{(14,15)} + 0.20x_{18} + \varepsilon$, with $\varepsilon \in N(0.1)$. Unit normally distributed random values with no link to the target variable are generated for variables 21 to 31 are included in the pool of potential variables

3.3.1 Orthogonal Transformation of Independent Variables

A high degree of correlation among independent variables (or regressors) can cause ordinary Gaussian least squares regression method to yield inaccurate coefficients for the correlated regressors. The inaccuracy can take the form of wrong coefficient signs or coefficient values that do not solely reflect the influence of individual regressors to the dependent variable but also the influence of one regressor to the other regressors it has high correlation with.

The basic idea of orthogonal transformation of regressors is to find an orthogonal basis to express the independent variables, perform regression calculations in this basis, then transform back to obtain regression coefficients in the original basis. Because the regression is done in an orthogonal basis where the transformed regressors are uncorrelated to one another (i.e. the covariance matrix of the regressors forms an identity matrix), the transformed coefficients calculated do not reflect the effect of the high degree of correlations among the real regressors. The next section explains the way orthogonal transformation is done in the experiments in this thesis and the

way the coefficients for the real regressors are calculated through this transformation process.

3.3.1.1 The Orthogonal Transformation Used

A. Problem Definition

We start with a set of observations of regressors

$$\mathbf{x_1}, \ldots, \mathbf{x_K} = \underset{N \times K}{\mathbf{X}} \tag{3.1}$$

the target variable

$$y_1, \ldots, y_N = \underset{N \times 1}{\mathbf{y}} \tag{3.2}$$

and we would like to infer a polynomial prediction model from the regressors

$$\underset{N \times 1}{\hat{\mathbf{y}}} = \underset{N \times K}{\mathbf{X}} \; \underset{K \times 1}{\beta} \tag{3.3}$$

with sum squares of the residuals defined as

$$\sum_{n=1}^{N} r_n^2 = \sum_{n=1}^{N} (y_n - \sum_{k=1}^{K} x_{nk}\beta_k)^2 \tag{3.4}$$

$$= (\underset{N \times 1}{\mathbf{y}} - \underset{N \times K}{\mathbf{X}} \; \underset{K \times 1}{\beta})^{\mathrm{T}} (\underset{N \times 1}{\mathbf{y}} - \underset{N \times K}{\mathbf{X}} \; \underset{K \times 1}{\beta}) \tag{3.5}$$

The task is to choose an orthogonal transformation $\underset{K\times K}{\mathbf{P}}$ (i.e. $\mathbf{P}^{\mathbf{T}}\,\mathbf{P} = \mathbf{P}\,\mathbf{P}^{\mathbf{T}} = \mathbf{I}$) that we can use to calculate a set of transformed regressors from the original regressor

$$\mathbf{w_1}, \ldots, \mathbf{w_K} = \underset{N\times K}{\mathbf{W}} = \underset{N\times K}{\mathbf{X}}\ \underset{K\times K}{\mathbf{P}} \tag{3.6}$$

which satisfies the following requirements:

1. All of the transformed regressors are uncorrelated, $\mathbf{W}^{\mathbf{T}}\,\mathbf{W} = N\,\mathbf{I}$, with $N =$ the number of observations, used as a scaling factor.

2. The sum squares of the prediction errors of the original polynomial prediction model, $\sum_{n=1}^{N} r_n^2$, and those of the transformed model, $\sum_{n=1}^{N} e_n^2$, should be the same.

The coefficients for the transformed regressors

$$\gamma_1, \ldots, \gamma_K = \underset{K\times 1}{\Gamma} \tag{3.7}$$

(3.8)

are to be used to calculate the coefficients for the original regressors of the polynomial prediction model to be found in the experiments in this chapter

$$\beta_1, \ldots, \beta_K = \underset{K\times 1}{\beta} \tag{3.9}$$

(3.10)

B. Solution: How to define the transformed regressors w_k, the transformed coefficients γ_k and the coefficients for the original regressors β_k

We first establish the fact that a given symmetric square matrix, $\frac{1}{N}\underset{K\times K}{\mathbf{A}} = \underset{N\times K}{\mathbf{X}^T}\ \underset{N\times K}{\mathbf{X}}$, by definition has K pairs of eigenvalues and eigenvectors, namely

$$\lambda_1, \mathbf{q_1} \quad \lambda_2, \mathbf{q_2} \quad \ldots \quad \lambda_K, \mathbf{q_K} \quad = \quad \underset{K\times K}{\mathbf{\Lambda}}\ , \underset{K\times K}{\mathbf{Q}} \tag{3.11}$$

where

$\underset{K\times K}{\mathbf{\Lambda}}$: a diagonal matrix with the eigenvalues of matrix $\frac{1}{N}\underset{K\times K}{\mathbf{A}}$ as its diagonal elements

$\underset{K\times K}{\mathbf{Q}}$: a K × K matrix with the eigenvectors of matrix $\frac{1}{N}\underset{K\times K}{\mathbf{A}}$ as its columns

With the existence of the eigenvalues and eigenvectors[3]. defined in Equation 3.11 above, the following equation holds

$$\frac{1}{N}\ \underset{K\times K}{\mathbf{A}}\ \underset{K\times K}{\mathbf{Q}} = \underset{K\times K}{\mathbf{Q}}\ \underset{K\times K}{\mathbf{\Lambda}} \tag{3.12}$$

since Q is orthonormal, square and assumed to be nonsingular $\mathbf{Q^{-1}} = \mathbf{Q^T}$ and $\mathbf{Q^T}\ \mathbf{Q} = \mathbf{I}$

$$\Rightarrow \frac{1}{N}\ \underset{K\times K}{\mathbf{A}} = \underset{K\times K}{\mathbf{Q}}\ \underset{K\times K}{\sqrt{\mathbf{\Lambda}}}\ \underset{K\times K}{\sqrt{\mathbf{\Lambda}}}\ \underset{K\times K}{\mathbf{Q}^T} \tag{3.13}$$

$$\Rightarrow N\ \underset{K\times K}{\mathbf{\Lambda}} = \underset{K\times K}{\mathbf{Q}^T}\ \underset{K\times K}{\mathbf{A}}\ \underset{K\times K}{\mathbf{Q}} \tag{3.14}$$

We will use the above equations in the proofs of the following results.

[3]In the experiments done in this thesis, the eigenvalues and eigenvectors of a square matrix are calculated using the Jacobian diagonalization of real symmetric matrix program `jacob.c` written by Chris Wallace. The program will terminate and return an error message if the eigenvalues of the square matrix equal zero

Result 3.1 If we define the new transformed regressors w_k as

$$\underset{N\times K}{\mathbf{W}} = \underset{N\times K}{\mathbf{X}}\ \underset{K\times K}{\mathbf{Q}}\ \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}} \tag{3.15}$$

$$\Rightarrow w_{n,k} = \sum_{m=1}^{K} \frac{x_{n,m}\ q_{m,k}}{\sqrt{\lambda_m}} \tag{3.16}$$

where

$w_{n,k}$: n^{th} data point of the transformed regressor w_k

$x_{n,m}$: n^{th} data point of the original regressors x_m

λ_m : the m^{th} diagonal element of $\underset{K\times K}{\mathbf{\Lambda}}$, the diagonal eigenvalue matrix of $\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}}$

$q_{m,k}$: the m^{th} element of column k of $\underset{K\times K}{\mathbf{Q}}$ the eigenvector matrix of $\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}}$

the regressors $w_1, \ldots, w_K$ will be uncorrelated to one another.

Proof. If regressors $\mathbf{x_i}$ and $\mathbf{x_j}$ are uncorrelated, then $\mathbf{x_i}^{\mathrm{T}} \cdot \mathbf{x_j} = 0$ if $i \neq j$ and $\mathbf{x_i}^{\mathrm{T}} \cdot \mathbf{x_k} = 1$

$$\underset{N\times K}{\mathbf{W}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{W}} = (\underset{N\times K}{\mathbf{X}}\ \underset{K\times K}{\mathbf{Q}}\ \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}})^{\mathrm{T}}\ (\underset{N\times K}{\mathbf{X}}\ \underset{K\times K}{\mathbf{Q}}\ \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}}) \tag{3.17}$$

$$\Leftrightarrow \underset{N\times K}{\mathbf{W}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{W}} = \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}}\ \underset{K\times K}{\mathbf{Q}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}}\ \underset{K\times K}{\mathbf{Q}}\ \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}}$$

$$\Leftrightarrow \underset{N\times K}{\mathbf{W}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{W}} = \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}}\ \underset{K\times K}{\mathbf{Q}^{\mathrm{T}}}\ \underset{K\times K}{\mathbf{A}}\ \underset{K\times K}{\mathbf{Q}}\ \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}}$$

$$\Rightarrow \underset{N\times K}{\mathbf{W}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{W}} = \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}}\ N\ \underset{K\times K}{\mathbf{\Lambda}}\ \underset{K\times K}{\mathbf{\Lambda}^{-\frac{1}{2}}} \qquad \text{by Equation 3.14}$$

$$\Leftrightarrow \underset{N\times K}{\mathbf{W}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{W}} = N\ \underset{K\times K}{\mathbf{I}} \tag{3.18}$$

q.e.d

Result 3.2 The polynomial prediction model inferred from the transform regressors defined in Equation 3.16 yields the same sum squares of prediction errors as the model inferred from the original regressors.

Proof. In the new orthogonal basis, the sum squares of prediction errors of the original model defined in Equation 3.5 transforms to those of the transformed model, that is

$$\sum_{n=1}^{N} r_n^2 = (\underset{N\times 1}{\mathbf{y}} - \underset{N\times K}{\mathbf{X}}\ \underset{K\times 1}{\beta})^{\mathrm{T}}\ (\underset{N\times 1}{\mathbf{y}} - \underset{N\times K}{\mathbf{X}}\ \underset{K\times 1}{\beta}) \tag{3.19}$$

transforms to

$$\sum_{n=1}^{N} e_n^2 = (\underset{N\times 1}{\mathbf{y}} - \underset{N\times K}{\mathbf{W}}\ \underset{K\times 1}{\Gamma})^{\mathrm{T}}\ (\underset{N\times 1}{\mathbf{y}} - \underset{N\times K}{\mathbf{W}}\ \underset{K\times 1}{\Gamma}) \tag{3.20}$$

where Γ are the coefficients for the transformed regressors.

As derived in Equations 2.13 and 2.15 on page 29, it is known that the sum squares of residuals $\sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N}(y_n - \sum_{k=1}^{K} w_{n,k}\gamma_k)^2$ is minimised when

$$\underset{N\times K}{\mathbf{W}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{W}}\ \underset{K\times 1}{\Gamma} = \underset{N\times K}{\mathbf{W}}^{\mathrm{T}}\ \underset{N\times 1}{\mathbf{y}} \tag{3.21}$$

$$\Rightarrow \underset{K\times 1}{\Gamma} = (\underset{N\times K}{\mathbf{W}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{W}})^{-1}\ \underset{N\times K}{\mathbf{W}}^{\mathrm{T}}\ \underset{N\times 1}{\mathbf{y}} \tag{3.22}$$

if we define $\mathbf{D} = \underset{N\times K}{\mathbf{W}}^{\mathrm{T}}\ \underset{N\times 1}{\mathbf{y}}$, we get

$$\Rightarrow \underset{K\times 1}{\Gamma} = (\underset{N\times K}{\mathbf{W}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{W}})^{-1}\ \mathbf{D} \tag{3.23}$$

We now define the sum squares of residuals in terms of the value of Γ derived in Equation 3.23

$$\sum_{n=1}^{N} e_n^2 = (\underset{N\times 1}{\mathbf{y}} - \underset{N\times K}{\mathbf{W}}\ \underset{K\times 1}{\Gamma})^T (\underset{N\times 1}{\mathbf{y}} - \underset{N\times K}{\mathbf{W}}\ \underset{K\times 1}{\Gamma}) \tag{3.24}$$

$$\Leftrightarrow \sum_{n=1}^{N} e_n^2 = \mathbf{y}^T\mathbf{y} - \Gamma^T\mathbf{W}^T\mathbf{y} - \mathbf{y}^T\mathbf{W}\Gamma + \Gamma^T\mathbf{W}^T\mathbf{W}\Gamma \tag{3.25}$$

Substituting $\Gamma = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y}$ from Equation 3.22 gives

$$\Rightarrow \sum_{n=1}^{N} e_n^2 = \mathbf{y}^T\mathbf{y} - [(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y}]^T\mathbf{W}^T\mathbf{y} - \mathbf{y}^T\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y} + [(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y}]^T\mathbf{W}^T\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y} \tag{3.26}$$

$$\Leftrightarrow \sum_{n=1}^{N} e_n^2 = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y} - \mathbf{y}^T\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y} + \mathbf{y}^T\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}[\mathbf{W}^T\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}]\mathbf{W}^T\mathbf{y} \tag{3.27}$$

Since $\mathbf{W}^T\mathbf{W}$ must be symmetric, $(\mathbf{W}^T\mathbf{W})^{-1}$ must also be symmetric.
Thus, $[(\mathbf{W}^T\mathbf{W})^{-1}]^T = (\mathbf{W}^T\mathbf{W})^{-1}$

$$\Leftrightarrow \sum_{n=1}^{N} e_n^2 = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{W}\Gamma \tag{3.28}$$

$$\Leftrightarrow \sum_{n=1}^{N} e_n^2 = \mathbf{y}^T\mathbf{y} - \mathbf{D}^T\Gamma \tag{3.29}$$

If we instead of doing the orthogonal transformation, we had used the original regressors, we would have got their coefficients as

$$\underset{K\times 1}{\beta} = (\underset{N\times K}{\mathbf{X}^T}\ \underset{N\times K}{\mathbf{X}})^{-1}\ \underset{N\times K}{\mathbf{X}^T}\ \underset{N\times 1}{\mathbf{y}} \tag{3.30}$$

and if we define $\underset{NXK}{\mathbf{C}} = \underset{NXK}{\mathbf{X}^T}\ \underset{NX1}{\mathbf{y}}$, we get

$$\Rightarrow \underset{KX1}{\beta} = (\underset{NXK}{\mathbf{X}^T}\ \underset{NXK}{\mathbf{X}})^{-1}\ \mathbf{C} \tag{3.31}$$

this would have given us sum squares of residuals as shown in Equation 3.5, which, derived in a similar manner as the sum squares of residuals of the transformed model shown in Equation 3.29, takes the form

$$\sum_{n=1}^{N} r_n^2 = \mathbf{y}^T\ \mathbf{y}\ - \mathbf{C}^T\ \beta \tag{3.32}$$

So, from Equations 3.22 and 3.32 we know that the sum squares of residuals of the transformed model should be the same as those of the original model if $\mathbf{D}^T\ \Gamma = \mathbf{C}^T\ \beta$. To prove this, we expand Equation $\mathbf{D}^T\ \Gamma$

$$\begin{aligned}
\mathbf{D}^T\ \Gamma &= (\mathbf{W}^T\mathbf{y})^T(\mathbf{W}^T\ \mathbf{W})^{-1}\ (\mathbf{W}^T\ \mathbf{y}) \\
&= \mathbf{y}^T\ \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\ \mathbf{W}^T\ \mathbf{y} \\
&= \mathbf{y}^T\ \mathbf{XP}\ (\mathbf{P}^T\ \mathbf{X}^T\ \mathbf{XP})^{-1}\mathbf{P}^T\ \mathbf{X}^T\ \mathbf{y} && \text{by Equation 3.6, } \mathbf{W} = \mathbf{XP} \\
&= \mathbf{y}^T\ \mathbf{XP}\ [\mathbf{P}^T\ (\mathbf{X}^T\ \mathbf{X})^{-1}\ \mathbf{P}]\ \mathbf{P}^T\ \mathbf{X}^T\ \mathbf{y} && \text{because } \mathbf{P}^{-1} = \mathbf{P}^T \\
&= \mathbf{y}^T\ \mathbf{XPP}^T\ (\mathbf{X}^T\ \mathbf{X})^{-1}\mathbf{PP}^T\ \mathbf{X}^T\ \mathbf{y} \\
&= \mathbf{y}^T\ \mathbf{X}(\mathbf{X}^T\ \mathbf{X})^{-1}\mathbf{X}^T\ \mathbf{y} && \text{because } \mathbf{PP}^T = \mathbf{I} \\
&= \mathbf{C}^T\ \beta
\end{aligned}$$

So,

$$\sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} r_n^2 \tag{3.33}$$

q.e.d

The definition of the transformed coefficients Γ shown in Equation 3.22 is correct if we use ordinary least squares method. As explained on page 43, we actually use a modified least squares shown in Equation 2.40 to calculate the coefficients for the regressors of all of the models considered in this thesis. Hence, with the transformation, the original parameter estimates $\hat{\sigma}^2$ and β shown in Equations 2.39 and 2.40 and, for clarity, replicated in Equations 3.34 and 3.35 below

$$\hat{\sigma}^2 = \frac{1}{N}\left(\sum_{n=1}^{N} r_n^2 + \sum_{k=1}^{K} \frac{\beta_k^2}{\alpha^2}\right) \tag{3.34}$$

and

$$\underset{K\times 1}{\hat{\beta}} = (\underset{N\times K}{\mathbf{X}}{}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{X}} + \frac{1}{\alpha^2}\ \underset{K\times K}{\mathbf{I}})^{-1}\underset{N\times K}{\mathbf{X}}{}^{\mathrm{T}}\ \underset{N\times 1}{\mathbf{y}} \tag{3.35}$$

transform to

$$\hat{\sigma}^2 = \frac{1}{N}\left(\sum_{n=1}^{N} e_n^2 + \sum_{k=1}^{K} \frac{\gamma_k^2}{\alpha^2}\right) \tag{3.36}$$

and

$$\underset{K\times 1}{\hat{\Gamma}} = [(N + \frac{1}{\alpha^2})\ \underset{K\times K}{\mathbf{I}}]^{-1}\underset{N\times K}{\mathbf{W}}{}^{\mathrm{T}}\ \underset{N\times 1}{\mathbf{y}} \tag{3.37}$$

Note that r_n shown in Equation 3.34 and e_n shown in Equation 3.36 are calculated using the coefficients shown respectively in Equations 3.35 and 3.37.

3.3.2 Performance Criteria

The performance criteria for the true model discovery task are whether or not a model selection method manages to select a model with the same set of variables and corresponding coefficients as those of the true model, as reflected in the following measures:

1. *K: The Size of the Discovered Model*

   The size of the discovered model is represented in the number of variables selected, K. Since the size of the true model is known, the size of the discovered model can be used as an indication of its degree of fit to the data.

2. *Model Error*

   Model error shows how close the coefficients of the discovered model, $\hat{\beta}_k$, are with those of the true model, β_k.

$$\frac{1}{K}\sum_{k=1}^{K}(\beta_k - \hat{\beta}_k)^2 \tag{3.38}$$

3. *Model Predictive Performance*

   Model predictive performance (on test data) is quantified by two measures:

   (a) Root of the mean of the sum of squared deviations of the predictions, $\hat{y}$, from the true values of the dependent/target variable, y.

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N} r_n^2} \tag{3.39}$$

   where:

r_n : the error of the estimate of the n^{th} data item of
the dependent/target variable y
$r_n = y_n - \sum_{k=1}^{K} \beta_k x_{nk}$

N : the number of sample data items

K : the number of independent variables

(b) Coefficient of determination which represents the percentage of the variability in y (as represented in $\sum_{n=1}^{N}(y_n - \bar{y})^2$ in Equation 3.40), that is explained by using x to predict y (as represented in $\sum_{n=1}^{N} r_n^2$ in Equation 3.40).

$$R^2 = 1 - \frac{\sum_{n=1}^{N} r_n^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{3.40}$$

3.4 Results and Discussions

The results of the experiments with artificial data sets for the true models 3.1, 3.2 and 3.3 are respectively given in Tables 3.2 and 3.3, 3.4 and 3.5, and 3.6 and 3.7. The results show that the search engine using model selection methods MML, MDL, CAICF, SRM or PMDL manages to home in to the true models (i.e. all of the variables with direct links to the target variable shown in the number of variables discovered. Due to a space constraint, the variables and their coefficients are not shown).

The other methods, namely AIC and BIC tend to choose wrong and much more complex models. The fact that the models selected by AIC and BIC for Model 2 and 3 have 70 variables for all of the sample sizes suggests that the search has

been stopped before convergence. From the performance criteria and the number of variables chosen for Model 1, 2 and 3, it is clear that AIC and BIC have *overfitted the training data.* Hence, this implies that in those model selection methods, the penalty for chosing a more complex model is too small compared to the reward of a better data fit.

Nonetheless, it has been observed that all of the methods selected some of the significant regressors early on in the search process and assigned relatively large coefficients to them and small coefficients to the variables chosen which do not exist in the true model. These results suggest that if a model selection procedure is to be fully automated, MML, MDL, CAICF, SRM and PMDL can reliably converge to the true model (if one exists), or to a reasonably parsimonious model estimate. The models selected by the AIC and BIC may need further judgements in deciding on the final model which can take two forms. First, a model can be chosen half way through the search process just before a more complex model with worse performance on some test data set is chosen. If this approach is taken, then it means we treat AIC and BIC as methods which need test data sets to decide on a model outlined in Section 2.3.2. Second, some of the variables with small coefficients are pruned out from the model. The need for these manual adjustments explains the real reason behind the traditional common practice of specifying beforehand the maximum number of variables for a model (e.g.[108], [6]).

3.5 Conclusion

The performance of the new Minimum Message Length model selection criterion proposed in Section 2.3.3.1 alongside with the other model criteria most commonly cited in the literature outlined in 2.3 are tested in their ability to recover true models

from artificial data sets with varying sample sizes and levels of noise using a common non-backtracking search strategy outlined in Section 3.2.2. The robustness of these model selection criteria in performing the task of selecting models that balance model complexity and goodness of fit is examined.

Based on the experiments with artificial data, it has been shown that MML, MDL, CAICF, SRM and PMDL methods are good candidates for fully automated model selection tasks. Given a noisy data set, the methods can reliably converge to a true model (if one exists) or to a reasonably parsimonious model.

The fact that AIC and BIC have overfitted the training data suggests that when comparing two models with different complexity, the increase in the penalty terms for model complexity is not sufficient compared to the decrease in the terms for goodness of fit. This prompted the doubt that the balancing mechanism of the methods might not be robust enough for automated model selection task.

Table 3.2: Performance of the different model selection methods on the task of discovering Model 1 using 500, 1000 and 2000 sample sizes

Sample Size	Method	Model 1 (nvar= 10)			
		nvar	ModelErr	RMSE	R^2
500	MML	6	0.0109	1.0399	0.7285
	MDL	6	0.0109	1.0399	0.7285
	CAICF	6	0.0109	1.0399	0.7285
	SRM	6	0.0109	1.0399	0.7285
	PMDL	8	0.0045	1.0320	0.7337
	AIC	40	0.0093	1.1858	0.6713
	BIC	40	0.0093	1.1858	0.6713
1000	MML	8	0.0049	1.0367	0.7567
	MDL	8	0.0049	1.0367	0.7567
	CAICF	8	0.0049	1.0367	0.7567
	SRM	8	0.0049	1.0367	0.7567
	PMDL	10	0.0057	1.0427	0.7544
	AIC	30	0.0035	1.0684	0.7473
	BIC	30	0.0035	1.0684	0.7473
2000	MML	9	0.0014	1.0103	0.7636
	MDL	7	0.0047	1.0224	0.7577
	CAICF	7	0.0047	1.0224	0.7577
	SRM	7	0.0047	1.0224	0.7577
	PMDL	9	0.0014	1.0103	0.7636
	AIC	31	0.0014	1.0316	0.7562
	BIC	31	0.0014	1.0316	0.7562

Table 3.3: Performance of the different model selection methods on the task of discovering Model 1 using 4000, 6000 and 10000 sample sizes

Sample Size	Method	Model 1 (nvar= 10)			
		nvar	ModelErr	RMSE	R^2
4000	MML	9	0.0005	1.0013	0.7743
	MDL	9	0.0005	1.0013	0.7743
	CAICF	9	0.0005	1.0013	0.7743
	SRM	9	0.0005	1.0013	0.7743
	PMDL	10	0.0009	1.0016	0.7742
	AIC	38	0.0011	1.0151	0.7697
	BIC	38	0.0011	1.0151	0.7697
6000	MML	10	0.0002	1.0106	0.7697
	MDL	10	0.0002	1.0106	0.7697
	CAICF	10	0.0002	1.0106	0.7697
	SRM	9	0.0006	1.0131	0.7686
	PMDL	10	0.0002	1.0106	0.7697
	AIC	32	0.0006	1.0182	0.7672
	BIC	32	0.0006	1.0182	0.7672
10000	MML	10	0.0001	1.0116	0.7702
	MDL	10	0.0001	1.0116	0.7702
	CAICF	10	0.0001	1.0116	0.7702
	SRM	10	0.0001	1.0116	0.7702
	PMDL	10	0.0001	1.0116	0.7702
	AIC	32	0.0005	1.0162	0.7686
	BIC	32	0.0005	1.0162	0.7686

Table 3.4: Performance of the different model selection methods on the task of discovering Model 2 using 500, 1000 and 2000 sample sizes

Sample Size	Method	Model 2 (nvar= 18)			
		nvar	ModelErr	RMSE	R^2
500	MML	14	0.0068	1.0563	0.9453
	MDL	14	0.0068	1.0563	0.9453
	CAICF	14	0.0170	1.0762	0.9432
	SRM	17	0.0048	1.0635	0.9449
	PMDL	22	0.0077	1.1264	0.9388
	AIC	70	0.0130	1.4311	0.9112
	BIC	70	0.0130	1.4311	0.9112
1000	MML	17	0.0012	0.9894	0.9539
	MDL	17	0.0012	0.9894	0.9539
	CAICF	17	0.0012	0.9894	0.9539
	SRM	17	0.0012	0.9894	0.9539
	PMDL	23	0.0024	1.0123	0.9521
	AIC	70	0.0059	1.1577	0.9403
	BIC	70	0.0059	1.1577	0.9403
2000	MML	17	0.0013	1.0011	0.9509
	MDL	17	0.0013	1.0011	0.9509
	CAICF	17	0.0013	1.0011	0.9509
	SRM	17	0.0013	1.0011	0.9509
	PMDL	20	0.0011	1.0090	0.9502
	AIC	70	0.0027	1.0767	0.9447
	BIC	70	0.0027	1.0767	0.9447

Table 3.5: Performance of the different model selection methods on the task of discovering Model 2 using 4000, 6000 and 10000 sample sizes

Sample Size	Method	Model 2 (nvar= 10)			
		nvar	ModelErr	RMSE	R^2
4000	MML	18	0.0004	0.9885	0.9535
	MDL	18	0.0004	0.9885	0.9535
	CAICF	18	0.0004	0.9885	0.9535
	SRM	18	0.0004	0.9885	0.9535
	PMDL	22	0.0009	0.9955	0.9529
	AIC	70	0.0013	1.0221	0.9509
	BIC	70	0.0013	1.0221	0.9509
6000	MML	18	0.0002	1.0018	0.9518
	MDL	18	0.0002	1.0018	0.9518
	CAICF	18	0.0002	1.0018	0.9518
	SRM	18	0.0002	1.0018	0.9518
	PMDL	19	0.0002	1.0026	0.9517
	AIC	68	0.0007	1.0245	0.9500
	BIC	68	0.0007	1.0245	0.9500
10000	MML	18	0.0001	1.0014	0.9513
	MDL	18	0.0001	1.0014	0.9513
	CAICF	18	0.0001	1.0014	0.9513
	SRM	18	0.0001	1.0014	0.9513
	PMDL	20	0.0002	1.0024	0.9512
	AIC	70	0.0005	1.0159	0.9502
	BIC	70	0.0005	1.0159	0.9502

Table 3.6: Performance of the different model selection methods on the task of discovering Model 3 using 500, 1000 and 2000 sample sizes

Sample Size	Method	Model 3 (nvar= 10)			
		nvar	ModelErr	RMSE	R^2
500	MML	10	0.0027	1.0245	0.9266
	MDL	10	0.0027	1.0245	0.9266
	CAICF	11	0.0050	1.0475	0.9234
	SRM	17	0.0108	1.1242	0.9129
	PMDL	20	0.0120	1.1624	0.9074
	AIC	70	0.0145	1.5098	0.8601
	BIC	70	0.0145	1.5098	0.8601
1000	MML	10	0.0008	1.0453	0.9256
	MDL	10	0.0008	1.0453	0.9256
	CAICF	10	0.0008	1.0453	0.9256
	SRM	13	0.0035	1.0722	0.9220
	PMDL	16	0.0049	1.0875	0.9200
	AIC	70	0.0062	1.2322	0.9029
	BIC	70	0.0062	1.2322	0.9029
2000	MML	9	0.0022	0.9945	0.9307
	MDL	10	0.0006	0.9857	0.9319
	CAICF	10	0.0006	0.9857	0.9319
	SRM	10	0.0006	0.9857	0.9319
	PMDL	19	0.0030	1.0025	0.9299
	AIC	70	0.0025	1.0745	0.9216
	BIC	70	0.0025	1.0745	0.9216

Table 3.7: Performance of the different model selection methods on the task of discovering Model 3 using 4000, 6000 and 10000 sample sizes

Sample Size	Method	Model 3 (nvar= 10)			
		nvar	ModelErr	RMSE	R^2
4000	MML	10	0.0002	0.9899	0.9316
	MDL	10	0.0002	0.9899	0.9316
	CAICF	10	0.0002	0.9899	0.9316
	SRM	10	0.0002	0.9899	0.9316
	PMDL	12	0.0005	0.9907	0.9315
	AIC	70	0.0011	1.0287	0.9272
	BIC	70	0.0011	1.0287	0.9272
6000	MML	10	0.0006	1.0055	0.9300
	MDL	10	0.0006	1.0055	0.9300
	CAICF	10	0.0006	1.0055	0.9300
	SRM	10	0.0006	1.0055	0.9300
	PMDL	10	0.0006	1.0055	0.9300
	AIC	70	0.0008	1.0318	0.9270
	BIC	70	0.0008	1.0318	0.9270
10000	MML	10	0.0001	1.0207	0.9290
	MDL	10	0.0001	1.0207	0.9290
	CAICF	10	0.0001	1.0207	0.9290
	SRM	10	0.0001	1.0207	0.9290
	PMDL	11	0.0002	1.0211	0.9290
	AIC	70	0.0005	1.0349	0.9275
	BIC	70	0.0005	1.0349	0.9275

Chapter 4

An Overview of Tropical Cyclone Intensity Forecasting Modeling

4.1 Introduction

Tropical cyclones, also known as typhoons or hurricanes, are severe weather systems in the form of intense circular vortices which account for the strongest sustained winds observed anywhere in the earth's atmosphere. For the Atlantic basin which covers the areas of North Atlantic Ocean, Caribbean Sea and Gulf of Mexico, tropical cyclone intensity is defined as the near-surface sustained wind speed (1 minute averaged speed) around its eye (center).

This chapter gives an overview of the current tropical cyclone (TC) intensity forecasting models used operationally in the Atlantic basin in particular those which are built using the multiple linear regression techniques, namely Statistical Hurricane Intensity FORcasting (SHIFOR) [10], SHIFOR94 (a modification of SHIFOR) [24], and Statistical Hurricane Intensity Prediction Scheme (SHIPS) [82]. The motivation

for the move of atmospheric scientists from numerical to stochastic modelling in the attempts to build tropical cyclone intensity forecasting is also explained.

The data used to build SHIFOR and SHIFOR94 are used in the experiments done in Chapter 5. For this reason, SHIFOR and SHIFOR94 are used as benchmark models for the experiments. This chapter gives an explanation of each variable of the data used.

4.2 TC Intensity Forecasting Models for the Atlantic basin

There have been a number of modeling techniques applied by atmospheric scientists in building tropical cyclone intensity forecasting models. As shown in Table 1.1 on page 13, the Atlantic basin is one of the seven tropical cyclone basins in the world. This section outlines two categories of forecasting modeling techniques used for tropical cyclone intensity in the Atlantic basin, namely numerical and multiple linear regression modeling techniques. There are a few other modeling techniques used which will not be discussed in this thesis, namely pattern recognition (e.g. the Dvorak technique [111] which uses visible satellite pictures of a tropical cyclone to estimate its intensities), expert systems (e.g [104]) and a method that combines the Dvorak technique and other forecaster rules (e.g. [39, pp.2.24–2.27]), among others.

4.2.1 Numerical Modeling of TC Intensity Forecasting

Whilst atmospheric scientists have been able to come up with good forecasts of tropical cyclone motions using numerous numerical and statistical models which have been in operational use at the various tropical cyclone forecast centres around

the world (documented in [65]), the state of the research in tropical cyclone intensity forecasting has not been as advanced. Jarvinen and Neumann [10] explains that the disparity is partly due to the difficulty in establishing cause and effect relationships for intensity changes because of the lack of historical data to be used for building the intensity change models.

Attempts to build tropical cyclone intensity forecasting model using deterministic numerical thermodynamic modelling techniques have not met with success when the models were tried in operational use[1]. The numerical model built by Bender et al. [85, 86] consistently underforecasts intensifying tropical cyclones and overestimates intensifying tropical storms [124]. Fitzpatrick [93] suspected that the errors were associated with the lack of consensus amongst experts on how to treat properly the cumulative effects of convective clouds on the large-scale temperature and moisture fields. He explains that tropical cyclone development is sensitive to the specification of grid-scale condensation and the parameterised condensation. Cumulus parameterisation schemes assume that there is a spectral gap between resolvable and sub-grid scales, which does not exist in nature. This makes partitioning their separate contributions to tropical cyclone intensity in a numerical model unclear [123].

The problem of numerical modelling of tropical cyclone intensity becomes more difficult with the observed sensitivity of tropical cyclone simulations to different control parameters in a particular parameterisation scheme. Baik et al. [53] in his simulation experiments using the Betts-Miller convective system [7] reported that a tropical cyclone would develop faster when an "adjustment time scale" was decreased or if a "stability weight" of the moist adiabat[2] in the lower atmosphere is

[1] A similar review can be found in [93]

[2] Adiabatic Process is defined as the thermodynamic transformation which occurs without the exchange of heat between a system and its environment [64, 122]

increased. However, adjustments to these parameters could not be confidently done since they had non-consistent effects to the grid-scale convection and parameterised convection.

Baik et al. [54] also found that simulated tropical cyclones evolve differently across separate parameterisation schemes. For example, in their experiments, tropical cyclones developed faster in the Kuo parameterisation schemes [51, 52, 102] than in the Betts-Miller scheme [7]. Other researchers (e.g. [66, 69, 67]) have reported similar observations with regards to the problems in adjusting control parameters and the different behaviours of simulated tropical cyclones in different parameterisation schemes. All of these reasons motivate atmospheric scientists to try stochastic modeling approaches like multiple linear regression.

4.2.2 Multiple Linear Regression Technique in TC Intensity Modeling

This section gives overviews of three forecasting models being operationally used to forecasting tropical cyclone intensities in the Atlantic basin, namely Statistical Hurricane Intensity FORcasting (SHIFOR) [10], SHIFOR94 (a modification of SHIFOR) [24], and Statistical Hurricane Intensity Prediction Scheme (SHIPS) [82]. These models were all built using the standard least squares method to multiple linear regression outlined in Section 2.3.1.1.

This section in particular describes the data sets used to build each of the models. The data sets used to build SHIFOR94[3] are used in the experiments to build the forecasting model reported in Chapter 5.

[3]The data sets were provided by Chris Landsea

4.2.2.1 Statistical Hurricane Intensity FORcasting (SHIFOR) Model

In 1979, Jarvinen and Neumann [10] used a data set extracted from the North Atlantic tropical cyclone data tape [11] of the National Hurricane Center (NHC) to build SHIFOR. The tape contains the dates, the tracks (the global positions), wind speeds and central pressure values (if available) for all of the tropical cyclones occurring from 1886 to 1977. The information on this tape was recorded at 6-hourly intervals and was based on post analyses of all available data. These are referred to as "best track" and "best wind" data. The dates and tracks of a cyclone are categorised as the *climatology* variables and its intensity is defined as *persistence*, due to the belief that if a storm has been strengthening/weakening in the past few hours, it is most likely to continue strengthening/weakening.

Tropical cyclone intensity is recorded in terms of wind speed and central pressure. Prior to the introduction of aircraft reconnaissance flights into tropical cyclones by the United States Air Force and Navy in 1944, the amount of central pressure data recorded is small. Maximum sustained wind speeds, on the other hand, have been measured or estimated by various means for all of the tropical cyclones, although in many cases, there has been doubt that the maximum wind speed values had really been obtained [10]. For this reason, wind speed is commonly used in modeling tropical cyclone intensity.

For the Atlantic basin, tropical cyclone wind speed is measured as the maximum sustained (1 minute) surface wind speed in knots. This is measured to the nearest 5 knots. Fitzpatrick [93] explains that this is because the best track data are often estimated from satellite images (e.g. using the Dvorak technique [111]) and it is difficult to determine a representative wind speed value in a large, often asymmetrical tropical cyclone. Miller and Fritsch [27] pointed out that the areal pixel counts may

be skewed in situations when the tropical cyclone is far from the satellite subpoint. Because of the low level of precision in the tropical cyclone data, it is common practice to predict intensity change rather than intensity in tropical cyclone forecasting models.

SHIFOR was built using the tropical cyclone data from 1900 to 1977. Data prior to 1900 were not used because of fragmented documentation [10]. The data set was divided into two parts: data from 1900 to 1972 as the training data set and data from 1973 to 1977 as the test data set. Forecasting models for 12 through to 72 hours into the future were built.

Several constraints were placed upon the data set:

1. Only the tropical cyclone data records within the geographical area bounded by 45°C latitude on the north, the equator on the south, 5°C longitude on the east, and the North, Central and South American Continents on the west, were accepted.

2. Only those tropical cyclone data which were not within 30 nautical miles of land and had previous 6- and 12-hour positions not within 30 nautical miles of land were accepted.

3. All wind values had to equal or exceed 35 knots (i.e. at tropical storm strength or greater).

4. If a storm moved inland and later moved our over water, its records that could not meet constraints 1, 2, or 3, were eliminated.

There are seven regressors basic regressors used in SHIFOR:

1 - **Julian Date** The number of days since the start of the year in Julian Calendar. The effect of leap years is neglected in the Julian date variable. Here the end of each month from December to November is represented respectively by the numbers 0, 31, 59, 90, 120, 151, 181, 212, 243, 273, 304 and 334. The Julian Date is calculated by adding the day of the cyclone data item to the number associated with the end of the previous month.

2,3 - **Initial Latitude and Longitude** The global position of a cyclone (in degrees North and West, respectively)

4 - **Average zonal speed past 12 hours** The average speed of a cyclone in the east to west direction (in knots)

5 - **Average meridional speed past 12 hours** The average speed of a cyclone in the north to south direction (in knots)

6 - **Current maximum sustained wind speed** The 1 minute sustained wind speed of a cyclone (in knots)

7 - **Previous 12 hour change in maximum sustained wind speed** The current wind speed minus the wind speed 12 hours prior (in knots)

The SHIFOR model in operational use in the Atlantic basin has been modified by Arthur Pike in the late 1980s[4]. This modified version of SHIFOR which includes the products of variables is used as a benchmark model in the experiments outlined in Chapter 5.

[4]Chris Landsea, personal communication

4.2.2.2 Modified Statistical Hurricane Intensity FORcasting (SHIFOR94) Model

The motivation of SHIFOR94 [24] is to find a forecasting model that can produce better forecasts 72 hours into the future than SHIFOR[5]. In addition to the climatology variables and persistence used in SHIFOR, Chris Landsea includes synoptic/environmental and seasonal data in the search space to find SHIFOR94. Seasonal variables have been reported (e.g. in [114, 115, 116, 118, 119, 120, 121]) to be influential to the frequency/activity of tropical cyclones in the Atlantic basin.

SHIFOR94 was built using tropical cyclone data from the period 1950-1994. The tropical cyclone data will not be used if the cyclone satisfies one of the following conditions (Chris Landsea, personal communication):

1. Tropical cyclones which did not last for 72 hours
2. Tropical cyclone records which were at either extra-tropical or sub-tropical stages
3. Tropical cyclone records with winds either at initial or final forecast time were less than 20 knots
4. Tropical cyclones which occurred before 1 June. This is because Atlantic basin tropical cyclone activity either before 1 June or after 30 November is nearly negligible and most cyclones occur between 1 August and 31 October [23].

Upon observations of the recordings of tropical cyclone central pressure and wind speed, Landsea [23] noticed that for specific wind speed categories, there has been a

[5]The limited published information on SHIFOR94 can only be found in [24], a short paper. Explanations on SHIFOR94 and the climatological, persistence and synoptic variables used to build it outlined in this section are derived from email communications with Chris Landsea, and the Fortran programs and data files he has made available to us.

shift to lower observed pressures for the decades of the 1940s to the 1960s. Since there have been no significant changes in the methodologies to measure actual and extrapolated surface pressures in tropical cyclones, Landsea deduced that any changes in the wind-pressure relationship might be due to alterations in the way sustained wind speeds were measured or estimated. He concluded that the decades of 1940s to the 1960s had overestimated wind speeds as compared to later years. The tropical cyclone data set is then corrected by removing the bias in the wind speed records for the years 1944 to 1969. The wind speeds between 100 to 124 knots were reduced by 5 knots and those greater and equal to 125 knots were reduced by 10 knots.

The tropical cyclone track position forecasting program called, CLIPER (CLimatology and PERsistence) [15], is used to get the forecasts of the latitude and longitude position and motion of a cyclone 72 hours in the future. CLIPER is the only purely statistical tropical cyclone track forecasting model still in operational use. The only inputs to CLIPER are the storm's current and previous positions, its motion and intensity, and the time of the year. Because CLIPER knows nothing about the meteorological situation surrounding a storm, the output of CLIPER is called "no skill" prediction. For the 72-hour forecast period, CLIPER requires that tropical cyclone must have existed in some form 18 hours prior. The cyclone data which have transited over land at anytime during the 72 hour period are removed from the data set.

There are in total 36 single variables in the pool of potential regressors to build SHIFOR94. The search space used to find SHIFOR94 includes the products and ratios of these single variables. These 36 single variables and their products are used in the experiments in Chapter 5. Thus, the second-order polynomial forecasting model tested takes the form as shown in Equation 2.1 on page 23.

The climatology and persistence variables are as follows:

1,2 - **Julian Date, JulOff** Variable Julian Date is the same as the Julian Date calculated in SHIFOR. JulOff measures the number of days from the peak of the hurricane season in the Atlantic basin. It is calculated as the absolute value of substracting 273 from Julian Date shown in the following formula:

$$\text{JulOff} = |\text{JulianDate} - 273| \tag{4.1}$$

The number 273 is associated with the date September 10, when the peak of the Atlantic tropical cyclone season in terms of the probability that a storm exists anywhere in the Atlantic basin is a maximum [82].

3,4 - **LatData, LonData** LatData and LonData are the values of the initial latitude and longitude global position of a cyclone (in degrees North and West, respectively) as used in SHIFOR.

5,6 - **Vmax, Del12V** Vmax is the current cyclone intensity, also known as "persistence". It is the current maximum (1 minute) sustained wind speed of a cyclone (in knots). Del12V is the previous 12 hour change in maximum sustained wind speed.

7,8,9 - **UCurr, VCurr, Speed** UCurr is the average zonal (i.e. in the east-west direction) speed/motion of a cyclone for the past 6 hours (in knots). VCurr is the average meridional (i.e. in the north-south direction) speed/motion of a cyclone for the past 6 hours (in knots). Speed is the resultant of UCurr and VCurr. Speed is included because the speed of a motion might be related to intensity change, independent of the direction of storm motion [82].

The synoptic environmental variables are:

10,11 - **POT, POTend** POT is the initial potential intensity (intensification potential) of a cyclone. Potential Intensity is defined as the theoretical maximum possible intensity that can be sustained for the current environmental conditions; normally related to ocean temperature and tropopause height and temperature [70].

Potential Intensity is calculated by taking the difference between the Maximum Potential Intensity and the current cyclone intensity. Maximum Potential Intensity (MPI) [105, 71] is the upper bound of the intensity of a cyclone estimated from the sea surface temperature, SST, at the location of the cyclone. For SHIFOR94, and thus the experiments done in Chapter 5 in this thesis, MPI is determined from the empirical relationship given below which was developed by DeMaria and Kaplan [81] from a 31-year sample of Atlantic cyclones.

$$\mathrm{MPI} = A + B^{C(\mathrm{SST}-\mathrm{SST_0})} \tag{4.2}$$

where:

A = 66.5 knots

B = 108.8 knots

C = $0.1813(^{\circ}\mathrm{C})^{-1}$

$\mathrm{SST_0}$ = 30.0°C

SST : the sea surface temperature in °C

Monthly mean values of sea surface temperature (SST) are available on a 1° latitude-longitude grid. These values are linearly extrapolated in space and

time to the position and date of each storm. The SST for the calculation of POT is averaged over the track of the cyclone during the forecast interval to account for SST variations along the storm track. For example, for the 72-hour forecast, the SST for Equation 4.2 is the average of the SST at 0-, 24-, 48- and 72-hour positions of the cyclone. For the calculation of POTEnd, SST is taken at the SST at the end of the forecast interval, i.e. at the position for the 72-hour forecast.

12 - **DelSST** DelSST is the difference between the current sea surface temperature and the SST at the end of the forecast interval.

13,14,15 - **SSST, SSSTend, DSSST** The climatological sea sub-surface temperature data available are averaged for the depths 10m, 30m and 50m. SSST is calculated as the average of the sub-surface temperatures at 0-, 24-, 48- and 72-hour positions of the cyclone. SSSTend is the sea sub-surface temperatures at the 72-hour position of the cyclone. DSSST is the difference between the sea sub-surface temperatures at the initial position of the cyclone and the position at the end of the forecast period.

16,17,18 - **UppSpd, Uppend, DUppSpd** Studies, e.g. [113, 106], have shown that the vertical shear of the horizontal wind has a negative influence on tropical cyclone intensification. It is, however, observed that there is no consensus on how the vertical shear that has an influence on tropical cyclone intensification is defined. For SHIFOR94, the shear is defined as the vertical shear between the tropical cyclone horizontal wind motion and the climatological 200mb wind. So, UppSpd is the vertical shear defined for SHIFOR94, averaged over the 0-, 24-, 48- and 72-hour positions of the cyclone. Uppend is the

shear at the position at the end of the forecast period. DUppSpd is the change in wind shear from the initial to the end of forecast period.

For SHIPS [82], which will be described in the next section, the shear is defined as the difference between the 850mb and 200mb climatological wind vectors [82]. The reason for the choice of these two levels to evaluate shear is because most of the satellite cloud track wind estimates that are used to calculate for the angular momentum variable used in SHIPS are assigned to these levels.

19,20,21 - **Stabil, Stabend, DelStab** Stabil is the moist static stability between 1000 and 200mb averaged over the 0-, 24-, 48- and 72-hour positions of the cyclone. Stabend is the moist static stability at the end of the forecast period. DelStab is the change in moist static stability from the initial to the end of forecast period.

22,23,24 - **200mbT, 200Tend, Del200T** 200mbT is the temperature at 200mb averaged over the 0-, 24-, 48- and 72-hour positions of the cyclone. 200Tend is the temperature at the end of the 72-hour forecast period. Del200T is the difference between the 200mb temperature at the initial position of the cyclone and the position at the end of the forecast period.

25,26 - **DisLand, Closest** DisLand is the initial distance of a cyclone from land based on the available land database. Closest is the closest approach of a cyclone to land.

27,28,29 - **200mbU, 200Uend, Del200U** 200mbU is the climatological 200mb wind averaged over the 0-, 24-, 48- and 72-hour positions of the cyclone. 200Uend is the climatological 200mb wind at the end of the forecast period. Del200U is the difference between the 200mb wind at the initial position of the cyclone and the position at the end of the forecast period.

Seasonal variables have been reported (e.g. in [114, 115, 116, 118, 119, 120, 121]) to be influential on the activity of tropical cyclones in the Atlantic basin. The Atlantic tropical cyclone activity is defined in 7 variables: the seasonal total numbers of named storms (NS), hurricanes (H), intense hurricanes (IH), named storm days (NSD), hurricane days (HD), intense hurricane days (IHD) and hurricane destruction potential (HDP). The definitions of these activity variables are contained in [118, 119, 23] and are summarised below:

Named Storms (NS): a hurricane or a tropical storm.

Hurricanes (H): a tropical cyclone with sustained low-level winds of 33 m/s or greater.

Intense hurricanes (IH): a hurricane reaching at some point in its lifetime a sustained low-level wind of at least 50 m/s. This constitutes a category 3 or higher on the Saffir-Simpson scale shown in Table 1.2 on page 14.

Named storm days (NSD): four 6-hour periods during which a tropical cyclone is observed or estimated to have attained tropical storm or hurricane intensity wind.

Hurricane days (HD): four 6-hour periods during which a tropical cyclone is observed or estimated to have hurricane intensity winds.

Intense hurricane days (IHD): four 6-hour periods during which a hurricane has intensity of Saffir-Simpson category 3 or higher.

Hurricane destruction potential (HDP): a measure of a hurricane's potential for wind and storm-surge destruction defined as the sum of the square of a hurricane's maximum wind speed for each 6-hour period of its existence. Values are given in $0.25 \times 10^4 (m/s)^2$.

The following are the seven seasonal variables included in the search space to find SHIFOR94. Figure 4.1 gives the locations of these variables. Whilst the influences of these seasonal variables on the activity of tropical cyclone in the Atlantic basin have been extensively studied in the past, their influences on the intensities of tropical cyclones in the basin are much less clear.

30 - **U50** U50 is the 50mb Quasi-Biennial Osciallation (QBO). QBO refers to variable east-west oscillating stratospheric winds which circle the globe near the equator. Strong stratospheric QBO easterly winds and strong lower-stratospheric vertical wind shear conditions inhibit lower-latitude tropical cyclone formation and intensification [114, 115, 80]. On average, there is nearly twice as much intense (i.e. category 3,4,5 on the Saffir-Simpson Scales shown in Table 1.2 on page 14) Atlantic basin hurricane activity during seasons when equatorial stratospheric winds at 30mb and 50mb (i.e. 23km and 20km altitude, respectively) are more westerly as compared to when they are from a more easterly direction [116].

31,32 - **RainS, RainG** The incidence of intense Atlantic hurricane activity is strongly enhanced during the seasons when the June–July African Western Sahel (5°W–15°W, 10°N–20°N) and previous year August–November Gulf of Guinea regions of West Africa (0°W–10°W, 5°N–10°N) have above average rainfall[121]. RainS and RainG refer to the rainfall indices in these two regions. Landsea and Gray [25] have found that the previous-year rainfall along the Gulf of Guinea and in the Sahel itself provides a very dependable indication of future Sahel rainfall. This rainfall is hypothesized to lead to earlier vegetation growth and greater amounts of soil moisture and evapotranspiration on the following August and September period. This extra moisture source appears to lead to

a higher percentage of stronger and more concentrated waves coming off West Africa to the Atlantic.

Landsea and Gray [25] have also found that there is a strong statistical relationship between western Sahelian rainfall and Caribbean basin upper-tropospheric zonal winds during the height of the hurricane season in September. Upper-tropospheric zonal winds are typically weak westerlies or actual easterlies during times of heavy western Sahelian rainfall. Inversely, they are typically strong westerly winds during western Sahelian drought conditions. The influence of the Caribbean basin upper-tropospheric zonal winds (ZWA) on the Atlantic tropical cyclone activity is explained in the next point. Gray et al. [120] conclude that interannual variability of easterly waves and the general circulation over the Atlantic appears to be the mechanism that ties the western Sahelian rainfall to Atlantic tropical cyclone activity.

33,34 - **SLPA, ZWA** SLPA and ZWA are the Caribbean Sea Level Pressure Anomalies and 200mb (12km altitude) Zonal Wind Anomalies, respectively. Stations located throughout the Caribbean basis show an inverse relationship of April and May sea level pressure (SLPA) to subsequent tropical cyclone activity [120]. In general, higher pressure precedes quiet conditions, while lower pressure indicates more activity to come [68, 61, 78, 115]. Surface pressure variations are observed to be associated with interannual shifts of the location and/or fluctuations of the intensity of the intertropical convergence zone (ITCZ)[6]. Negative SLPA values typically indicate a shift of ITCZ farther to the north of its normal position, which might favour tropical cyclone genesis. It is presumed than anomalous conditions occuring in April and May

[6]intertropical convergence zone (ITCZ) is a discontinuous belt of thunderstorms paralleling the equator and marking the convergence of the northern and southern hemisphere surface trade winds [90]

have a tendency to persist through the main months of the hurricane season (August–October) [120].

Tropospheric vertical wind shear has long been recognised as a major inhibiting factor for tropical cyclonegenesis and intensification [113]. Gray et al. [120] explains that because of the circulation regime of the tropical North Atlantic, tropospheric vertical shear is dominated by the variations in the upper troposphere. Thus with the nearly constant trade-wind flow (e.g. easterlies) near the surface, 200mb ZWA adequately describe vertical wind shear. That is, positive anomalies (westerly) indicate enhanced shear and less tropical cyclone activity, while negative anomalies (easterly) indicate reduced shear and more tropical cyclone activity.

Gray et al. [120] further explains that, similarly to the Caribbean SLPA, the April and May ZWA are useful as predictors because of their tendency to persist into the heart of the hurricane season. Also, because of the anomalous circulation forced by ENSO events (see the explanation on ENSO in the next point) and the location of the Caribbean stations chosen, the 200mb ZWA are good measures of the ENSO effect upon the Caribbean and the western North Atlantic [95, 114].

35,36 - **ElNino, SOI** ElNino is the ENSO (El Niño-Southern Oscillation) influence. ENSO characterises the sea surface temperature anomalies (SSTA) in the eastern equatorial Pacific and the surface pressure gradient between Tahiti and Darwin, i.e. the sea level pressure of Tahiti minus that of Darwin, also known as the Southern Oscillation Index (SOI). As explained in [121], a moderate or strong El Niño event (i.e. warm water and low surface pressure gradient between Tahiti and Darwin) in the eastern equatorial Pacific is observed to

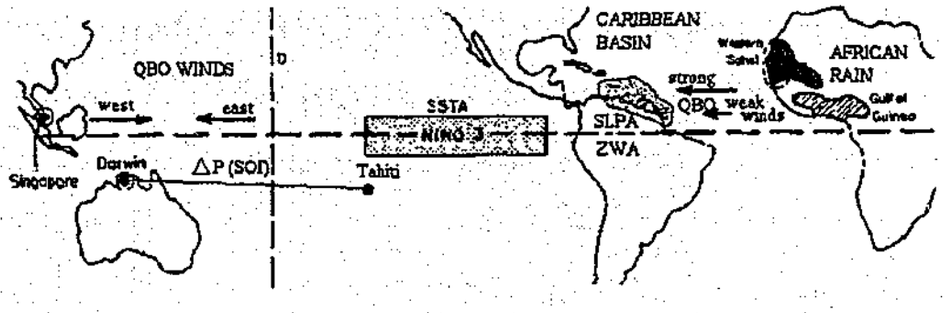


Figure 4.1: Locations of the seasonal/environmental variables reported to be influential on tropical cyclones in the Atlantic basin. *Source:* [121]

have inhibiting effects on the Atlantic basin hurricane activity. On the other hand, more active Atlantic basin hurricane seasons happen during La Niña years (i.e. cold water and high surface pressure gradient between Tahiti and Darwin).

The above observation is explained in a wider context in [95, 114, 79, 117, 119, 120]. These papers explain that there is a strong relationship between ENSO, the 200mb (12km) circulation over the Carribbean basin and tropical Atlantic upper-tropospheric zonal winds. In the seasons of warm eastern Pacific SSTA or when low values of SOI occur, the Caribbean basin and tropical Atlantic typically have positive ZWA. This inhibits Atlantic activity. Opposite or weak upper-tropospheric zonal winds occur in the seasons of cold easter Pacific SSTA and high SOI. This latter conditions enhance Atlantic activity.

4.2.2.3 Statistical Hurricane Intensity Prediction Scheme (SHIPS) Model

SHIPS [82] was built using the same climatological, persistence and synoptic[7] variables which would later be used in building SHIFOR94 outlined in the previous

[7]Except that wind shear for SHIPS is defined differently from that for SHIFOR94 as discussed in Section 4.2.2.2

section. Instead of using CLIPER [15], VICBAR [83], is used to get the tropical cyclone track forecasts. VICBAR replaces the complex dynamics of the atmosphere with a layer of fluid with constant density representing the average motions between 850mb and 200mb.

Three additional variables were also used in SHIPS namely, the 200mb relative and planetary angular momentum flux convergence, REFC and PEFC respectively, and the 850mb relative angular momentum, SIZE. The data used are for the period 1989 to 1992 with a few additional cases from 1982 and 1988. DeMaria and Kaplan [82] explain that the momentum flux variables (REFC and PEFC) are included to account for positive interactions between the tropical cyclone and synoptic-scale weather systems (i.e. 50km – 5,000km). They further explain that the theoretical results in [41] and the observational results in [55] suggest that when a storm interacts with the large-scale flow in a way that makes the upper-level flow more cyclonic, the intensification rate of the storm may be increased. The integrated relative angular momentum, SIZE, is included as a measure of the extent of the outer circulation of a tropical cyclone. SIZE is calculated using satellite cloud track winds at 850mb representing the outer circulation of a storm. SHIPS is not used as a benchmark model in this thesis because the data required to calculate the three additional independent variables mentioned above were not available.

4.3 Conclusion

This chapter gives an overview of some of the models used to forecast tropical cyclone intensity change in the Atlantic basin. Of particular importance to this thesis are the explanations of the 36 variables used to build SHIFOR94 [24]. The same data set used to build SHIFOR94 is used in the experiments to build a new forecasting

model done in Chapter 5. Of particular interest are the seasonal variables (variables 30–36), i.e. QBO, African rainfall, Current ENSO conditions, Caribbean pressures and 200mb winds. These variables had never previously been used to build tropical cyclone intensity forecasting model. This is despite the strong observed relationships between these seasonal variables and the Atlantic tropical cyclone activity, as has become obvious from the explanations of each individual variable in this chapter. Chapter 5 gives the comparisons between the new model proposed in this thesis and SHIFOR and SHIFOR94.

Chapter 5

A New Tropical Cyclone Intensity Forecasting Model: Balancing Complexity and Quality of Fit

5.1 Introduction

Building forecasting models for tropical cyclone intensity is one of the most challenging area in tropical cyclone research. As explained in Section 4.2 (see also [24]), numerical models have not been able to demonstrate real-time improvement over no-skill predictions (i.e. simple extrapolation of the trend in tropical cyclone intensity data, known as *persistence*) because of the strong interactions between mesoscale and synoptic features in the atmosphere leading to tropical cyclone intensity change. Tropical cyclone intensity data are recorded in sparse intervals of 5 knots prompting the common use of intensity change (strengthening or weakening) within a specified

amount of time into the future, e.g. 12, 24, 48, 60, or 72 hours, as the dependent variable of tropical cyclone forecasting models.

As explained in Section 4.2, there are three statistical tropical cyclone forecasting models in operational use in the Atlantic basin: SHIFOR [10]), SHIFOR94 [24] and SHIPS [82]. All of the three statistical tropical cyclone forecasting models have been built using the least squares method, a variant of the Maximum Likelihood (ML) approach outlined in Section 2.3.1.1, which belongs to the first category of model selection criteria outlined in Section 2.3.2. With this method, separate 'test' (i.e. 'semi' independent) data sets are needed to guide the search for a set of independent variables to form an optimum forecasting model. This implies that the predictive performance reported on the test data sets should not be seen as reflecting performance on a completely independent data sets. The need to partition data into two sets for model development is seen to be one of the drawbacks of the ML approach in the face of limited available data.

This chapter[1] uses the four complexity-penalised model selection criteria, namely MML, MDL, CAICF and SRM to build tropical cyclone intensity forecasting models. These criteria which are outlined in Section 2.3.3, have been proven to be robust for fully automated model selection tasks in the experiments using artificial data in Chapter 3. The data sets used to build SHIFOR94 [24] are used in the experiments in this chapter. Each variable in the data sets is explained in Section 4.2.2.2 in Chapter 4. The models discovered in this chapter are compared with SHIFOR [10] and SHIFOR94 [24].

The experiments in this chapter started with running the common non-backtracking optimisation search algorithm using each of the four model selection criteria on 10

[1] An earlier version of this chapter has been published in [43]

pairs of training-test data sets. The four criteria have produced competitive forecasting models, making it hard to choose which model to pick. This has prompted us to propose a new model selection strategy that builds new forecasting models based on the results of the four model selection criteria.

5.2 Building the TC Intensity Forecasting Models

The forecasting models to be built typically take the form of a polynomial regression model to the second-order by considering products of two single variables as well as the single variables as shown in Equation 2.1 on page 23. Table 5.1 summarizes the set of methods used in this chapter. The non-backtracking optimisation search algorithm explained in Section 3.2.2 is used for all of the methods. The performance criteria for model comparison are

1. Parsimony, reflected in the number of regressors chosen and the model cost calculated using any of the cost functions given in Table 5.1 (on training data). In this chapter; the MML message length is used to reflect model cost.

2. Model predictive performance (on test data), the square root of the mean of the sum of squared deviations (RMSE) and the coefficient of determination (R^2) respectively given in Equations 3.39 on page 69 and 3.40 on page 70.

Two types of experiments are done in this chapter. The first is to find forecasting models by running the search method with one of the four model selection methods as the cost function on a number of data sets. The second is to find forecasting models based on the results of the first type of experiments. Procedure 1 gives the proposed procedure for the second type of experiments. This procedure is proposed to be used in a situation where we have a number of competitive model selection

methods and all of them have produced different but equally good forecasting models. To help us make a decision as to what model to use operationally, we can build new models based on the the results of all of the competitive model selection methods. This way we can make an informed decision based on how far the model selection methods agree on a model.

Procedure 1 Searching for a forecasting model using an integrated approach consisting of a number of model selection criteria

Step 1 Using random subsampling, create m pairs of training and test data sets.

Step 2 For each of the methods in Table 5.1, search for forecasting models on each training data set using the optimisation search algorithm proposed in Chapter 3.

Step 3 Count the frequency of each independent variable's being chosen in any of the forecasting models discovered in the previous step. Create a set of new models where each new model comprises a set of variables which have been chosen in the forecasting models at least a certain total number of times. Calculate the set of coefficients for each model using all of the available data.

Step 4 Compare the models with increasing complexity based on all of the performance criteria.

Although subsampling is used to produce both training and test data sets in Step 1 in Procedure 1 above, the test data is not used in Procedure 1. However results from running the models found using Procedure 1 on the test data are presented in this chapter to see if the models generalize well.

Table 5.1: Summary of the model selection criteria used in this chapter for the task of finding a tropical cyclone intensity change forecasting model from a set of climatology, persistence, synoptic/environmental and seasonal data sets.

Method		Reference	Equation	Page No.
Minimum Message Length	MML	[19]	2.32	40
Predictive Minimum Description Length	PMDL	[59]	2.43	45
Minimum Description Length	MDL	[59]	2.44	46
Structured Risk Minimisation	SRM	[109]	2.57	53

5.3 Experimental Design

For the task of building hurricane intensity change forecasting models, it is actually possible to reduce the search space considerably by using only the combinations of variables that would likely be influential to the target variable and, after consultations with meteorologists, removing the combinations that are implausible. This in fact has been the common approach in hurricane research due to the limitations of the multiple linear regression methods used.

Based on the results of their experiments, most atmospheric scientists have come to the conclusion that the search space should be limited to potentially significant variables because it is believed that even random numbers will inadvertently be selected as significant predictors, e.g. [10, 96, 82]. This approach however is not taken in this chapter for two reasons:

- It has been proven using artificial data sets in Chapter 3 that the methods used in this chapter managed to choose parsimoniously a handful of significant regressors amongst a large pool of variables

- Part of the goal of these experiments is to test whether or not the above conclusion applies to real and more complex problem domains like the tropical cyclone systems.

5.3.1 Potential Predictors

The task of building tropical cyclone intensity change forecasting models is difficult due to the following reasons:

- The fact that exact relationships amongst the variables are not known
- A lot of the independent variables are highly preprocessed, either using a thermodynamic formula chosen from a number of possibilities (e.g. Potential Intensity as a function of sea surface temperature, see [82] or a completely stochastic procedure)
- Noise inherent in observational data.

Table 5.2 gives the summary of the independent variables used to build the tropical cyclone intensity change forecasting models which can be categorized into four groups as explained in Section 4.2.2.2:

1. **Persistence:** the tropical cyclone intensity in knots (Regressors 5,6)
2. **Climatology:** Julian date (Regressors 1,2), global position (Regressors 3,4) and motion in knots (Regressors 7,8,9)
3. **Synoptic environmental features:** potential intensities (Regressors 10,11), sea surface temperature (Regressor 12), sea sub-surface temperatures (Regressors 13–15), wind shear (Regressors 16–18), moist stability (Regressors 19–21),

Table 5.2: Basic regressors used to build the Atlantic tropical cyclone intensity change forecasting models summarised from Section 4.2.2.2. The target variable is the change of intensity (wind speed) 72 hour into the future. To get the *average, at end* and *change* values of a variable, the tropical cyclone track/location (longitude and latitude) forecast out to 72 hours is required: *average* means the average of the values at the location forecast at 0, 24, 48 and 72 hours, *at end* means the value at 72 hours and *change* means the difference between the current value and the value at 72 hours. Figure 4.1 on page 4.1 illustrates the location of the seasonal variables, i.e. variable 30 to 36

No	Basic Regressor: acronym and explanation
1,2	Julian, JulOff – date: Julian, |Julian - 253|
3,4	LatData, LonData – position: latitude (deg N), longitude (deg W)
5,6	Vmax, Del12V – intensity (in knots): initial, previous 12 hour change
7,8,9	UCurr, VCurr, Speed – motion (in knots): east-west, north-south, resultant
10,11	POT, POTEnd – potential intensity: initial, at end
12	DelSST – change of sea surface temperature
13,14,15	SSST, SSSTend, DSSST – sea sub-surface temperature: average, at end, change
16,17,18	UppSpd, Uppend, DUppSpd – windspeed at 200mb: average, at end, change
19,20,21	Stabil, Stabend, DelStab – moist stability 1000mb to 200mb: average, at end, change
22,23,24	200mbT, 200Tend, Del200T – temperature at 200mb: average, at end, change
25,26	DisLand, Closest – distance from land: initial, closest approach
27,28,29	200mbU, 200Uend, Del200U – east-west motion at 200mb: average, at end, change
30	U50 – 50mb Quasi-Biennial Oscillation (QBO) zonal winds
31	RainS – African Western Sahel rainfall index (5W-15W, 10N-20N)
32	RainG – African Gulf of Guinea rainfall index (0W-10W, 5N-10N)
33	SLPA – April-May Caribbean basin Sea Surface Pressure Anomaly
34	ZWA – April-May Caribbean basin Zonal Wind Anomaly at 200mb (12 km)
35	ElNino – Sea surface temperature anomaly in the eastern equatorial pacific
36	SOI – Surface pressure gradient between Darwin and Tahiti

temperature at 200mb (Regressors 22–24), distance from land (25–26), and motion in the east-west direction (Regressors 27-29)

4. **Seasonal variables:** QBO, Rainfall (RainS and RainG), SLPA, ZWA, ElNino and SOI (30–36)

In this chapter, the target variable is the intensity change 72 hours into the future. The tropical cyclone track forecasting model CLIPER [1.5] is used to provide future forecast positions for which the climatological independent variables are calculated as explained in Section 4.2.2.2.

5.3.2 Sample Sets

There are in total 4347 data items available. The experiments are conducted on data sets built using two types of sampling method. Ten training–test data sets were built using a 2 : 1 random sampling method. An eleventh data set was built so that the training data was taken from the years 1950–1987 and the test data from the years 1988–1994. Convenient separation of data based on consecutiveness is common practice in hurricane intensity change forecasting. SHIFOR (modified by Pike from [10]) and SHIFOR94 [24] have both been built using training data from 1950 to 1987 and test data from 1988 to 1994. The purpose of using these two categories of data sets in this chapter is to see whether or not the possible changes in atmospheric dynamics from year to year should be taken into account in experimental design.

We have been unable to come up with exactly the same parameter coefficients for the variables in SHIFOR and SHIFOR94 as reported in [10] and [24] despite the fact that the same data sets and least squares method have been used. It is not clear how the coefficients in SHIFOR and SHIFOR94 were normalized/standardized. Within this thesis, coefficients are standardized by subtracting from each variable value its mean and then scaling the result by its standard deviation as shown in Equations 2.2 and 2.3 on page 24. Because of the previously mentioned problem, with respect to the benchmark models, it was decided to run two types of experiment. In the first type of experiments, new coefficients for the variables were found using each training data set and tested using the corresponding test set. The new coefficients were calculated using the orthogonal transformation and the hyper-parameter α shown in Equation 3.37 on page 68, as it was done with the other methods considered in this thesis. The benchmark models with the new coeficients are named SHIFOR'

and SHIFOR94'. In the second type of experiments, the models SHIFOR and SHIFOR94 as reported in the papers were simply run on the test data sets. Because the Fortran program made available to us to run these experiments only calculates RMSE, R^2 values are not reported for these experiments. Because message length is calculated using standardized data and coefficients and it is not clear how the data were standardized for SHIFOR and SHIFOR94, the message lengths for these models cannot be calculated.

The fact that the models were built using the maximum likelihood method (with the test data being used to guide the search for model) implies that SHIFOR and SHIFOR94 have 'seen' all of the data items available. This means that the resulting predictive performance should be less conservative than if the models have been tested on completely independent data sets.

5.4 Results and Discussions

Tables 5.3 and 5.4 show the models chosen by each method for each of the 10 data sets. Although Predictive Minimum Description Length criterion (PMDL) has been proven to be a good candidate for automated model discovery using artificial data in Chapter 3, it failed to converge to an optimum model before the maximum number of variables set for a model has been reached. Therefore, it is decided not to use nor show the results of PMDL.

Comparing the performance of MML against the other three methods of the complexity-penalised criteria in Tables 5.3 and 5.4, we see that MML consistently picked less complex models with either better or slightly worse performances based on message length and generalisation ability on the test data. The fact that the search strategy using MML as a cost function chose less complex models with worse

performance in data sets 4 and 10 for example, shows that the search strategy did not find model with the minimum message length. With improvements in the message length calculation (like by using Equation 2.25 instead of 2.26 on pages 35 and 36, respectively) and the search strategy by letting it to start from different points in the search space, MML might have chosen slightly more complex models than the ones it had picked but with better performance than the ones chosen by the other three methods.

The consistent performance of SHIFOR and SHIFOR94 across all of the data sets including the last data set (where the training and test data sets are taken from consecutive years) should be taken with caution because of the way the models were built. As has been explained in Section 5.1, the performance on test data set has been used to guide the search strategy in building SHIFOR and SHIFOR94. Hence, the test data has been used as a part of the 'training' data set. This implies that the performance of SHIFOR and SHIFOR94 on the 'test' data set cannot be seen as an indication of its performance on a completely independent data set. So, it is not surprising to see that there no significant difference between the performances of SHIFOR' and SHIFOR94' on the training data and their performances on the test data.

All of the complexity-penalised methods perform substantially worse on the last data set than on the first 10 data sets. This phenomenon can be observed both by comparing their predictive performance between the training and test data for the last data set and by comparing the level of model parsimony and predictive performance on test data across the 11 data sets. The difference in RMSE values of MML, MDL, CAICF and SRM between training and test data sets for data sets 1 to 10 is always much less than 1 knot. The difference for data set 11, on the other hand, is around 6 knots. This difference between the performance on the training

data and that on the test data is more pronounced in the R^2 values. It is clear from these observations, that partitioning training and test data based on consecutiveness like in data set 11, the data set used to build SHIFOR and SHIFOR94, has resulted in two non-homogeneous data sets. A lot of the regularities learned by a model from the training data set are not present in the test data set resulting in much reduced performance.

Test data sets are created in Step 1 of Procedure 1 not for the purpose of model development but to see whether or not the performance of a method on unseen data is much different from that on training data. For each of the first 10 data sets, there is not too much difference between the predictive performance of MML, MDL, CAICF and SRM on the training data and on the test data. This confirms the findings in Chapter 3 and strengthens the belief that all of the available data can be used as training data since overfitting is not a problem for these methods. The competitive performance amongst all of the four methods on the first 10 data sets prompts us to propose Procedure 1 shown on page 102 to make use of the results of the four methods to build a new model.

Following the third step of Procedure 1, Table 5.5 shows the models built by categorizing variables based on the frequency of being chosen in all of the 10 data sets in Tables 5.3 and 5.4. It is not surprising that Potential Intensity (variable 10 - POT) and intensity change during the previous 12 hours (variable 6 - Del12V) have been chosen by all of the methods for the 10 data sets. Potential Intensity as a function of sea surface temperature indicates the maximum wind speed to which a tropical cyclone can intensify should the system not be perturbed by dampening factors in the environment.

Table 5.6 shows the performance of the models built in Table 5.5 using all of the available data following Step 3 of Procedure 1. All of the cost values calculated using MML, MDL, CAICF and SRM show that Model_{18} is the best model. However, since both SHIFOR and SHIFOR94 comprise 9 variables, Table 5.7 compares them with Model_7, the model with the best 9 variables. All of the three models, Model_7, SHIFOR and SHIFOR94, only agree on one compound intensity variable, variable (6,5) or (Del12V*Vmax). Among the independent variables explained in Section 5.3.1, SHIFOR was built using persistence and climatology. SHIFOR94, on the other hand, was built using all of the variables in Table 5.2. On top of the single and products of variables as considered in this thesis, both SHIFOR and SHIFOR94 also used ratios of variables in the search space.

As explained above, it is not a surprise that POT and Del12V are included in both SHIFOR94 and the new model. As explained in Section 4.2.2.2, the vertical shear of the horizontal wind has been reported to have a negative influence on tropical cyclone. One of the shear variables between the 200mb climatological wind and the cyclone horizontal wind, Uppend is chosen in SHIFOR94. None of the three shear variables is chosen in the new model. Instead, two of the three variables of the climatological horizontal wind namely, 200mbUend and Del200U, are included in the new model. As explained in Section 4.2.2.2, there is no consensus on how the shear should be defined. Therefore, we think that the inclusion of variables (200mbUend,Vmax) with coefficient -0.119977 and Del200U with coefficient 0.109351 in the new model can be seen as a representation of the influence of shear on tropical cyclone intensification.

It is interesting to note, that in contrast of the prevalence of the seasonal predictors in the new model, none of them is chosen for SHIFOR94 despite the fact that they have successfully been used as predictors of Atlantic cyclone activity (in

terms of named storms, named storm days, hurricanes, hurricane days, intense hurricanes, intense hurricane days, etc) as described in Section 4.2.2.2. For example, it has been reported that the effect of moderate or strong ElNino (warm water) and low SOI values reduces Atlantic basin hurricane activity. This is because during ElNino seasons, ZWA and SLPA are enhanced creating strong vertical shear over the Atlantic. By contrast, cold water and high SOI values (i.e. La Nina) enhances Atlantic basin hurricane activity. The correlation between SOI and the Atlantic tropical cyclone activity is picked in the new model in the inclusion of compound variables (SOI*SOI) and (ElNino,Del200U). The negative coefficient of -0.123373 of (SOI*SOI) might mean that with high values of SOI, the frequency of cyclones in the Atlantic basin increases. More cyclones may mean that the intensity of each cyclone might be reduced as a consequence.

Another example is, that according to Gray [116], Gulf of Guinea rainfall during the prior autumn season (August to November) is likely to be related to the strength of the West African monsoon (June to July) in the following year through positive feedbacks of evapotranspiration and soil moisture. RainS and RainG combined (known as "early season combination rainfall index") is very good predictor of intense hurricane activity during the period from August to October. When the western Sahel region has above-average rainfall, Atlantic hurricane activity is greatly enhanced. The inclusions of RainG with a coefficient of 0.067429 and (SLPA,RainG) with a coefficient of 0.108001 in the new model is consistent with this finding.

To see the models the search algorithm could come up with if we used all of the available data, we ran more experiments using MML, MDL, CAICF and SRM individually as the cost function. Tables 5.8 and 5.9 show the results of these experiments. The results of Model_7, Model_{18} and SHIFOR94 that have been shown in Table 5.7 are included in Table 5.8 for clarity of comparison. We can see from

Table 5.8, that all of the penalty-penalised methods can improve on the performances of Model_7 and indeed on SHIFOR94 when they are allowed to select a more complex model than a model with 9 variables. The fact that MML has chosen a less complex model with slightly worse performance that the models found by MDL and CAICF is consistent with the results shown in Tables 5.3 and 5.4. These results have been discussed earlier in the second paragraph of this section.

We can see in Table 5.9 that all of the variables chosen for Model_7 are a subset of each of the variable sets of the models chosen by MML, MDL, CAICF. The variables are also a subset of the SRM model except for one variable, (SLPA, RainG). This confirms that Step 3 of Procedure 1, in which we create new models with increasing complexity from all of the models found independently by MML, MDL, CAICF and SRM in the order of most frequently occuring to least frequently occuring variables, is a good way of presenting good models with increasing complexity. The decision to be made by a human decision maker is how complex a model is allowed to be. The point is that using a good optimisation search algorithm, a good complexity-penalised model selection criterion will eventually converge to an optimum model, and will never overfit the training data set. If the optimum model, e.g. Model_{18}, is seen to be too complex, the decision of which less complex model will be chosen should be based on the economic principle, i.e. whether or not the improvement on predictive performance warrants the increase in model complexity. The model selection procedure outlined in Procedure 1 allows a human decision maker to make an informed decision on this matter.

Tables 5.8 and 5.9 also show that Model_{18} with 27 variables which is found as the best model using Procedure 1 on the 10 data subsets has worse performance than the model with 26 variables found by CAICF using all of the available data. It is possible that if we use all of the available data in Step 1 of Procedure 1 and run a

number of experiments using different starting points in the search space instead of using the random subsampling method, we will be able to come up with a better model. As seen in Table 5.8, the four complexity-penalised methods do not overfit the data. Hence, it is definitely possible to forget about test data sets and use these methods to build the forecasting models using all of the available data.

This thesis proposes Model_7 shown in Table 5.7 for two reasons. First, it consists of 9 variables, the same model complexity as SHIFOR and SHIFOR94, the benchmark models. If, however, the model is allowed to be more complex that 9 variables, then the model with 26 variables found by CAICF should be used. Note that this model also has the minimum message length. Second, the prevalence of the seasonal variables in Model_7 has already provided evidence that the seasonal variables indeed influence tropical cyclone intensity in the Atlantic basin. This finding is consistent with the findings published in the atmospheric science literatures which report that the seasonal variables influence tropical cyclone activity in the basin.

5.5 Conclusion

In this chapter we start with having four model selection criteria, i.e. MML, MDL, CAICF and SRM, and a non-backtracking search strategy which have been proven to perform well on artificial data in the experiments reported in Chapter 3. We use these methods in two types of experiments to build models for tropical cyclone intensity change forecasting. The first is finding models by running the search strategy using one of the four model selection method as the cost function. The second is finding new models by ranking the regressors included in the models found by the four model selection methods in the first approach, from the most to the least popular.

We notice that the search strategy will not always converge to a global optimum. For example, the best model according to MML message length (i.e. the model with the shortest message length) was found by using CAICF as the cost function. This is despite the fact the non-backtracking search strategy has been designed to be more exhaustive than a simple greedy search in that it considers every available variable when trying to add a variable to a model or delete a variable from a model. This confirms the known fact that any optimisation search strategy will run the risk of getting trapped in a local minimum.

In all of the experiments done in this chapter, all of the good models selected by all of the four model selection methods have large number of regressors. For example, Model_{18} yielded by the experiments using Procedure 1 and the best model with the shortest MML message length and CAICF cost yielded by the experiments using all of the data have 27 and 26 variables, respectively. This is not unreasonable. The regressors were chosen in the first place because they are thought to be useful by the experts in the atmospheric science community. The large number of regressors in the good models is justifiable for two reasons. First, by only using 3.85% of all of the possible regressors, we can be assured that the models are not overly complex. The models have around 27 variables out of 702 possible regressors in 2^{702} possible models. Second, the good models were more likely to use single/simple variables variables than compound variables, given that there are 36 single variables and 666 compound variables.

In a nutshell, the experiments reported in this chapter highlight the following important points:

1. That the four model selection methods, i.e. the proposed MML method, MDL, CAICF and SRM, together with the proposed non-backtracking search strategy managed to discover good forecasting models in an automated manner. These models are a bit too complex to have been discovered by human inspection.

2. That the proposed procedure of building and ranking new models based on the degrees of popularity of the regressors included in the models found by the search strategy with each of the four methods used as the cost function, can be used to study how each variable contributes to the forecasting ability of a model.

3. That the often unavoidable practice of using a non-exhaustive search strategy on large search space introduces the influence of selection bias in determining to which local optimum a model selection method will converge. There is no guarantee that a non-exhaustive search strategy will converge to a local global optimum all the time. At times, it can get stuck in a local minimum.

4. That it is important to have homogeneous data (i.e. those coming from the same probability distribution) for training and test data sets for a model selection method to pick up regularities in the training data that can be extrapolated into the test data set and beyond.

5. That in contrast to SHIFOR94, there is strong presence of the seasonal predictors in the new model discovered using the proposed procedure. These predictors have been proven in the literature to have strong correlation with to the Atlantic cyclone activity.

5.6 Epilogue

Although it is impossible to explain all of the interactions between the variables stochastically chosen for tropical cyclone intensity change forecasting models, one would hope that, to a certain extent, the reasons why some variables were chosen and some were not could be explained. The total absence of the seasonal predictors, which have been proven to be influential to Atlantic tropical cyclone systems, from the tropical cyclone intensity forecasting models being used in operation begs a closer look into the way the models have been built. New models built using scientifically better methods like the ones proposed in this chapter should be tested in operational use over a period of time.

Table 5.3: Models selected for Atlantic hurricane intensity change forecasting. See the caption of Table 5.4 for further explanations

Data set	Method	Tot Reg	Message Length	training data		test data	
				RMSE	R^2	RMSE	R^2
1	MML	14	3568.3599	21.5337	0.4451	21.9360	0.4578
	MDL	22	3571.9245	21.0042	0.4734	21.7157	0.4719
	CAICF	14	3568.3599	21.5337	0.4451	21.9360	0.4578
	SRM	14	3568.3599	21.5337	0.4451	21.9360	0.4578
	SHIFOR'	9	4146.1623	26.3637	0.1668	27.0951	0.1695
	SHIFOR94'	9	3675.8142	22.9202	0.3703	23.1839	0.3920
	SHIFOR	9	n/a	24.64	n/a	25.08	n/a
	SHIFOR94	9	n/a	22.50	n/a	22.97	n/a
2	MML	17	3470.3483	21.1317	0.4874	22.0990	0.3960
	MDL	17	3494.2895	21.1978	0.4842	22.2196	0.3894
	CAICF	17	3494.2895	21.1978	0.4842	22.2196	0.3894
	SRM	17	3494.2895	21.1978	0.4842	22.2196	0.3894
	SHIFOR'	9	4121.8865	26.6905	0.1800	26.3002	0.1392
	SHIFOR94'	9	3609.8455	22.8881	0.3970	23.2677	0.3263
	SHIFOR	9	n/a	25.15	n/a	25.24	n/a
	SHIFOR94	9	n/a	22.58	n/a	22.78	n/a
3	MML	16	3501.8237	21.3013	0.4741	21.7768	0.4272
	MDL	22	3525.6188	21.0137	0.4892	21.9255	0.4220
	CAICF	22	3526.8604	21.0067	0.4896	21.9845	0.4189
	SRM	14	3541.3063	21.6404	0.4569	22.1669	0.4055
	SHIFOR'	9	4113.0345	26.4903	0.1843	26.7255	0.1326
	SHIFOR94'	9	3631.5392	22.9458	0.3883	23.1232	0.3506
	SHIFOR	9	n/a	24.78	n/a	25.17	n/a
	SHIFOR94	9	n/a	22.59	n/a	22.76	n/a
4	MML	15	3544.6713	21.5543	0.4585	21.3028	0.4581
	MDL	22	3527.0303	20.9527	0.4895	21.4421	0.4539
	CAICF	22	3527.0303	20.9527	0.4895	21.4421	0.4539
	SRM	34	3529.3299	20.2208	0.5264	21.2702	0.4677
	SHIFOR'	9	4160.5295	26.8368	0.1589	25.9770	0.1904
	SHIFOR94'	9	3670.4534	23.1806	0.3725	22.5890	0.3878
	SHIFOR	9	n/a	24.95	n/a	25.39	n/a
	SHIFOR94	9	n/a	22.85	n/a	22.13	n/a
5	MML	16	3517.5771	21.0877	0.4698	22.4091	0.4321
	MDL	22	3526.3994	20.7077	0.4897	22.3168	0.4394
	CAICF	21	3515.7079	20.7330	0.4883	22.2483	0.4424
	SRM	29	3534.6611	20.3360	0.5090	21.6955	0.4731
	SHIFOR'	9	4137.0499	26.3244	0.1718	27.1136	0.1642
	SHIFOR94'	9	3649.0758	22.7536	0.3812	23.5696	0.3684
	SHIFOR	9	n/a	24.41	n/a	24.88	n/a
	SHIFOR94	9	n/a	22.39	n/a	23.20	n/a
6	MML	19	3541.8138	21.1563	0.4723	21.9039	0.4460
	MDL	22	3551.1410	21.0008	0.4806	21.9859	0.4431
	CAICF	22	3551.1410	21.0008	0.4806	21.9859	0.4431
	SRM	28	3548.9379	20.6145	0.5005	21.4804	0.4709
	SHIFOR'	9	4144.5904	26.5266	0.1677	26.6490	0.1736
	SHIFOR94'	9	3658.2100	22.9404	0.3775	23.2089	0.3732
	SHIFOR	9	n/a	25.60	n/a	25.10	n/a
	SHIFOR94	9	n/a	22.68	n/a	22.53	n/a

Table 5.4: ... continued from Table 5.3. SHIFOR and SHIFOR94 are the original benchmark models. SHIFOR' and SHIFOR94' are models with the same variables as those of the original models but with coefficients recalculated to fit the training data of each data set. The last set has training data from the year 1950 to 1987 and test data from the year 1988 to 1994.

Data set	Method	Tot Reg	Message Length	training data		test data	
				RMSE	R^2	RMSE	R^2
7	MML	16	3550.5365	21.4161	0.4570	21.6605	0.4609
	MDL	20	3555.6959	21.1497	0.4711	21.4657	0.4722
	CAICF	20	3555.6959	21.1497	0.4711	21.4657	0.4722
	SRM	20	3555.6959	21.1497	0.4711	21.4657	0.4722
	SHIFOR'	9	4129.8202	26.3546	0.1757	27.0755	0.1531
	SHIFOR94'	9	3664.3147	22.9488	0.3750	23.1047	0.3833
	SHIFOR	9	n/a	25.10	n/a	25.01	n/a
	SHIFOR94	9	n/a	22.64	n/a	22.64	n/a
8	MML	16	3547.1234	21.4495	0.4593	21.3224	0.4683
	MDL	18	3557.4183	21.3424	0.4651	21.3072	0.4699
	CAICF	18	3557.4183	21.3424	0.4651	21.3072	0.4699
	SRM	18	3557.4183	21.3424	0.4651	21.3072	0.4699
	SHIFOR'	9	4150.8674	26.6377	0.1642	26.3814	0.1816
	SHIFOR94'	9	3670.8837	23.0851	0.3723	22.8126	0.3880
	SHIFOR	9	n/a	25.31	n/a	24.99	n/a
	SHIFOR94	9	n/a	22.73	n/a	22.42	n/a
9	MML	17	3516.6721	21.3600	0.4738	21.3381	0.4442
	MDL	22	3522.0461	21.0540	0.4896	21.0350	0.4620
	CAICF	17	3513.9074	21.3405	0.4747	21.2465	0.4489
	SRM	27	3525.8268	20.7391	0.5056	20.8957	0.4711
	SHIFOR'	9	4122.9112	26.6371	0.1795	26.4400	0.1413
	SHIFOR94'	9	3637.2253	23.0412	0.3860	22.9308	0.3541
	SHIFOR	9	n/a	25.37	n/a	25.17	n/a
	SHIFOR94	9	n/a	22.42	n/a	22.80	n/a
10	MML	14	3545.9281	21.7222	0.4507	21.7164	0.4333
	MDL	18	3542.4352	21.4043	0.4674	21.3084	0.4561
	CAICF	18	3542.4352	21.4043	0.4674	21.3084	0.4561
	SRM	18	3545.8094	21.4065	0.4673	21.3807	0.4524
	SHIFOR'	9	4155.6197	26.8147	0.1616	26.0096	0.1839
	SHIFOR94'	9	3645.6768	23.0106	0.3826	22.9506	0.3646
	SHIFOR	9	n/a	25.97	n/a	25.24	n/a
	SHIFOR94	9	n/a	22.73	n/a	22.42	n/a
years:	MML	23	4131.9872	20.1489	0.5162	26.6464	0.2345
1950-87	MDL	26	4127.9430	19.9376	0.5267	27.7543	0.1735
1988-94	CAICF	26	4127.9430	19.9376	0.5267	27.7543	0.1735
	SRM	26	4127.9430	19.9376	0.5267	27.7543	0.1735
	SHIFOR'	9	5036.0730	26.6455	0.1507	26.3322	0.2361
	SHIFOR94'	9	4432.1653	22.8928	0.3731	23.7897	0.3765
	SHIFOR	9	n/a	25.18	n/a	24.64	n/a
	SHIFOR94	9	n/a	22.44	n/a	23.74	n/a

Table 5.5: Models as collections of variables with the same minimum frequency of being chosen to form a model by MML, MDL, CAIF, or SRM for the 10 data sets in Table 5.3 and 5.4

Model	Size	Min. Freq.	Commonly chosen regressors in models (the + sign means plus the variables in the rows above)
1	2	40	6 10
2	3	38	+ (36,36)
3	4	36	+ (28,5)
4	5	33	+ (33,32)
5	7	30	+ (6,5) 29
6	8	28	+ 32
7	9	25	+ (35,29)
8	10	24	+ 31
9	11	23	+ (3,2)
10	12	19	+ (32,11)
11	13	18	+ (34,11)
12	15	17	+ 7 (32,15)
13	16	16	+ (9,3)
14	17	15	+ 2
15	18	12	+ 4
16	19	11	+ (30,22)
17	25	10	+ 3 9 (9,4) 13 25 (29,29)
18	27	9	+ 15 (32,31)
19	30	8	+ (4,4) (8,6) 35
20	35	7	+ 5 (5,5) (9,1) (22,9) (35,33)
21	42	6	+ (4,1) (11,5) 30 (34,17) (34,21) (34,28) (35,28)
22	43	5	+ 18
23	50	4	+ (7,2) (13,2) (13,13) (27,14) (31,3) (32,21) (36,35)
SHIFOR	9	n/a	7 (3,1) (5,1) (6,1) (4,3) (5,3) (7,5) (5,5) (6,5)
SHIFOR94	9	n/a	10 11 5 16 (16/4) 25 (10,10) (4,5) (6,3)

Table 5.6: Performance of each model in Table 5.5 calculated using all of the available data (following Step 3 of Procedure 1)

Model	Size	MML	MDL	CAICF	SRM	RMSE	R^2
1	2	5227.1275	5215.0073	5225.0946	0.6944	23.2945	0.3588
2	3	5213.8092	5195.3430	5207.7612	0.6937	23.1557	0.3666
3	4	5202.4689	5178.5372	5193.1246	0.6929	23.0321	0.3735
4	5	5183.3942	5154.5672	5171.1571	0.6891	22.8707	0.3823
5	7	5144.7799	5109.8437	5130.0198	0.6810	22.5659	0.3990
6	8	5137.7611	5101.6132	5123.4250	0.6808	22.4872	0.4033
7	9	5128.5315	5088.4975	5111.8294	0.6788	22.3830	0.4089
8	10	5116.4937	5075.6970	5100.4513	0.6767	22.2806	0.4145
9	11	5095.5560	5051.2319	5077.2850	0.6707	22.1187	0.4231
10	12	5069.3557	5021.7370	5048.9831	0.6630	21.9322	0.4329
11	13	5060.4083	5009.7104	5038.1178	0.6606	21.8345	0.4381
12	15	5070.1827	5016.4281	5047.0496	0.6646	21.7933	0.4405
13	16	5063.9349	5007.5211	5039.1089	0.6627	21.7109	0.4448
14	17	5058.9406	5002.6561	5035.1638	0.6619	21.6487	0.4481
15	18	5040.9346	4984.9951	5018.3082	0.6571	21.5229	0.4546
16	19	5045.0687	4986.7331	5020.8790	0.6581	21.4933	0.4563
17	25	5001.3226	4942.4237	4980.2451	0.6460	21.0455	0.4794
18	27	4996.8836	4937.5439	4976.2822	0.6444	20.9444	0.4846
19	30	5000.6515	4939.6237	4979.5095	0.6444	20.8375	0.4902
20	35	5031.1451	4966.6092	5007.9117	0.6505	20.7698	0.4941
21	42	5070.0102	5002.0126	5044.0818	0.6569	20.6588	0.5003
22	43	5069.8073	5004.1261	5046.1716	0.6567	20.6285	0.5019
23	50	5111.5253	5042.4039	5083.8683	0.6625	20.5259	0.5077
SHIFOR'	9	5871.9392	5825.8890	5850.9167	0.9529	26.5225	0.1701
SHIFOR94'	9	5179.8152	5178.5030	5202.0420	0.7075	22.8515	0.3840

Table 5.7: $Model_7$ (consisting the best 9 variables), SHIFOR94 and SHIFOR: variable names and their respective coefficients

Chosen Variable		Normalized Coefficient		
No.	Acronym	$Model_7$	SHIFOR94'	SHIFOR'
(6,5)	(Del12V,Vmax)	-0.099445	-0.085522	0.900026
10	POT	0.649806	0.644563	
6	Del12V	0.165545	0.180562	
(36,36)	(SOI,SOI)	-0.123373		
(28,5)	(200Uend,Vmax)	-0.119977		
29	Del200U	0.109351		
(33,32)	(SLPA,RainG)	0.108001		
(35,29)	(ElNino,Del200U)	0.075820		
32	RainG	0.067429		
12	DSSST		-0.105468	
(10,10)	(POT,POT)		-0.101357	
26	Closest		0.074916	
17	Uppend		0.023286	
(4,7)	(LonData,UCurr)		0.013477	
(17/6)	(Uppend/Vmax)		-0.006972	
(5,5)	(Vmax,Vmax)			-0.314015
7	UCurr			-0.294385
(5,3)	(Vmax,LatData)			-0.069982
(3,1)	(LatData,Julian)			-0.069132
(4,3)	(LonData,LatData)			0.069030
(7,5)	(UCurr,Vmax)			-0.054350
(5,1)	(Vmax,Julian)			-0.017290
(6,1)	(Del12V,Julian)			0.012620

Table 5.8: Performances of $Model_7$, SHIFOR94' and $Model_{18}$ (from Table 5.6) compared with those of MML, MDL, CAICF and SRM when the search algorithm is run on all of the available data. Columns 3 to 6 show the costs of each model based on the calculations of the four model selection methods. Columns 7 and 8 show RMSE and R^2.

Model	Size	MML	MDL	CAICF	SRM	RMSE	R^2
$Model_7$	9	5128.5315	5088.4975	5111.8294	0.6788	22.3830	0.4089
SHIFOR94'	9	5179.8152	5178.5030	5202.0420	0.7075	22.8515	0.3840
$Model_{18}$	27	4996.8836	4937.5439	4976.2822	0.6444	20.9444	0.4846
MML	18	4978.9114	4938.7904	4971.0337	0.6427	21.2326	0.4692
MDL	27	4969.3988	4896.1398	4934.6116	0.6323	20.7458	0.4944
CAICF	26	4968.0812	4896.3251	4934.3544	0.6324	20.7851	0.4923
SRM	18	4980.5410	4951.9334	4985.1020	0.6391	21.2257	0.4696

Table 5.9: $Model_7$ and $Model_{18}$ (from Table 5.7) and the models yielded by the search on all of the available data shown in Table 5.8: variable names and their respective coefficients. The order of the variables from the top to the bottom of the table follows the order in which each variable appears in the models shown in Table 5.5

Chosen Variable		Normalized Coefficient					
No.	Acronym	$Model_7$	$Model_{18}$	MML	MDL	CAICF	SRM
6	Del12V	0.165545	0.167516	0.170748	0.154972	0.155474	0.170557
10	POT	0.649806	0.762634	0.749576	0.739638	0.737958	0.760658
(36,36)	(SOI,SOI)	-0.123373	-0.116110	-0.121128	-0.284544	-0.285706	-0.135621
(28,5)	(200Uend,Vmax)	-0.119977	-0.151762	-0.157228	-0.150183	-0.150318	-0.164793
(33,32)	(SLPA,RainG)	0.108001	0.095510	0.079034	0.074778	0.076354	
(6,5)	(Del12V,Vmax)	-0.099445	-0.097937	-0.099032	-0.086772	-0.092528	-0.099625
29	Del200U	0.109351	0.112490	0.116458	0.090777	0.090835	0.103970
32	RainG	0.067429	0.093559	0.082605	0.086814	0.087316	0.092783
(35,29)	(ElNino,Del200U)	0.075820	0.065408	0.064796	0.140820	0.145080	0.060216
31	RainS		0.056894	0.077050	0.106116	0.105726	0.079129
(3, 2)	(LatData, JulOff)		0.081782	0.074244	0.075362	0.076064	0.075711
(32,11)	(RainG, POTend)		0.072127	0.130398	0.140085	0.138693	0.128291
(34,11)	(ZWA, POTend)		0.088700	0.102347	0.118228	0.117873	0.104532
7	UCurr		-0.061550				
(32,15)	(RainG, DSSST)		0.047407		0.068531	0.067274	
(9, 3)	(Speed, LatData)		-0.033333	-0.071669	-0.066931	-0.065315	-0.062871
2	JulOff		-0.046875	-0.059054			-0.052871
4	LonData		-0.094644				
(30,22)	(U50, 200mbT)		0.068787		0.063724	0.062727	
3	LatData		0.156105				
9	Speed		0.045668				
(9,4)	(Speed, LonData)		-0.060584				
13	SSST		0.003643	-0.156523	-0.163369	-0.161464	-0.164559
25	DisLand		-0.049317	0.090532	0.116527	0.117286	0.106286
(29,29)	(Del200U, Del200U)		-0.077878				
15	DSSST		-0.138307				
(32,31)	(RainG, RainS)		-0.049317				
(8, 6)	(VCurr, Del12V)				-0.045868		
35	ElNino			-0.052226			
(35,33)	(ElNino, SLPA)				0.091674	0.091991	
(35,28)	(ElNino, 200Uend)				-0.123700	-0.128778	
(32,16)	(RainG, UppSpd)				0.076416	0.075250	
(33,22)	(SLPA, 200mbT)				0.063399	0.063001	0.092112
(35, 2)	(ElNino, JulOff)				0.071662	0.070609	
33	SLPA				0.073865	0.074714	
(36,35)	(SOI, ElNino)				-0.257329	-0.257280	
(26,26)	(Closest, Closest)				-0.061172	-0.063483	-0.058905

Chapter 6

Sampling of Highly Correlated Data for Polynomial Regression and Model Discovery

6.1 Introduction

The experiments done in Chapter 3 tested the abilities of various model selection criteria to recover true second-order polynomial models from artificial data. The experiments done in Chapter 5 used the model selection criteria that had passed the tests done in Chapter 3 to build a tropical cyclone intensity change forecasting model from a set of real climatological and environment data gathered from the North Atlantic tropical cyclone basin. The experiments in this chapter[1] test the model selection criteria used to build the forecasting model in Chapter 5 in their abilities to recover the forecasting model from artificial data sets of a set of variables whose

[1] An earlier version of this chapter has been published in [44]

covariance matrix is the same as the covariance matrix of the real climatological and environment data from which the model has been inferred. The aim is to see if the model selection criteria will still infer the same model as the one proposed in Chapter 5 in the face of varying sample sizes and levels of noise in the target variable.

This chapter derives a formula to generate the artificial data sets of a pool of variables from which the model is to be inferred. The artificial data sets are generated using a combination of random data generated from the standard normal distribution and the covariance matrix of the real observation data.

6.2 The Covariance Matrix and the True Model

The source of the set of observation data whose covariance matrix is used in this chapter is the 36 climatological and environmental variables used to build the tropical cyclone forecasting model in Chapter 5. The summary of the variables can be found in Table 5.2 on page 105.

The tropical cyclone intensity forecasting model proposed in Chapter 5 consists of 9 regressors. The model and the explanations of the regressors are respectively shown in Tables 6.1 and 6.2. These tables are extracted from Tables 5.7 on page 121 and 5.2 on page 105. As outlined in Section 5.3.1, the regressors are chosen from the 36 meteorological and environmental variables and their products. The search space thus consists of 36 single and 666 products of variables. As mentioned in Section 5.3.2, the sample size of the observation data is 4347 data points.

The forecasting model combined with a degree of independent random noise $r \sim N(0, \sigma)$ is used to calculate the data sets of the target variable from the artificial

Table 6.1: The Atlantic tropical cyclone intensity change forecasting model with 9 regressors found in the experiments done in Chapter 5. Compound variable (Variable1, Variable2) represents a product of two variables. The meanings of the variables are given in Table 6.2

No	Chosen Regressor	Coefficient
1	Del12V	0.550029
2	POT	0.551974
3	Del200U	0.473030
4	RainG	2.757764
5	(200Uend,Vmax)	-3.478769
6	(Del12V,Vmax)	-2.899739
7	(ElNino,Del200U)	2.314881
8	(SLPA,RainG)	2.829764
9	(SOI,SOI)	-2.622052
	constant	-30.787608

data of the corresponding regressors generated in this chapter. The task of each model selection criterion tested in this chapter is to find a model that can predict this target variable well. It is then to be observed if the model found consists of the same set of regressors as the 'true' model, the proposed forecasting model.

6.3 The Model Selection Criteria Tested

The model selection criteria which were involved in the formation of the forecasting model shown in Table 6.1 as described in the integrated model discovery procedure proposed in Section 1, are individually tested in the experiments in this chapter. These criteria are Minimum Message Length Principle (MML, see Section

Table 6.2: The basic regressors of the Atlantic tropical cyclone intensity change forecasting model shown in Table 6.1. The target variable is the change of intensity (wind speed) 72 hours into the future

Regressor	Explanation
Vmax	the initial cyclone intensity (in knots)
Del12V	the change of cyclone intensity in the past 12 hours (in knots)
POT	cyclone initial potential intensity
Del200U	the forecast of the change of eastward wind motion at 200mb
RainG	African Gulf of Guinea rainfall index (0W-10W, 5N-10N)
200Uend	the forecast of eastward wind motion at 200 mb at the end of 72 hours
SOI	Surface pressure gradient between Darwin and Tahiti
SLPA	April-May Caribbean basin Sea Surface Pressure Anomaly
ElNino	Sea surface temperature anomaly in the eastern equatorial Pacific

2.3.3.1), Minimum Description Length Principle (MDL, see Section 2.3.3.3), Corrected Akaike's Information Criterion (CAICF, see Section 2.3.3.5) and Structured Risk Minimization (SRM, see Section 2.3.3.7).

The existence of high degree of correlations among regressors and the introduction of varying degrees of noise in the values of the target variable have been known to hinder the discovery of true model. In the experiments in Chapter 3, it has been shown that the orthogonal transformation of regressors method outlined in Section 3.3.1 manages to find a true model from a search space with highly correlated regressors. This method will also be used in the experiments in this chapter.

6.4 Methodology

This section derives the formula to be used to generate artificial data of a set of variables whose covariance matrix is the same as the covariance matrix of a set of

observation data. This is done by using the combination of the covariance matrix of the set of observation data and random data generated from the standard normal distribution.

This section then outlines a model discovery procedure using a set of artificial data of the variables that make up the covariance matrix, a data set of the target variable calculated from a known true model and a model selection criterion.

6.4.1 Generating Regressor Data from Covariance Matrix

A. Problem Definition

Suppose we are given a set of observation data of K variables,

$$\mathbf{x}_1, \ldots, \mathbf{x}_K = \underset{N \times K}{\mathbf{X}} \tag{6.1}$$

where $\mathbf{x}_k$ is a vector of N data points for the observation variable x_k,

and a set of random independent data generated from the standard normal distribution $N(0,1)$

$$\mathbf{u}_1, \ldots, \mathbf{u}_K = \underset{N \times K}{\mathbf{U}} \tag{6.2}$$

where $\mathbf{u}_k$ is a vector of N data points for the variable u_k; $u_{n,k} \sim N(0,1)$.

We would like to find matrix $\underset{K \times K}{\mathbf{B}}$ that can transform $\underset{N \times K}{\mathbf{U}}$ into a new set of data

$$\mathbf{z}_1, \ldots, \mathbf{z}_K = \underset{N \times K}{\mathbf{Z}} \tag{6.3}$$

where $\mathbf{z}_k$ is a vector of N data points for the transformed variable z_k whose covariance matrix, $E(\underset{N\times K}{\mathbf{Z}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{Z}})$, equals the covariance matrix of the observation data $\underset{N\times K}{\mathbf{X}}$

$$E(\underset{N\times K}{\mathbf{Z}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{Z}}) = E(\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}}) \tag{6.4}$$

The covariance matrix of the observation matrix is by definition:

$$\begin{aligned} E(\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}}) = \frac{1}{N} \times & \begin{bmatrix} (x_{1,1}-\mu_1) & (x_{1,2}-\mu_2) & \dots & (x_{1,K}-\mu_K) \\ (x_{2,1}-\mu_1) & (x_{2,2}-\mu_2) & \dots & (x_{2,K}-\mu_K) \\ \dots & \dots & \dots & \dots \\ (x_{N,1}-\mu_1) & (x_{N,2}-\mu_2) & \dots & (x_{N,K}-\mu_K) \end{bmatrix}^{\mathrm{T}} \\ \times & \begin{bmatrix} (x_{1,1}-\mu_1) & (x_{1,2}-\mu_2) & \dots & (x_{1,K}-\mu_K) \\ (x_{2,1}-\mu_1) & (x_{2,2}-\mu_2) & \dots & (x_{2,K}-\mu_K) \\ \dots & \dots & \dots & \dots \\ (x_{N,1}-\mu_1) & (x_{N,2}-\mu_2) & \dots & (x_{N,K}-\mu_K) \end{bmatrix} \end{aligned} \tag{6.5}$$

where $\mu_1, \dots, \mu_K$ are the means of variables $x_1, \dots, x_K$.

B. Solution: Finding the Transformation Matrix $\underset{K\times K}{\mathbf{B}}$

Data of variables $\mathbf{u}_i$ and $\mathbf{u}_j$ are independent and identically distributed as defined in Equation 6.2. Hence, the following equations hold

$$E(\mathbf{u}_i^T \mathbf{u}_j) = 0 \tag{6.6}$$

$$E(\mathbf{u}_i^T \mathbf{u}_i) = 1 \tag{6.7}$$

$$\Rightarrow E(\underset{N\times K}{\mathbf{U}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{U}}) = \underset{K\times K}{\mathbf{I}} \tag{6.8}$$

where $\underset{K\times K}{\mathbf{I}}$: the identity matrix K × K

We then find a constant transformation matrix $\underset{K\times K}{\mathbf{B}}$ to transforms the matrix of the original unit normal data $\underset{N\times K}{\mathbf{U}}$ to a new data matrix $\underset{N\times K}{\mathbf{Z}}$

$$\underset{N\times K}{\mathbf{Z}} = \underset{N\times K}{\mathbf{U}}\ \underset{K\times K}{\mathbf{B}} \tag{6.9}$$

The covariance matrix of the new set of data is thus

$$E(\underset{N\times K}{\mathbf{Z}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{Z}}) = E\left((\underset{N\times K}{\mathbf{U}}\ \underset{K\times K}{\mathbf{B}})^{\mathrm{T}}(\underset{N\times K}{\mathbf{U}}\ \underset{K\times K}{\mathbf{B}})\right) \tag{6.10}$$

and because $\underset{K\times K}{\mathbf{B}}$ is constant

$$\Leftrightarrow E(\underset{N\times K}{\mathbf{Z}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{Z}}) = \underset{K\times K}{\mathbf{B}}^{\mathrm{T}}\ E(\underset{N\times K}{\mathbf{U}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{U}})\ \underset{K\times K}{\mathbf{B}} \tag{6.11}$$

By Equation 6.8

$$\Leftrightarrow E(\underset{N\times K}{\mathbf{Z}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{Z}}) = \underset{K\times K}{\mathbf{B}}^{\mathrm{T}}\ \underset{K\times K}{\mathbf{I}}\ \underset{K\times K}{\mathbf{B}} \tag{6.12}$$

$$\Leftrightarrow E(\underset{N\times K}{\mathbf{Z}}^{\mathrm{T}}\ \underset{N\times K}{\dot{\mathbf{Z}}}) = \underset{K\times K}{\mathbf{B}}^{\mathrm{T}}\ \underset{K\times K}{\mathbf{B}} \tag{6.13}$$

This means that by choosing $\underset{K\times K}{\mathbf{B}}$ correctly, we can find the new set of data defined in Equation 6.3 that makes Equation 6.4 hold.

To do this, we define the covariance matrix of the observation data $E(\mathbf{X}^T\mathbf{X})$ in terms of its eigenvalues and eigenvectors

$$E(\underset{N\times K}{\mathbf{X}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{X}}) = \underset{K\times K}{\mathbf{Q}}^{\mathrm{T}}\ \underset{K\times K}{\mathbf{\Lambda}}\ \underset{K\times K}{\mathbf{Q}} \tag{6.14}$$

$$\Leftrightarrow E(\underset{N\times K}{\mathbf{X}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{X}}) = \underset{K\times K}{\mathbf{Q}}^{\mathrm{T}}\ \sqrt{\underset{K\times K}{\mathbf{\Lambda}}}\ \sqrt{\underset{K\times K}{\mathbf{\Lambda}}}\ \underset{K\times K}{\mathbf{Q}} \tag{6.15}$$

where

$\underset{K\times K}{\mathbf{\Lambda}}$: a diagonal matrix with the eigenvalues of the covariance matrix $E(\underset{N\times K}{\mathbf{X}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{X}})$ as its diagonal elements

$\underset{K\times K}{\mathbf{Q}}$: a K × K matrix with the eigenvectors of the covariance matrix $E(\underset{N\times K}{\mathbf{X}}^{\mathrm{T}}\ \underset{N\times K}{\mathbf{X}})$ as its columns

By Equations 6.4, 6.13 and 6.15, we get

$$\underset{K\times K}{\mathbf{B}} = \sqrt{\underset{K\times K}{\mathbf{\Lambda}}}\ \underset{K\times K}{\mathbf{Q}} \tag{6.16}$$

Finally by Equations 6.16 and 6.9, we get

$$\underset{N\times K}{\mathbf{Z}} = \underset{N\times K}{\mathbf{U}}\ \sqrt{\underset{K\times K}{\mathbf{\Lambda}}}\ \underset{K\times K}{\mathbf{Q}} \tag{6.17}$$

That is, a data point for a transformed variable z_k can be calculated using the following formula[2].

$$z_{n,k} = \sum_{m=1}^{K} u_{n,m}\sqrt{\lambda_m}\ q_{m,k} \tag{6.18}$$

[2]The unit normal random data sets $u_1,\ldots,u_K$ used in this thesis are generated using the random number generator program `FastNorm2.c` written by Chris Wallace [17]. The eigenvalues and eigenvectors of the covariance matrix of the observation data are calculated using the Jacobian diagonalization of real symmetric matrix program `jacob.c` also written by Chris Wallace.

where

$z_{n,k}$: n^{th} data point of the transformed variable z_k

$u_{n,m}$: n^{th} data point of the random independent variable u_m generated from the standard normal distribution, $u_{nm} \sim N(0,1)$

λ_m : the m^{th} diagonal element of $\underset{K\times K}{\mathbf{\Lambda}}$, the diagonal eigenvalue matrix of the covariance matrix $E(\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}})$

$q_{m,k}$: the m^{th} element of column k of $\underset{K\times K}{\mathbf{Q}}$ the eigenvector matrix of the covariance matrix $E(\underset{N\times K}{\mathbf{X}^{\mathrm{T}}}\ \underset{N\times K}{\mathbf{X}})$

6.4.2 Model Discovery Process

Procedure 2 is the procedure to generate artificial data for a pool of potential regressors, to calculate their associated target variable from a known true model and to use a search engine to recover the true model from the pool of potential regressors.

6.5 Experiments and Results

We use the model given in Table 6.1 as the true model that we would like to recover from the newly generated data. We generate five categories of data sets with increasing levels of noise $r \sim N(0,\sigma^2)$, denoted by `DataSet1, ..., DataSet5`. For each category, we generate train-test data sets each comprising 100, 500, 1000, 2000, 4000, 6000 and 10000 data points. We then follow the procedure outlined in Section 6.4.2 to compare the performance of each model selection criterion named in Section 6.3 for the task of recovering the true model from the data sets.

Procedure 2 Recover model from artificial data of regressors whose covariance matrix equals that of the observation data

Step 1 Calculate the covariance matrix of the set of multivariate observation data.

Step 2 Generate artificial data for the pool of potential regressors using the formula given in Equation 6.18.

Step 3 Calculate the values for the target variable of the true model (e.g. Table 6.1. This is done by multiplying the value of each regressor with its link weight/coefficient connected to the target variable plus an independent random noise value $r \sim N(0, \sigma^2)$. The values of the target variable will be used to compare the performance of the models found.

Step 4 For each model selection criterion named in Section 6.3, run a search mechanism, like the optimisation search algorithm given in Section 3.2.2 on page 57, to recover the true model from the data generated in Step 2. To eliminate the effect of correlations among regressors in the calculation of the coefficients of the recovered model, use the orthogonal transformation outlined in Section 3.3.1.

Step 5 Compare the models found by each criterion using the performance criteria given in Section 3.3.2 on page 69.

Tables 6.3 and 6.4 show the result of the experiments. It is shown that the results of all of the model selection criteria are quite uniform in that they all managed to recover the true model with similar degrees of accuracy. Also we can observe that all of the criteria managed to recover the true model of 9 regressors with a data set as small as 100 data points regardless of the levels of noise in the target variable. Increasing the size of the data set as the level of noise is increased does not significantly increase the accuracy of the results. This finding indicates that most of the potential regressors do not have direct influence on the target variable.

This finding is valuable in that it confirms that the integrated model discovery procedure outline in Chapter 5 indeed has the ability to select a parsimonious subset

of variables from a pool of potential regressors in the absence of prior knowledge of the problem domain.

6.6 Conclusion

We start with:

1. A set of observation data (e.g. a set data of tropical cyclone, climatological and environmental factors outlined in Section 5.3.1)

2. A set of model selection criteria that have been previously (e.g. in Chapter 3) empirically proven to be robust for the task of polynomial model discovery

3. A model that has previously (e.g. in Chapter 5) been empirically found to be the right model inferred from the data using an integrated approach involving all of the model selection criteria mentioned above

4. A non-backtracking search engine (outlined in Section 3.2.2)

5. An orthogonal matrix transformation method (outlined in Section 3.3.1)

We would like to be able to replicate the observation data so as to be able to conduct multiple experiments to further test the robustness of the integrated model discovery procedure outlined in Section 1 on page 102 and the ability of the individual model selection criterion involved in the integrated approach to recover the model that is seen to be the right model in the face of varying sample sizes and levels of noise.

In this chapter, we outline a procedure to generate artificial data of a set variables from the covariance matrix of a set of observation data, to calculate the values of the

target variable based on a true model and to recover the model from artificial data. The findings from the experiments confirm that the model found by the integrated model discovery procedure is indeed likely to be a good enough model for the target variable inferred from the observation data since it is recovered by all of the model selection criteria individually.

The fact that the criteria only require a relatively small size of data set to recover the true model, even in the presence of high level of noise in the target variable, suggests that the non-backtracking search engine used for the experiments covers a search space extensive enough to converge to an optimum model. It also suggests that the integrated model discovery procedure indeed has the ability to select a parsimonious subset of variables from a pool of potential regressors in the absence of prior knowledge of the problem domain.

Table 6.3: Model discovered in data sets `DataSet1` (noise $\sigma = 2$)

Sample Size	Method	DataSet1			
		nvar	ModelErr	RMSE	R^2
100	MML MDL CAICF SRM	9	0.0023	1.1968	0.9966
500	MML MDL CAICF SRM	9	0.0009	1.0173	0.9972
1000	MML MDL CAICF SRM	9	0.0006	1.0076	0.9973
2000	MML MDL CAICF SRM	9	0.0004	1.0035	0.9973
4000	MML MDL CAICF SRM	9	0.0005	1.0106	0.9972
6000	MML MDL CAICF SRM	9	0.0003	1.0024	0.9973
10000	MML MDL CAICF SRM	9	0.0001	1.0119	0.9973

Table 6.4: Model discovered in data sets `DataSet2` (noise $\sigma = 3$)

Sample Size	Method	DataSet2			
		nvar	ModelErr	RMSE	R^2
100	MML, MDL, CAICF, SRM	9	0.0100	1.2717	0.9963
500	MML, MDL, CAICF, SRM	9	0.0008	1.0014	0.9970
1000	MML, MDL, CAICF, SRM	9	0.0006	1.0222	0.9973
2000	MML, MDL, CAICF, SRM	9	0.0001	1.0047	0.9974
4000	MML, MDL, CAICF, SRM	9	0.0002	1.0130	0.9973
6000	MML, MDL, CAICF, SRM	9	0.0001	0.9982	0.9974
10000	MML, MDL, CAICF, SRM	9	0.0001	0.9944	0.9974

Table 6.5: Model discovered in data sets `DataSet3` (noise $\sigma = 5$)

Sample Size	Method	DataSet3			
		nvar	ModelErr	RMSE	R^2
100	MML MDL CAICF SRM	9	0.0075	1.1785	0.9959
500	MML MDL CAICF SRM	9	0.0010	1.0050	0.9972
1000	MML MDL CAICF SRM	9	0.0013	0.9705	0.9976
2000	MML MDL CAICF SRM	9	0.0005	0.9861	0.9974
4000	MML MDL CAICF SRM	9	0.0001	1.0116	0.9973
6000	MML MDL CAICF SRM	9	0.0001	1.0069	0.9973
10000	MML MDL CAICF SRM	9	0.0000	1.0163	0.9973

Table 6.6: Model discovered in data sets `DataSet4` (noise $\sigma = 10$)

Sample Size	Method	DataSet4			
		nvar	ModelErr	RMSE	R^2
100	MML MDL CAICF SRM	9	0.0049	1.0821	0.9968
500	MML MDL CAICF SRM	9	0.0017	1.0816	0.9972
1000	MML MDL CAICF SRM	9	0.0015	1.0576	0.9969
2000	MML MDL CAICF SRM	9	0.0004	0.9926	0.9974
4000	MML MDL CAICF SRM	9	0.0001	1.0013	0.9974
6000	MML MDL CAICF SRM	9	0.0002	1.0075	0.9973
10000	MML MDL CAICF SRM	9	0.0001	1.0176	0.9973

Table 6.7: Model discovered in data sets `DataSet5` (noise $\sigma = 25$)

Sample Size	Method	DataSet5			
		nvar	ModelErr	RMSE	R^2
100	MML MDL CAICF SRM	9	2.5592	1.4493	0.9950
500	MML MDL CAICF SRM	9	0.0028	1.0112	0.9975
1000	MML MDL CAICF SRM	9	0.0022	1.0383	0.9972
2000	MML MDL CAICF SRM	9	0.0007	1.0095	0.9973
4000	MML MDL CAICF SRM	9	0.0001	1.0102	0.9973
6000	MML MDL CAICF SRM	9	0.0001	1.0027	0.9973
10000	MML MDL CAICF SRM	9	0.0001	1.0067	0.9973

Chapter 7

Future Work

The experiments using the real data of atmospheric variables in Chapter 5 show that when using the proposed MML method as the cost function, the search strategy consistently picked less complex models with either better or slightly worse performance based on the message length and generalisation ability on the test data. Improvements on the message length calculation as shown in Equation 2.26 on page 36 will penalise more complex models less severely. This will allow MML to choose more complex models with longer message length but better fit to the data.

All optimisation search strategies other than the exhaustive brute-force approach have the potential of getting stuck in local minima. The search strategy proposed in this thesis has been designed to cover a large search space by trying all possible variables in making a decision to add/delete a variable to/from the model under consideration. However, improvements to the search strategy can be made by allowing it to try different starting points in the search space and allowing it to backtrack in its search process, i.e. try different paths with higher initial costs with the hope that it would jump particular local minima and converge to an even lower local minimum.

Non-homogeneity between the tropical cyclone data of the years 1950–1987 (used as the training data to built SHIFOR94) and 1988–1994 (used as the test data to built SHIFOR94) is suspected, based on the experiments on 11 subsets of the available data done in Chapter 5. With the assumption that the nature of climate may have changed over time and the way observational data have been collected may not be uniform from 1950 to 1994, further experiments can be done to see if we can actually build better forecasting models by not using all of the data from the earlier years.

There is actually no reason why we have to restrict ourselves to second order polynomial models. We did this because we did not want to have a larger search space than the search space used to built the benchmark models, SHIFOR and SHIFOR94. MML and the other three criteria can readily be used on a wide variety of models.

The widespread use of linear least squares regression analysis is merely due to the availability of easy to use tools/softwares, not because it is the best approach to take. We can consider taking the total least squares curve fitting approach (e.g. see [36]) which is a better and more robust approach. Instead of trying to fit regressor data points which are assumed to be correct, the total least squares curve fitting approach allows for errors in the regressor data points. With the real possibility of the introduction of human errors in the collection and recording of observational data required to build forecasting models, an approach that is not affected too much by errors in the data is likely to yield better models.

The model selection strategy proposed in this thesis can readily be applied to build tropical cyclone intensity forecasting models for the other basins in the Pacific and Indian oceans. The database produced by the NCEP/NCAR reanalysis project

[100, 32, 91] can be used as a source of high quality historical atmospheric data. The project is a cooperation between the US NCEP and NCAR (National Centers for Environmental Prediction and National Center for Atmospheric Research) to produce a 51-year (1948-1998) record of global analyses of atmospheric variables in support of the needs of the research and climate monitoring communities [100]. The effort involves setting up a global database as complete as possible of land surface, ship, rawinsonde, pibal, aircraft, satellite, and other data. Contributors to the database come from various different countries and organisations which also provide observations not available in real time for operations. The project also involves quality controlling and assimilating these data with a data assimilation system that is kept unchanged over the reanalysis period 1948-1998. This ensures that researchers can reliably compare recent anomalies with those in the earlier decade. An updated reanalysis using a state-of-the-art system every five years or so is planned.

Potential intensity (POT) is known to play a major part in tropical cyclone intensity forecasting. Therefore the use of a good thermodynamic model to the calculation of Maximum Potential Intensity (MPI) is imperative to statistical tropical cyclone intensity forecasting modelling. An immediate extension to this thesis is to use the NCEP/NCAR data with the new MPI thermodynamic model reported in [40] to build forecasting models for tropical cyclone intensity 24 hours into the future for the Australian basin. The new MPI model uses a combination of analysis of available observations and models (e.g.[81, 72, 33]). More studies on relating tropical cyclone intensity changes with seasonal/environmental aspects, like the study done in this thesis, and with climate change, e.g. [2], are needed.

Chapter 8

Conclusions

This thesis has contributed to the development of polynomial model selection research in two ways. First, by proposing a new second-order polynomial model selection model method based on the Minimum Message Length principle derived in Chapter 2. Second, by proposing a new model selection strategy that uses the combined results of four model selection methods, namely, MML, MDL, CAICF and SRM and a common search mechanism in Chapter 5. It has been shown in Chapter 5 that the methods, most notably MML and CAICF, and the model selection strategy help us explore large variable space and find both justifiably good models and variables which are common to good models.

This thesis has contributed to tropical cyclone research in two ways. The first contribution is that new tropical cyclone intensity forecasting models that are better than the models in operational use namely, SHIFOR and SHIFOR94 have been found in an automated manner in Chapter 5. With the ability to rank the performances of good models with increasing complexity, the model selection strategy proposed in this thesis allows a human expert to make the final decision on which one of the good models found will ultimately be selected to be used operationally. A model can be

selected because it has the right number of regressors according to a pre-determined model complexity. A model can also be chosen because it has the minimum MML message length, i.e. it is the best model found by the automated procedure.

The second contribution is that for the first time, to my knowledge, pronounced influence of seasonal variables are shown in a tropical cyclone intensity forecasting model. These seasonal variables have been reported to influence tropical cyclone activity in extensive studies in atmospheric science as explained in Chapter 4. This thesis shows for the first time that they also influence tropical cyclone intensity, as has been expected but has never been able to be proven before.

The four model selection methods and the optimisation search strategy included in the proposed model selection strategy have passed the tests in two sets of experiments. First, the experiments of recovering true models from artificial data sets with varying sample sizes from three true models in Chapter 3. Second, the experiments of recovering the proposed forecasting model from artificial data sets of varying sample size and levels of noise, generated from the covariance matrix of the real atmospheric data from which the model has been inferred. These experiments address the question of whether or not we have enough data to induce the model that we did induce in Chapter 5. These experiments are outlined in Chapter 6 together with the new procedure to recover model from artificial data of regressors whose covariance matrix equals that of the observation data.

Appendix A

Cost of Encoding Model Structure for the New MML Model Selection Criterion

The equation below is Equation 2.22 found on page 34. The equation shows the cost of encoding the structure of a second-order polynomial model comprising single and compound variables.

$$L_s = -\log h(\nu, j) - \log h(\xi, l) - \log {}_K C_k$$

A.1 Prior Probability Functions of Sending Single and Compound Variables

The prior probability function of sending single variables, $h(\nu, j)$, and compound variables, $h(\xi, l)$, are both proposed to follow the geometric series. The difference

between the two functions is in the assumption that it is cheaper to send a single variable than a compound variable. Therefore, the term ν is given a bigger value than ξ in the experiments.

$$\text{Given} \qquad h(\nu, j) = c\,\nu^j \qquad \text{where} \qquad \nu < 1 \tag{A.1}$$

$$\Rightarrow \qquad \sum_{j=0}^{J} c\,\nu^j = 1$$

$$= c\,(\nu^0 + \nu^1 + \nu^2 + \ldots + \nu^J) = 1$$

$$= c\,S = 1 \tag{A.2}$$

For the calculation of $h(\nu, j)$, the term c must be given in terms of the known terms ν, j and J.

$$\begin{aligned} S &= \nu^0 + \nu^1 + \nu^2 + \ldots + \nu^J \\ \nu\,S &= \quad + \nu^1 + \nu^2 + \ldots + \nu^J + \nu^{(J+1)} \\ \hline (1-\nu)\,S &= 1 \qquad\qquad\qquad\qquad - \nu^{(J+1)} \end{aligned}$$

$$\Leftrightarrow \qquad S = \frac{1-\nu^{(J+1)}}{1-\nu}$$

Substituting S in Equation A.2 gives

$$\sum_{j=0}^{J} c\,\nu^j = c \times \frac{1-\nu^{(J+1)}}{1-\nu} = 1$$

$$c = \frac{1-\nu}{1-\nu^{(J+1)}}$$

Finally, substituting c in Equation A.1 gives

$$h(\nu, j) = \frac{(1-\nu)\ \nu^j}{1-\nu^{(J+1)}}$$

A.2 Prior Probability Function of Sending Combination of Single and Compound Variables

The denominator of the prior probability function $\frac{1}{{}_K C_k}$ in Equation 2.22 is the number of combinations of K variables taken k at a time

$${}_K C_k = \frac{K!}{k!(K-k)!} \qquad \text{where:} \quad K = J + L \quad \text{and} \quad k = j + l$$

To avoid having to calculate the factorial of big numbers before taking the logarithm in Equation 2.22, the following simplification is done

$$\begin{aligned}
-\log \frac{1}{{}_K C_k} &= \log {}_K C_k \\
\Leftrightarrow \log {}_K C_k &= \log \frac{K!}{k!(K-k)!} \\
&= \log K! - \log(K-k)! - \log k! \\
&= \log[(K-0)\times(K-1)\times(K-2)\times\cdots\times\{K-(k-1)\}\times(K-k)!] - \\
&\quad - \log(K-k)! - \log[(k-0)\times(k-1)\times(k-2)\times\cdots\times 1] \\
&= \log[(K-0)\times(K-1)\times(K-2)\times\cdots\times\{K-(k-1)\}] - \\
&\quad - \log[(k-0)\times(k-1)\times(k-2)\times\cdots\times\{k-(k-1)\}] \\
&= \sum_{i=0}^{k-1}\Big[\log(K-i) - \log(k-i)\Big] \qquad \text{(A.3)}
\end{aligned}$$

Appendix B

The Fisher Information for the Polynomial Model

The Fisher information $I(\theta)$ is the determinant of the expected second derivative of the negative log likelihood function, $-\log P(y|\theta)$, given in Equation 2.10 on page 28, rewritten below:

$$L = -\log P(y|\theta) = -\log P(y|\sigma, \{\beta_k\}) = \frac{N}{2}\log 2\pi + N\log\sigma + \sum_{n=1}^{N}\frac{r_n^2}{2\sigma^2} \qquad \text{(B.1)}$$

where: $r_n = y_n - \sum_{k=1}^{K}\beta_k x_{nk}$

The Fisher information is derived as follows[1]:

[1] Similar derivations can be found in [21] and [97]

$$I(\theta) = I(\sigma, \{\beta_k\}) = \left| -E\frac{\partial^2}{\partial\theta_i\theta_j}\log P(y|\theta) \right| = \left| -\int P(y|\theta)\frac{\partial^2}{\partial\theta_i\theta_j}\log P(y|\theta)dx \right| \quad \text{(B.2)}$$

where: $\theta = \{\theta_k\} = \{\sigma, \beta_k\}$ and $i, j = 0, \ldots, k$

The first partial derivatives of Equation B.1 with respect to σ and β_k respectively are:

$$\frac{\partial L}{\partial \sigma} = \frac{N}{\sigma} - \frac{1}{\sigma^3}\sum_{n=1}^{N} r_n^2 \quad \text{(B.3)}$$

and

$$\frac{\partial L}{\partial \beta_k} = \frac{1}{2\sigma^2}\Big(2\sum_{n=1}^{N} r_n(-x_{nk})\Big) = \frac{1}{\sigma^2}\Big(-\sum_{n=1}^{N} r_n x_{nk}\Big) \quad \text{(B.4)}$$

The second partial derivatives are:

$$\frac{\partial^2 L}{\partial \sigma^2} = -\frac{N}{\sigma^2} + \frac{3}{\sigma^4}\sum_{n=1}^{N} r_n^2 \quad \text{(B.5)}$$

and

$$\frac{\partial^2 L}{\partial \beta_i \beta_j} = \frac{1}{\sigma^2}\sum_{n=1}^{N} x_{ni}x_{nj} \quad \text{(B.6)}$$

Based on Equation B.5, the expected second derivative of Equation B.1 with respect to σ is

$$E_y\left(\frac{\partial^2 L}{\partial \sigma^2}\right) = -\frac{N}{\sigma^2} + \frac{3}{\sigma^4}\sum_{n=1}^{N} E_y r_n^2 \tag{B.7}$$

The residual r_n comes from a Normal distribution, $r_n \sim N(0, \sigma^2)$. Therefore $E_y r_n^2 = \sigma^2$. Hence Equation B.7 becomes

$$E_y\left(\frac{\partial^2 L}{\partial \sigma^2}\right) = -\frac{N}{\sigma^2} + \frac{3}{\sigma^4} N\sigma^2 = \frac{2N}{\sigma^2} \tag{B.8}$$

The expected second derivative of Equation B.1 with respect to β_k remains unchanged

$$E_y\left(\frac{\partial^2 L}{\partial \beta_i \beta j}\right) = \frac{1}{\sigma^2}\sum_{n=1}^{N} x_{ni} x_{nj} \tag{B.9}$$

The expected derivative of Equation B.4 with respect to σ is null since $E_y(\sum_{n=1}^{N} r_n) = 0$

$$E_y\left(\frac{\partial^2 L}{\partial \sigma \partial \beta_k}\right) = E_y\left(\frac{-2}{\sigma^3}\sum_{n=1}^{N} r_n x_{nk}\right) = 0 \tag{B.10}$$

This means the values of the off-diagonal blocks of the first row and column in the Fisher information matrix are null.

The Fisher information is therefore

$$I(\theta) = I(\sigma, \{\beta_k\}) = \begin{vmatrix} \frac{2N}{\sigma^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma^2}x_1 \cdot x_1 & \frac{1}{\sigma^2}x_1 \cdot x_2 & \dots & \frac{1}{\sigma^2}x_1 \cdot x_K \\ 0 & \frac{1}{\sigma^2}x_2 \cdot x_1 & \frac{1}{\sigma^2}x_2 \cdot x_2 & \dots & \frac{1}{\sigma^2}x_2 \cdot x_K \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \frac{1}{\sigma^2}x_K \cdot x_1 & \frac{1}{\sigma^2}x_K \cdot x_2 & \dots & \frac{1}{\sigma^2}x_K \cdot x_K \end{vmatrix} \tag{B.11}$$

where $x_i = (x_{1i}, x_{2i}, \dots, x_{Ni})(i = 1, 2, \dots, K)$

Equation B.11 can also be expressed as

$$I(\theta) = I(\sigma, \{\beta_k\}) = \frac{2N}{\sigma^2} \times \frac{1}{\sigma^{2K}} \left| x_i \cdot x_j \right|_{K \times K} = 2N\sigma^{-2(K+1)} \left| x_i \cdot x_j \right|_{K \times K}$$

as given in Equation 2.29 on page 38.

Appendix C

Non-backtracking Search Algorithm

The following is a pseudo-code of the non-backtracking search algorithm outlined in Section 3.2.2 on page 57.

Procedure 3 Search for model in a search space of regressors using a cost function to compare two models

---

module nonBacktrackingSearch() {

input training data

input test data

standardize variables of training data

compute product of 2 standardized variables /* e.g. $(var_0 * var_0)$, $(var_0 * var_1)$, etc */

standardize product of 2 standardized variables

store all potential regressors to regressorList /* single and compound variables */

initialise modelList to empty

initialise lowestCost to the upper limit of floating point value

```
    findFirstModel(regressorList, modelList, lowestCost) /* find the first model */

    loop {

        if (regressorList is empty OR model with the lowest cost is found) then
            exit loop

        addANewVariableToModel(regressorList, modelList, hasBeenAdded, newAddition, lowestCost)

        if (hasBeenAdded is true) then
            deleteAVariableFromModel(regressorList, modelList, newAddition, lowestCost)

    } /* end loop */


} /* end of module nonBacktrackingSearch() */


submodule findFirstModel(regressorList, modelList, lowestCost) {

    use the first variable from regressorList to make a model

    loop { /* traverse regressorList */

        calculate the parameters of the model
        calculate the cost of the model modelCost

        if (modelCost is less than lowestCost) then {
            set lowestCost to modelCost
            record the new model newModel
        } /* end if */

        if (at the end of regressorList) then
            exit loop
        else
            use the next variable from regressorList to make a model

    } /* end loop */
```

```
    put the chosen variable to modelList
    remove the chosen variable from regressorList

} /* end of submodule findFirstModel()


submodule addANewVariableToModel (regressorList, modelList,
hasBeenAdded, newAddition, lowestCost) {

    initialise hasBeenAdded to false
    copy the first variable from regressorList to modelList
    copy the variable just added to modelList to justAdded

    loop { /* traverse regressorList */

        calculate the parameters of the new model
        calculate the cost of the new model modelCost

        if (modelCost is less than lowestCost) then {
            set lowestCost to modelCost
            record justAdded in newAddition
            set hasBeenAdded to true
        }

        if (at end of regressorList) then
            exit loop
        else {
            delete justAdded from modelList
            copy the next variable from regressorList to modelList
            copy the variable just added to modelList to justAdded
        } /* end else */

    } /* end loop */

    if (hasBeenAdded is true) then
        remove newAddition from regressorList
```

```
} /* end of addANewVariableToModel()

submodule deleteAVariableFromModel(regressorList, modelList,
newAddition, lowestCost) {

    initialise deletionPossible to false
    use the first variable from model for deletion as toBeDeleted

    loop { /* traverse modelList */

        if (toBeDeleted = newAddition) then
            disregard toBeDeleted
        else {

            delete toBeDeleted from modelList
            calculate the parameters of the new model
            calculate the cost of the new model modelCost

            if (modelCost is less than lowestCost) then {
                set lowestCost to modelCost
                record the variable in newAddition
                        set deletionPossible to true
            } /* end if */

        } /* end else */

        if (at end of modelList) then
            exit loop
        else {
            put toBeDeleted back onto modelList
            use the next variable from model for deletion as toBeDeleted
        } /* end else */

    } /* end of loop */
```

```
    if (deletionPossible is true) then {
        delete toBeDeleted from modelList
        put toBeDeleted back onto regressorList
    } /* end if */

} /* end of deleteAVariableFromModel() */
```

Appendix D

Sample Trace of the Non-backtracking Search Algorithm

The following is an example of the trace of an experiment using the non-backtracking search algorithm given in Appendix C. It shows how the search algorithm adds and deletes a variable in the process of finding a model with the lowest cost. In this example, the cost function used is the MML method. The result of this experiment is shown for data set 1 in Table 5.3 on page 117. This kind of trace record might be of interest to domain experts in the tropical cyclone research to see which variables get added/deleted and in which order they are added/deleted.

```
====================
Start time: 16:22:20 on Thursday, 15 April 1999
====================
Start Search: 16:22:30 on Thursday, 15 April 1999
Duration (hr:min:sec)= 0:0:10.000
====================
TRACE RECORD

Search Algorithm:  Non-backtracking Optimisation Search
Matrix Computation:  Jacobi Rotation
Parameter estimates: minimize ML
```

D. SAMPLE TRACE OF THE NON-BACKTRACKING SEARCH ALGORITHM

```
Objective Function:  Message Length

Cases: Training= 3042

Dependent Var : DelVmax
Independent Vars:
1-Julian 2-JulOff 3-LatData 4-LonData 5-Vmax 6-Del12V 7-UCurr
8-VCurr 9-Speed 10-POT 11-POTend 12-DelSST 13-SSST
14-SSSTend 15-DSSST 16-UppSpd 17-Uppend 18-DUppSpd 19-Stabil
20-Stabend 21-DelStab 22-200mbT 23-200Tend 24-Del200T 25-DisLand
26-Closest 27-200mbU 28-200Uend 29-Del200U 30-U50 31-RainS
32-RainG 33-SLPA 34-ZWA 35-ElNino 36-SOI

insert item = 10
modelList: 10 NULL
nvar= 1 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.9997 S/N ratio = 1.0006
Part1ML= 14.7083 Part2ML= 3710.3460 TotML= 3725.0543
Training Data: MAE= 18.8861 MSE= 558.4768 RMSE= 23.6321 R^2= 0.3288
trainErrMin = -73.2008 trainErrMax = 79.6092

normCoeffs: 0.573186

insert item = 6
modelList: 6 10 NULL
nvar= 2 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.8194 S/N ratio = 1.4895
Part1ML= 24.1115 Part2ML= 3659.0049 TotML= 3683.1164
Training Data: MAE= 18.4746 MSE= 540.1087 RMSE= 23.2402 R^2= 0.3510
trainErrMin = -83.3501 trainErrMax = 83.4524

normCoeffs: 0.149365 0.579024

insert item = (11 5)
modelList: (11 5) 6 10 NULL
nvar= 3 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.8057 S/N ratio = 1.5404
Part1ML= 37.0855 Part2ML= 3627.4175 TotML= 3664.5031
Training Data: MAE= 18.1410 MSE= 529.1762 RMSE= 23.0038 R^2= 0.3644
trainErrMin = -84.1088 trainErrMax = 89.1333

normCoeffs: 0.120492 0.133868 0.547719

insert item = 31
modelList: 31 (11 5) 6 10 NULL
nvar= 4 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7974 S/N ratio = 1.5727
Part1ML= 45.9499 Part2ML= 3603.7014 TotML= 3649.6513
Training Data: MAE= 17.9423 MSE= 521.1563 RMSE= 22.8288 R^2= 0.3742
trainErrMin = -83.4998 trainErrMax = 91.1975

normCoeffs: 0.0994645 0.12669 0.131041 0.550648

insert item = 13
modelList: 13 31 (11 5) 6 10 NULL
nvar= 5 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7912 S/N ratio = 1.5975
Part1ML= 54.4731 Part2ML= 3586.4292 TotML= 3640.9023
Training Data: MAE= 17.9660 MSE= 515.4383 RMSE= 22.7033 R^2= 0.3813
trainErrMin = -81.1858 trainErrMax = 90.6144

normCoeffs: -0.107349 0.10717 0.118044 0.144514 0.618286

insert item = (36 36)
```

D. SAMPLE TRACE OF THE NON-BACKTRACKING SEARCH ALGORITHM

```
modelList: (36 36) 13 31 (11 5) 6 10 NULL
nvar= 6 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7866 S/N ratio = 1.6164
Part1ML= 66.6803 Part2ML= 3566.2705 TotML= 3632.9509
Training Data: MAE= 17.7305 MSE= 508.8159 RMSE= 22.5569 R^2= 0.3895
trainErrMin = -79.9251 trainErrMax = 93.5904

normCoeffs: -0.0910103 -0.116782 0.10663 0.119114 0.140929 0.631133

insert item = (6 5)
modelList: (6 5) (36 36) 13 31 (11 5) 6 10 NULL
nvar= 7 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7814 S/N ratio = 1.6379
Part1ML= 78.4700 Part2ML= 3546.8198 TotML= 3625.2898
Training Data: MAE= 17.5745 MSE= 502.5126 RMSE= 22.4168 R^2= 0.3972
trainErrMin = -71.8852 trainErrMax = 97.152

normCoeffs: -0.100056 -0.0895028 -0.116281 0.105045 0.111905 0.187857 0.644023

insert item = 7
modelList: 7 (6 5) (36 36) 13 31 (11 5) 6 10 NULL
nvar= 8 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7763 S/N ratio = 1.6593
Part1ML= 86.7060 Part2ML= 3532.0455 TotML= 3618.7516
Training Data: MAE= 17.5117 MSE= 497.8166 RMSE= 22.3118 R^2= 0.4030
trainErrMin = -72.6425 trainErrMax = 95.2591

normCoeffs: -0.090999 -0.102919 -0.0859313 -0.154795 0.0991577 0.122989
0.181805 0.630475

insert item = (28 5)
modelList: (28 5) 7 (6 5) (36 36) 13 31 (11 5) 6 10 NULL
nvar= 9 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7726 S/N ratio = 1.6754
Part1ML= 98.1906 Part2ML= 3516.3351 TotML= 3614.5257
Training Data: MAE= 17.5035 MSE= 492.8608 RMSE= 22.2005 R^2= 0.4092
trainErrMin = -73.0541 trainErrMax = 95.754

normCoeffs: -0.101474 -0.0951706 -0.118847 -0.090541 -0.196298 0.097955
0.0643029 0.187051 0.665149

delete item = (11 5)
modelList: (28 5) 7 (6 5) (36 36) 13 31 6 10 NULL
nvar= 8 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.9017 S/N ratio = 1.2298
Part1ML= 86.7112 Part2ML= 3522.6119 TotML= 3609.3231
Training Data: MAE= 17.6388 MSE= 494.7368 RMSE= 22.2427 R^2= 0.4067
trainErrMin = -73.455 trainErrMax = 94.1551

normCoeffs: -0.13836 -0.0901097 -0.127962 -0.0919673 -0.213527 0.0958663
0.197083 0.692944

insert item = (13 13)
modelList: (13 13) (28 5) 7 (6 5) (36 36) 13 31 6 10 NULL
nvar= 9 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7702 S/N ratio = 1.6859
Part1ML= 98.2003 Part2ML= 3499.0281 TotML= 3597.2284
Training Data: MAE= 17.4754 MSE= 487.2813 RMSE= 22.0744 R^2= 0.4159
trainErrMin = -73.1377 trainErrMax = 95.0036

normCoeffs: -0.145716 -0.145369 -0.113623 -0.125753 -0.0988746 -0.345855
0.101908 0.195974 0.711732

insert item = (35 29)
```

D. SAMPLE TRACE OF THE NON-BACKTRACKING SEARCH ALGORITHM

```
modelList: (35 29) (13 13) (28 5) 7 (6 5) (36 36) 13 31 6 10 NULL
nvar= 10 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7640 S/N ratio = 1.7134
Part1ML= 109.4572 Part2ML= 3480.0766 TotML= 3589.5338
Training Data: MAE= 17.4447 MSE= 481.4029 RMSE= 21.9409 R^2= 0.4231
trainErrMin = -75.5681 trainErrMax = 93.8188

normCoeffs: 0.0861333 -0.15767 -0.148664 -0.111276 -0.124994 -0.107793
-0.358079 0.102841 0.190231 0.712692

insert item = (32 11)
modelList: (32 11) (35 29) (13 13) (28 5) 7 (6 5) (36 36) 13 31 6 10 NULL
nvar= 11 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7593 S/N ratio = 1.7343
Part1ML= 120.5235 Part2ML= 3465.9627 TotML= 3586.4862
Training Data: MAE= 17.3745 MSE= 477.1113 RMSE= 21.8429 R^2= 0.4284
trainErrMin = -74.4932 trainErrMax = 95.4231

normCoeffs: 0.0740721 0.0781943 -0.154267 -0.146956 -0.11136 -0.125287
-0.103434 -0.343371 0.0990517 0.188432 0.704921

insert item = (34 11)
modelList: (34 11) (32 11) (35 29) (13 13) (28 5) 7 (6 5) (36 36) 13 31
6 10 NULL
nvar= 12 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7559 S/N ratio = 1.7503
Part1ML= 131.4335 Part2ML= 3447.0052 TotML= 3578.4387
Training Data: MAE= 17.2348 MSE= 471.3538 RMSE= 21.7107 R^2= 0.4355
trainErrMin = -72.6279 trainErrMax = 95.8296

normCoeffs: 0.0926964 0.108976 0.0659722 -0.146224 -0.147629 -0.108669
-0.124631 -0.109306 -0.34562 0.0971674 0.187384 0.705484

insert item = 2
modelList: 2 (34 11) (32 11) (35 29) (13 13) (28 5) 7 (6 5) (36 36) 13 31
6 10 NULL
nvar= 13 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7509 S/N ratio = 1.7736
Part1ML= 139.4338 Part2ML= 3432.8788 TotML= 3572.3125
Training Data: MAE= 17.1138 MSE= 467.1480 RMSE= 21.6136 R^2= 0.4407
trainErrMin = -72.4637 trainErrMax = 94.731

normCoeffs: -0.0750469 0.102027 0.116134 0.0649468 -0.147106 -0.156491
-0.0943276 -0.124807 -0.112183 -0.353268 0.0984102 0.19198 0.718202

delete item = (35 29)
modelList: 2 (34 11) (32 11) (13 13) (28 5) 7 (6 5) (36 36) 13 31
6 10 NULL
nvar= 12 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.9236 S/N ratio = 1.1723
Part1ML= 128.5313 Part2ML= 3443.6745 TotML= 3572.2059
Training Data: MAE= 17.1084 MSE= 470.3221 RMSE= 21.6869 R^2= 0.4368
trainErrMin = -69.8368 trainErrMax = 95.7729

normCoeffs: -0.0759514 0.111955 0.126686 -0.137223 -0.154106 -0.0955914
-0.125321 -0.105926 -0.343325 0.0971979 0.195955 0.717003

insert item = 35
modelList: 35 2 (34 11) (32 11) (13 13) (28 5) 7 (6 5) (36 36) 13 31
6 10 NULL
nvar= 13 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7498 S/N ratio = 1.7787
Part1ML= 136.3430 Part2ML= 3432.3754 TotML= 3568.7184
Training Data: MAE= 17.0446 MSE= 466.9933 RMSE= 21.6100 R^2= 0.4409
```

```
trainErrMin = -71.0721 trainErrMax = 93.3476

normCoeffs: -0.0675735 -0.0764607 0.113963 0.132555 -0.144827 -0.149843
-0.0972769 -0.124134 -0.0898984 -0.352892 0.0938223 0.193789 0.713505

insert item = (35 29)
modelList: (35 29) 35 2 (34 11) (32 11) (13 13) (28 5) 7 (6 5) (36 36)
13 31 6 10 NULL
nvar= 14 meanY= 9.4001 stddevY = 28.8350
normstddevY= 1.0000 normstddevErr= 0.7470 S/N ratio = 1.7921
Part1ML= 147.2432 Part2ML= 3421.1167 TotML= 3568.3599
Training Data: MAE= 17.0201 MSE= 463.7004 RMSE= 21.5337 R^2= 0.4451
trainErrMin = -75.1345 trainErrMax = 92.2469

normCoeffs: 0.066083 -0.068704 -0.0755489 0.103895 0.121917 -0.155011
-0.152198 -0.0960192 -0.123591 -0.095996 -0.363169 0.0949993 0.189708 0.714666

Nothing more to insert/delete - end search

Trace Summary:

nvar= 14 meanY=     9.4001 stddevY =    28.8350
normstddevY=     1.0000 normstddevErr=     0.7440 S/N ratio =     1.8066
Penalty WallaceMML:  3568.3599 Part1ML=   147.2432 Part2ML=  3421.1167
TotML=  3568.3599
Training Data: MAE= 17.0201 MSE= 463.7004 RMSE= 21.5337 R^2= 0.4451
Test Data:     MAE= 17.3790 MSE= 481.1881 RMSE= 21.9360 R^2= 0.4578
trainErrMin =   -75.1345 trainErrMax =    92.2469
testErrMin  =   -77.2743 testErrMax  =    84.8163


No.   VarNum   VarName                  NormCoeff  RealCoeff
---   ------   --------------------     ---------  ---------
  1  (35,29)  (   ElNino,  Del200U)   0.066083   2.035125
  2       35                 ElNino  -0.068704  -0.031417
  3        2                 JulOff  -0.075549  -0.121078
  4  (34,11)  (      ZWA,    POTend)   0.103895   3.111376
  5  (32,11)  (    RainG,    POTend)   0.121917   3.506426
  6  (13,13)  (     SSST,      SSST)  -0.155011  -1.682023
  7  (28, 5)  (  200Uend,      Vmax)  -0.152198  -4.347148
  8        7                  UCurr  -0.096019  -0.336593
  9  ( 6, 5)  (   Del12V,      Vmax)  -0.123591  -3.542231
 10  (36,36)  (      SOI,       SOI)  -0.095996  -2.040737
 11       13                  SSST  -0.363169  -4.290428
 12       31                 RainS   0.094999   4.231230
 13        6                Del12V   0.189708   0.620980
 14       10                   POT   0.714666   0.606140

Constant: 79.922024
==================
End time: 16:26:38 on Thursday, 15 April 1999
Duration (hr:min:sec)= 0:4:18.000
==================
```

References

[1] Footy Tipping Competition. http://www.csse.monash.edu.au/~footy, 2002.

[2] A. Henderson-Sellers, et al. Tropical Cyclones and Global Climate Change: Reports from the WMO/CAS/TMRP Committee on Climate Change Assessment (Project TC-2). http://www.bom.gov.au/bmrc/meso/Project/tcclimat.htm, 1997.

[3] A.E. Hoerl. Applications of Ridge Analysis to Regression Problems. *Chemical Engineering Progress*, 58:54–59, 1962.

[4] A.E. Hoerl and R.W. Kennard. Ridge Regression: Applications to NonOrthogonal Problems. *Technometrics*, 12:69–82, 1970. Erratum: vol 12, p.723.

[5] A.E. Hoerl and R.W. Kennard. Ridge Regression: Biased Estimation for NonOrthogonal Problems. *Technometrics*, 12:55–67, 1970.

[6] A.J. Miller. *Subset Selection in Regression.* Chapman and Hall, London, 1990.

[7] A.K. Betts. A New Convective Adjustment Scheme. Part I: Observational and Theoretical Basis. *Journal of Royal Meteorological Society Quarterly*, 112:677–691, 1986.

[8] A.U. White. Global Summary of Human Response to Natural Hazards: Tropical Cyclone. In Gilbert F. White, editor, *Natural Hazards: Local, National, Global*, pages 255–265. Oxford University Press, New York, 1974.

[9] B. Crowder. *The Wonders of the Weather*, chapter 9, pages 169–199. Australian Government Publishing Service, Canberra, Australia, 1995.

[10] B.R. Jarvinen and C.J. Neumann. Statistical Forecasts of Tropical Cyclone Intensity. Technical Report Tech. Memo. NWS NHC-10, National Oceanic and Atmospheric Administration (NOAA), Miami, Florida, 1979.

[11] B.R. Jarvinen and E.L. Caso. A Tropical Cyclone Data Tape for the North Atlantic Basin, 1886–1977: Contents, Limitations and Uses. Technical Report Tech. Memo. NWS NHC-6, National Oceanic and Atmospheric Administration (NOAA), Fort Worth, Texas, 1978.

[12] Bureau of Meteorology – Queensland Regional Office. Understanding Cyclones. Booklet B92/22364 Cat.No.92 2738 1, 1992.

[13] C. Mallows. Some Comments on C_p. *Technometrics*, 15:661–675, 1973.

[14] C.E. Shannon. A Mathematical Theory of Communication. *Bell Systems Technical Journal*, 27:379–423, 1948.

[15] C.J. Neumann. An Alternate to the HURRAN tropical cyclone forecasting system. Technical Memo NWS-62, NOAA, 1972.

[16] C.S. Wallace. Inference and Estimation by Compact Coding. Technical Report 46, School of Computer Science and Software Engineering, Monash University, 1984.

[17] C.S. Wallace. Fast Pseudorandom Generators for Normal and Exponential Variates. *ACM Transactions on Mathematical Software*, 22(1):119–127, March 1996.

[18] C.S. Wallace and D.M. Boulton. An Information Measure for Classification. *Computer Journal*, 11:185–195, 1968.

[19] C.S. Wallace and K. Korb. Learning Linear Causal Models by MML Sampling. Technical Report 310, School of Computer Science and Software Engineering, Monash University, 1997.

[20] C.S. Wallace and P.R. Freeman. Estimation and Inference by Compact Coding. *Journal of the Royal Statistical Society B*, 49(1):240–252, 1987.

[21] C.S. Wallace, K. Korb and H. Dai. Causal Discovery via MML. In *Proceedings of the Thirteenth International Conference of Machine Learning*, pages 516–524. Morgan Kaufmann Publisher, Bari, 1996.

[22] C.S. Wallace, K. Korb and H. Dai. MML Induction of Causal Models by Sampling Posterior Probabilities. unpublished technical report - School of Computer Science and Software Engineering, Monash University, 1996.

[23] C.W. Landsea. A Climatology of Intense (or Major) Atlantic Hurricanes. *Monthly Weather Review*, 121:1703–1713, June 1993.

[24] C.W. Landsea. SHIFOR94 - Atlantic Tropical Cyclone Intensity Forecasting. In *Proceedings of the 21st Conference on Hurricanes and Tropical Meteorology*, pages 365–367, Miami, Florida, 1995. American Meteorological Society.

[25] C.W. Landsea and W.M. Gray. The Strong Association between Western Sahelian Monsoon Rainfall and Intense Atlantic Hurricanes. *Journal of Climate*, 5:435–453, May 1992.

[26] D. Dowe and N. Krusel. A Decision Tree Model of Bushfire Activity. In *Proc. of 6th Australian Joint Conf. on Artificial Intelligence*, 1993.

[27] D. Miller and J.M. Fritsch. Mesoscale Convective Complexes in the Western Pacific Region. *Monthly Weather Review*, 119:2978–2992, 1991.

[28] D.M. Allen. The Relationship between Variable Selection and Data Augmentation and a Method for Prediction. *Ann. Inst. Statist. Math.*, 21:243–247, 1974.

[29] D.M. Boulton. Numerical Classification based on an Information Measure. Master's thesis, Basser Computing Department, University of Sydney, Sydney, Australia, 1970.

[30] D.M. Boulton. *The Information Measure Criterion for Intrinsic Classification.* PhD thesis, Department of Computer Science, Monash University, Melbourne, Australia, 1975.

[31] D.R. Powell, D.L. Dowe, L. Allison, T.I. Dix. Discovering Simple DNA Sequences by Compression. In *Pacific Symp. on Biocomputing (PSB98)*, pages 595–606, 1998.

[32] E. Kalnay, et al. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, 77:437–470, March 1996.

[33] E. Kleinschmidt, Jr. Grundlagen Einer Theorie des Tropischen Zyklonen. *Archiv für Meteorologie, Geophysik und Bioklimatologie, Ser. A*, 4:53–72, 1952.

[34] E. Simiu and R.H. Scanlon. *Wind Effects on Structures.* Wiley–Interscience, New York, 1978.

[35] E.F. Halpern. Polynomial Regression from a Bayesian Approach. *Journal of the American Statistical Association*, 68:137–143, 1973.

[36] G. H. Golub and C. F. Van Loan. An Analysis of the Total Least Squares Problem. *SIAM Journal of Numerical Analysis*, 17:883–893, 1980.

[37] G. Rumantir, A. Henderson-Sellers and G. Holland. Causal Bayesian Reasoning for Tropical Cyclone Risk Analysis: Research in Progress. In *Proceedings of the Third Workshop of Tropical Cyclone Coastal Impacts Project (TCCIP)*, pages 26–27, Brisbane, 1996. Tropical Cyclone Coastal Impacts Project.

[38] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6:461–464, 1978.

[39] G.J. Holland, editor. *Global Guide to Tropical Cyclone Forecasting.* Technical Documentation: WMO/TD - No. 560, Tropical Cyclone Programme Report No. TCP-31. World Meteorological Organization, Geneva - Switzerland, 1993.

[40] G.J. Holland. The maximum Potential Intensity of Tropical Cyclones. *Journal of Atmospheric Science*, 54:2519–2541, 1997.

[41] G.J. Holland and R.T. Merrill. On the Dynamic of Tropical Cyclone Structure Changes. *Quarterly Journal of Royal Meteorological Society*, 110:723–745, 1984.

[42] G.W. Rumantir. Minimum Message Length Criterion for Second-order Polynomial Model Discovery. In T. Terano, H. Liu, A.L.P. Chen, editor, *Knowledge*

Discovery and Data Mining: Current Issues and New Applications, PAKDD 2000, LNAI 1805, pages 40–48. Springer–Verlag, Berlin Heidelberg, 2000.

[43] G.W. Rumantir. Tropical Cyclone Intensity Forecasting Model: Balancing Complexity and Goodness of Fit. In R. Mizoguchi and J. Slaney , editor, *PRICAI 2000 Topics in Artificial Intelligence, LNAI 1886*, pages 230–240. Springer–Verlag, Berlin Heidelberg, 2000.

[44] G.W. Rumantir and C.S. Wallace. Sampling of Highly Correlated Data for Polynomial Regression and Model Discovery. In F. Hoffmann, et al., editor, *Advances in Intelligent Data Analysis, LNCS 2189*, pages 370–377. Springer–Verlag, Berlin Heidelberg, 2001.

[45] G.W. Rumantir and C.S. Wallace. Minimum Message Length Criterion for Second-order Polynomial Model Selection Applied to Tropical Cyclone Intensity Forecasting. In M.R. Berthold, et al., editor, *Advances in Intelligent Data Analysis, LNCS2810*, pages 486–496. Springer–Verlag, Berlin Heidelberg, 2003.

[46] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In B.N. Petrov and F. Csaki, editor, *Proc. of 2nd Int. Symp. Information Thy.*, pages 267–281, 1973.

[47] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC-19:716–723, 1974.

[48] H. Akaike. On the Entropy Maximization Principle. In P.R. Krishniah, editor, *Applications on Statistics*, pages 27–41. North Holland, Amsterdam, 1977.

[49] H. Akaike. Prediction and Entropy. In A.C. Atkinson and S.E. Fiendberg, editor, *A Celebration of Statistics*, pages 1–24. Springer Verlag, New York, 1985.

[50] H. Bozdogan. Model Selection and Akaike's Information Criterion (AIC): the General Theory and its Analytical Extensions. *Psychometrika*, 52(3):345–370, 1987.

[51] H.L. Kuo. On Formation and Intensification of Tropical Cyclones through Latent Heat Release by Cumulus Convection. *Journal of Atmospheric Science*, 22:40–63, 1965.

[52] H.L. Kuo. Further Studies of the Parameterization of the Influence of Cumulus Convection on Large-scale Flow. *Journal of Atmospheric Science*, 31:1232–1240, 1974.

[53] J. Baik, M. DeMaria and S. Raman. Tropical Cyclone Simulations with the Betts Convective Adjustment Scheme. Part I: Model Description and Control Simulation. *Monthly Weather Review*, 118:513–528, 1990.

[54] J. Baik, M. DeMaria and S. Raman. Tropical Cyclone Simulations with the Betts Convective Adjustment Scheme. Part III: Comparisons with the Kuo Convective Parameterization. *Monthly Weather Review*, 119:2889–2899, 1991.

[55] J. Molinari and D. Vollaro. External Influences on Hurricane Intensity. Part 1: Outflow Layer Eddy Angular Momentum Fluxes. *Journal of Atmospheric Science*, 46:1093–1105, 1989.

[56] J. Patrick and C.S. Wallace. Stone Circles: A Comparative Analysis of Megalithic Geometry. In *Proc. of 48th ANZA AS Conf.*, 1977.

[57] J. Rissanen. Modeling by Shortest Data Description. *Automatica*, 14:465–471, 1978.

[58] J. Rissanen. Stochastic Complexity and Modeling. *Annal of Statistics*, 14:1080–1100, 1986.

[59] J. Rissanen. Stochastic Complexity. *Journal of the Royal Statistical Society B*, 49(1):223–239, 1987.

[60] J.D. Patrick. *An Information Measure Comparative Analysis of Megalithic Geometries.* PhD thesis, Department of Computer Science, Monash University, Melbourne, Australia, 1979.

[61] J.F. Brennan. Relation of May-June Weather Conditions in Jamaica to the Caribbean Tropical Disturbances of the Following Season. *Monthly Weather Review*, 63:13–14, 1935.

[62] J.H. Conway and N.J.A. Sloane. *Sphere Packings, Lattices and Groups.* Springer–Verlag, New York, 1988.

[63] J.J. Oliver and D.J. Hand. Introduction to Minimum Encoding Inference. Technical Report 205, School of Computer Science and Software Engineering, Monash University, 1994.

[64] J.M. Wallace and P.V. Hobbs. *Atmospheric Science: an Introductory Survey.* Academic Press, Inc., 1977.

[65] J.R. Hope and C. J. Neumann. A survey of worldwide tropical cyclone prediction models. In *Proceedings of the 21st Conference on Hurricanes and Tropical Meteorology*, pages 367–374, Miami, Florida, 1977. American Meteorological Society.

[66] J.S. Hobgood and J.N. Rayne. A Test of Convective Parameterizations in a Tropical Cyclone Model. *Monthly Weather Review*, 117:1221–1226, 1989.

[67] J.S. Kain and J.M. Fritsch. The Impact of Model Physics on Numerical Simulations of Tropical Cyclone Irma. In *Proceedings of the 20st Conference on Hurricanes and Tropical Meteorology*, pages 135–138, Miami, Florida, 1993. American Meteorological Society.

[68] K. Puri and M.J. Miller. Relation of Tropical Cyclone Frequency to Summer Pressures and Ocean Surface-Water Temperatures. *Monthly Weather Review*, 63:10–12, 1935.

[69] K. Puri and M.J. Miller. Sensitivity of ECMWF Analyses - Forecasts of Tropical Cyclone to Cumulus Parameterization. *Monthly Weather Review*, 118:1709–1741, 1990.

[70] K.A. Emanuel. An Air-Sea Interaction Theory for Tropical Cyclones. Part 1: Steady-state Maintenance. *Journal of Atmospheric Science*, 43:585-604, 1986.

[71] K.A. Emanuel. The Maximum Intensity of Hurricanes. *Journal of Atmospheric Science*, 45:1143–1155, 1988.

[72] K.A. Emanuel. The Theory of Hurricanes. *Annual Review of Fluid Mechanics*, 23:179–196, 1991.

[73] K.C. Li. Asymptotic Optimality for C_p, C_L, Cross-validation, and Generalized Cross-validation: Discrete Index Set. *Annal of Statistics*, 15:958–975, 1987.

[74] L. Allison. Inductive Inference and machine learning using Minimum Message Length MML encoding. http://www.csse.monash.edu.au/~lloyd/tildeMML, 2002.

[75] L. Allison, L. Stern, T. Edgoose, T.I. Dix. Sequence Complexity for Biological Sequence Analysis. *Computers and Chemistry*, 24(1):43–55, January 2000.

[76] L. Allison, T. Edgoose, T.I. Dix. Compression of Strings with Approximate Repeats. In *Intelligent Systems in Molecular Biology (ISMB'98)*, pages 8–16, Montreal, 1998.

[77] L. Stern, L. Allison, R.L. Coppel, T.I. Dix. Discovering Patterns in Plasmodium Falciparum Genomic DNA. *Molecular and Biochemical Parasitology*, 118(2):175–186, December 2001.

[78] L.J. Shapiro. Hurricane Climate Fluctuations. Part II: Relation to Large-scale Circulation. *Monthly Weather Review*, 110:1014–1023, 1982.

[79] L.J. Shapiro. Month-to-month Variability of the Atlantic Tropical Circulation and its Relationship to Tropical Storm Formation. *Monthly Weather Review*, 115:1545–1552, 1987.

[80] L.J. Shapiro. The Relationship of the Quasi-Biennial Oscillation to Atlantic Tropical Storm Activity. *Monthly Weather Review*, 117:1545–1552, 1989.

[81] M. DeMaria and J. Kaplan. Sea surface temperature and the maximum potential intensity of atlantic tropical cyclones. *Journal of Climate*, 7:1324–1334, 1994.

[82] M. DeMaria and J. Kaplan. A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic Basin. *Weather and Forecasting*, 9(2):209–220, June 1994.

[83] M. DeMaria, S.D. Aberson, K.V. Ooyama and S.J. Lord. A Nested Spectral Model for Hurricane Track Forecasting. *Monthly Weather Review*, 120:1628–1643, 1992.

[84] M. Ezekiel. *Methods of Correlation Analysis.* Wiley, New York, 1930.

[85] M.A. Bender, I. Ginis and Y. Kurihara. Numerical Simulations of Tropical Cyclone Ocean Interactions with a High-resolution Model. *Monthly Weather Review*, 121:2046–2061, 1993.

[86] M.A. Bender, R.J. Ross, R.E. Tuleya and Y. Kurihara. Improvements in Tropical Cyclone Track and Intensity Forecasts Using the GFDL Initialization System. *Monthly Weather Review*, 98:245–263, 1993.

[87] M.A. Efroymson. Multiple Regression Analysis. In A. Ralston and H.S. Wilf, editor, *Mathematical Methods for Digital Computers*, pages 191–203. Wiley, New York, 1960.

[88] M.P. Georgeff and C.S. Wallace. A General Selection Criterion for Inductive Inference. In *Proc. of 6th European Conference on Artificial Intelligence*, pages 473–482, 1993.

[89] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[90] National Aeronautics and Space Administration (NASA). Earth Observatory:Reference. http://earthobservatory.nasa.gov:81/Library. Accessed: September 1, 2003.

[91] NOAA-CIRES Climate Diagnostics Center. The NCEP/NCAR Reanalysis Project at the NOAA-CIRES Climate Diagnostics Center . http://www.cdc.noaa.gov/cdc/reanalysis/. Accessed: September 17, 2003.

[92] P. Gray, W. Hart, L. Painton, C. Phillips, M. Trahan, J. Wagner. A Survey of Global Optimization Methods. http://www.cs.sandia.gov/opt/survey/sa.html, 1997.

[93] P. J. Fitzpatrick. *Understanding and Forecasting Tropical Cyclone Intensity Change.* PhD thesis, Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, 1996.

[94] P. Tischer. MML and the Science of Information or Artificially Assisted Intelligence.

[95] P.A. Arkin. The Relationship between Interannual Variability in the 200mb Tropical Wind Field and the Southern Oscillation. *Monthly Weather Review*, 110:1393–1401, 1982.

[96] P.J. Fitzpatrick. Forecasting Cyclone Intensity Change in the West Pacific. In *Proceedings of the 21st Conference on Hurricanes and Tropical Meteorology*, pages 94–96, Miami, Florida, 1995. American Meteorological Society.

[97] R. Baxter. *Minimum Message Length Inference: Theory and Applications.* PhD thesis, School of Computer Science and Software Engineering, Monash University, Melbourne, Australia, 1997.

[98] R. Baxter and D. Dowe. Model Selection in Linear Regression using the MML Criterion. Technical Report 206, School of Computer Science and Software Engineering, Monash University, 1994.

[99] R. Baxter and J.J. Oliver. MDL and MML: Similarities and Differences. Technical Report 207, School of Computer Science and Software Engineering, Monash University, 1994.

[100] R. Kistler, et al. The NCEP-NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bulletin of the American Meteorological Society*, 82:247–268, 2001.

[101] R. Shibata. An Optimal Selection of Regression Variables. *Biometrika*, 68(1):45–54, 1981.

[102] R.A. Anthes. Hurricane Model Experiments with a New Cumulus Parameterization Scheme. *Monthly Weather Review*, 105:513–528, 1990.

[103] R.H. Simpson. The Hurricane Disaster Potential Scale. *Weatherwise*, 27:169–186, 1974.

[104] R.M. Zehr and R. Phillips. EXPERT System for Tropical Cyclone Intensity Forecasts. In *Proceedings of the 7th Conference on Satellite Meteorology and Oceanography*, pages 71–74. American Meteorological Society, 1997.

[105] R.T. Merrill. An Experiment in Statistical Prediction of Tropical Cyclone Intensity Change. Technical Memo NWS-NHC-34, NOAA, 1987.

[106] R.T. Merrill. Environmental Influences on Hurricane Intensification. *Journal of Atmospheric Science*, 45:1678–1687, 1988.

[107] S. Muggleton. Bayesian Inductive Logic Programming. In *COLT-94*, pages 3–11. The Association for Computing Machinery, New York, 1994.

[108] T. Ryan. *Modern Regression Methods.* John Wiley & Sons, New York, 1997.

[109] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[110] V. Vapnik. Structure of Statistical Learning Theory. In A. Gammerman, editor, *Computational Learning And Probabilistic Reasoning*, pages 3–32. John Wiley & Sons, New York, 1996.

[111] V.F. Dvorak. Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery. *Monthly Weather Review*, 103:420–430, 1975.

[112] G.R. Walker. Tropical cyclones – the potential for insurance loss. *The Insurance Record*, pages 514–523, 1989.

[113] W.M. Gray. Global View of the Origin of Tropical Disturbances and Storms. *Monthly Weather Review*, 96:669–700, 1968.

[114] W.M. Gray. Atlantic Seasonal Hurricane Frequency: Part 1. El Nino and 30 mb Quasi-Biennial Oscillation Influences. *Monthly Weather Review*, 112:1649–1668, 1984.

[115] W.M. Gray. Atlantic Seasonal Hurricane Frequency: Part 2. Forecasting its Variability. *Monthly Weather Review*, 112:1669–1683, 1984.

[116] W.M. Gray and C.W. Landsea. African Rainfall as a Precursor of Hurricane-related Destruction on the US East Coast. *Bulletin American Meterological Society*, 73(9):1352–1364, September 1992.

[117] W.M. Gray and J.D. Sheaffer. El Niño Quasi-Biennial Oscillation Influence on Seasonal Atlantic Hurricane Activity. In M.H. Glantz, R.W. Katz, and N.

Nicholls, editor, *ENSO Teleconnection Linking World Wide Climate Anomalies: Scientific Basis and Societal Impact*, pages 257–283. Cambridge University Press, London, 1991.

[118] W.M. Gray, C.W. Landsea, P.W. Mielke and K.J. Berry. Predicting Atlantic Seasonal Hurricane Activity 6-11 months in Advance. *Weather and Forecasting*, 7:440–455, 1992.

[119] W.M. Gray, C.W. Landsea, P.W. Mielke and K.J. Berry. Predicting Atlantic Basin Seasonal Hurricane Activity by 1 August. *Weather and Forecasting*, 8:73–86, 1993.

[120] W.M. Gray, C.W. Landsea, P.W. Mielke and K.J. Berry. Predicting Atlantic Basin Seasonal Hurricane Activity by 1 June. *Weather and Forecasting*, 9:103–115, 1994.

[121] W.M. Gray, C.W. Landsea, P.W. Mielke, Jr. and K.J. Berry. Summary of 1995 Atlantic Tropical Cyclone Activity and Verification of Authors' Seasonal Predictors. http://typhoon.atmos.colostate.edu/forecasts/1995/ver.long.dec95/ennov95.html, 1995.

[122] World Meteorological Organization. *International Meteorological Vocabulary. WMO-No.182.* Geneva, 1992.

[123] W.R. Cotton and R.A. Anthes. *Storm and Cloud Dynamics.* Academic Press, Inc., 1989.

[124] Y. Kurihara, M.A. Bender and R.E. Tuleya. Performance Evaluation of the GFDL Hurricane Prediction System in the 1994 Hurricane Season. In *Proceedings of the 21st Conference on Hurricanes and Tropical Meteorology*, pages 41–43, Miami, Florida, 1995. American Meteorological Society.