Data-Driven Methods for the Assessment and Improvement of Forecasts

Suja M. Aboukhamseen

Department of Statistical Science

University College London

Thesis submitted for the degree of Doctor of Philosophy

Faculty of Science, University of London

September 2001

Abstract

This thesis uses data-driven techniques to analyse and assess both point and probability forecasts within a prequential framework. Point forecasts are assessed using recursive residuals. Examination of the properties of the recursive residual found them to be unique to this residual. Recursive residuals for the hidden state of HMM are also defined by taking the difference between the one step ahead forecast and the forecast's filtered update. The quality of forecasts generated from different models can be assessed by comparing the information content in their corresponding residuals. When faced with model misspecification it is shown how this residual can be modelled to correct this misspecification, thereby improving forecasts. It is also shown how the residual content can be used to judge the predictive sufficiency of alternative forecasting methods. Using the theory of probability forecasting, the technique of forecasting assessment by calibration is extended to HMM's to assess how well the one step ahead forecast is explained by its filtered update. A test statistic to test the empirical calibration of the forecasts is also defined and applied to the real world problem of CpG island detection in Human DNA sequences. The distribution of the test statistic is investigated using a prequential frame of reference and is found to be $N(0, 1)$. Calibration of HMMs is also examined using smoothed predictions and cross-validation forecasts. A test statistic is defined for this scenario and the its distribution is examined using a cross-validation frame of reference. A prequential estimation algorithm for HMMs is also developed.

Abstract

This thesis uses data-driven techniques to analyse and assess both point and probability forecasts within a prequential framework. Point forecasts are assessed using recursive residuals. Examination of the properties of the recursive residual found them to be unique to this residual. Recursive residuals for the hidden state of HMM are also defined by taking the difference between the one step ahead forecast and the forecast's filtered update. The quality of forecasts generated from different models can be assessed by comparing the information content in their corresponding residuals. When faced with model misspecification it is shown how this residual can be modelled to correct this misspecification, thereby improving forecasts. It is also shown how the residual content can be used to judge the predictive sufficiency of alternative forecasting methods. Using the theory of probability forecasting, the technique of forecasting assessment by calibration is extended to HMM's to assess how well the one step ahead forecast is explained by its filtered update. A test statistic to test the empirical calibration of the forecasts is also defined and applied to the real world problem of CpG island detection in Human DNA sequences. The distribution of the test statistic is investigated using a prequential frame of reference and is found to be $N(0,1)$. Calibration of HMMs is also examined using smoothed predictions and cross-validation forecasts. A test statistic is defined for this scenario and the its distribution is examined using a cross-validation frame of reference. A prequential estimation algorithm for HMMs is also developed.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Prof. A. P. Dawid for his guidance, patience, and time. For the opportunity of being his student, I feel truly privileged.

I would also like to thank my parents, and my brothers and sisters for their unwavering faith, unfaltering tolerance, and continuous encouragement.

To my fellow students and staff at the Department of Statistical Science, and to all the friends that I have made all over the UK, I would like to express my sincere thanks for their friendship, encouragement and help. A special thank you to my colleague Kai-Ming Chang who has shared the Ph.D. experience with me.

Lastly, I would also thank my sponsors, Kuwait University, for providing me with the opportunity to carry out this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis will address the problems of forecasting improvement when faced with model inadequacy and the problem of forecasting assessment in an information-restricted situation represented by hidden Markov Modelling.

The motivation behind this work draws heavily on the distinction between statistical models and the physical reality these models attempt to represent. A model is proposed in the hope of providing an explanation for a real world problem: a coherent statistical representation based on the modeller's subjective interpretation of the data generating system. Since the *true* mechanics of a data source are not known, the only link between the physical world and the statistical world used to represent it is the data observed. Based on this, the focus of this thesis is on the use of data-driven techniques in statistical analysis as a method of assessing and improving forecasts.

The purpose of statistical data analysis, as it is presented here, is to provide a valid explanation for a sequence of observations in the hope of producing the best possible forecasts for uncertain future outcomes of a real world problem. In turn, the forecasts themselves are assessed by their empirical success at explaining their forecast events. The statistical methods used

will be judged by the quality of the forecasts they generate at each intermediate point in time for the next observation, based on analysis of earlier outcomes. In this thesis, both point forecasts and probability forecasts are constructed and analysed within the prequential framework (Dawid, 1984), using the formalisms of probability forecasting (Dawid, 1986).

The prequential approach to data analysis (Dawid, 1984) is customised for the data-driven analysis of real world problems and the sources that emit them (Dawid, 1992). Therefore, the prequential approach, being an essentially data analytic approach, is adopted as the general theoretical framework for the concepts developed.

This thesis is dedicated to the development of new empirical methods for the analysis of data and the assessment of forecasts, and the extension of already existing methods of empirical assessment in various applications, hidden Markov models in particular (refer to section 1.1). In the case when point forecasts are made, the data are analysed by defining and using recursive residuals. In the case of probability forecasts, the field of probability forecasting has developed a rich literature of probability forecasting assessment techniques, the primary focus of which is calibration (explained in section 1.3). Here, these calibration techniques are extended to applications in hidden Markov models.

1.1 Hidden Markov Models

Hidden **M**arkov **M**odels (HMMs) are used to represent data in which there are sequences of two or more variables linked together by a causality rule. Specifically, the data sequence is believed to consist of noisy or obscured observations emitted from a higher level unobserved variable.

Researchers in many different fields have found treating a sequence of observations, as part of a causal formation such as the one described above to be very useful. Although these modelling techniques have been in use in various fields of engineering for some time now. interest was rekindled with Rabiner's 1989 article on HMMs. Since then HMM modelling has been applied in variety of different fields. In econometrics, Hamilton (1988, 1989, 1990, 1993), Harvey (1993), and McCullock and Tsay (1994) use *switching-state space models* to model data where the dynamics of an observed time series are considered to change according to a non-observed Markov chain. Hamilton (1989) and Kitagawa (1987) have both developed variations of filtering and smoothing algorithms for this sort of model. HMM modelling has also be applied extensively in the fields of computational biology (Krogh 1994, 1998) and speech and pattern recognition (Jaung and Rabiner,1991). A review of the use of HMMs in protein and DNA sequencing can be found in *Biological Sequence Analysis* by Durbin et al (1998). Churchill (1992) and Crowley et al (1997) also give examples of other HMM applications in genetics.

The unobserved variables can be modelled using any number of different statistical techniques. Linear dynamic models (Harrison and West, 1997), factor analysis and principle component analysis (Everitt, 1984, Hinton et al, 1995), and hidden Markov models (Rabiner, 1990) are examples of some of the various techniques used. Much work has been done showing the close correspondence of these different methods and how they relate to other statistical concepts. Roweis and Ghahramani (1998) show how these techniques can be expressed as variants of one general underlying model. In Smyth et al (1997) HMMs are explored within the general framework of probabilistic independence networks and Ghahramani (1997) shows how HMMs can be

viewed as examples of a dynamic Bayesian network.

Throughout the thesis, the term HMM will be used as a general term to encompass all these modelling techniques.

1.2 Prequential Analysis

The prequential approach to statistics (Dawid, 1984, 1996) is characterised by three main features:

1. The formalisation of the procedure involved in making forecasts for the future and assessing these methods on their empirical success at this task.

2. Offering suitable measures of uncertainty for unknown events where uncertainty is expressed in the form of a numerical probability.

3. Considering the sequential nature of the forecasting task.

The basis of this approach to statistics is the "appropriate manipulation of the data currently available so as to produce a specific probability distribution for the next observation" (Dawid, 1985) under the supposition that the data arrive in sequence. The prequential method can also accommodate situations which only require a point forecast or a decision problem. At any time i, a probability distribution, P_{i+1}, is formulated expressing uncertainty about the outcome of the next observation, A_{i+1}, in the light of the outcomes observed so far. In the case when a point forecast is needed, this formulation can be applied to solve the problem at hand. The term *prequential* refers to the combination of probability forecasting with sequential prediction.

The general framework of prequential analysis requires that statistical methods be judged solely by the forecasts they generate. With this in mind,

estimation is considered only in its capacity to improve the predictive performance of the prequential forecasts generated. The success of the estimation task is determined by the quality of the forecasts it helped to produce. The predictive performance is assessed through the comparison of the forecasts with their outcomes.

1.3 Probability forecasting and calibration

The development of probability forecasting as a theoretical discipline came about through the work of meteorologists in their use of probabilistic weather forecasting. The uncertain nature of the weather requires forecasters to quantify their degree of belief about the outcome of rain (**P**robability **o**f **P**recipitation) on any given day. Each day a weather forecaster issues a PoP for the next day using all the information available. Come the next day, the outcome of yesterday's uncertain event is now known. Adding this newly acquired information to the forecaster's information base, the forecaster again repeats the task of issuing a PoP for the following day. The demands placed upon weather forecasters in issuing daily PoPs has motivated much of the development of the theory and practice of probability forecasting. A detailed review of probability forecasting is given in Dawid (1983). Of primary interest here are the contributions made in the development of methods for the empirical assessment and comparison of a sequence of forecasts in the light of the outcomes of the forecast events.

Further developments in the theory of probability forecasting were made by Dawid (1983, 1986). The basis of the prequential approach to statistics, stimulated by the applications in weather forecasting, stems from the methodology of probability forecasting applications in meteorology.

In the prequential framework, the probability forecasts issued are constructed from what is called a *prequential forecasting system.* Let $\mathbf{a} = (a_1, a_2, \ldots)$ denote the sequential outcomes of uncertain events $\mathbf{A} = (A_1, A_2, \ldots)$ where $A_i = 1$ if the event occurs and $A_i = 0$ if the event does not occur $(i \geq 1)$. In light of the observed outcomes at time i, a probability, P_{i+1} must be assigned to the outcome of the next event A_{i+1}. Any method of constructing sequential forecasts for every i and $\mathbf{a}^{(i)} = (a_1, a_2, \ldots, a_i)$ in this way is called a forecasting system Dawid (1985). A prequential forecasting system is defined by a rule which associates a choice of P_{i+1} for every i and with any possible outcome $\mathbf{a}^{(i)} = (a_1, a_2, \ldots, a_i)$ of $\mathbf{A}^{(i)} = (A_1, A_2, \ldots, A_i)$.

Probability forecasts are assessed by determining how successful a forecasting system, F, which constructs the sequence of forecasts, is in explaining the sequence of outcomes. In the case when $A_i \in \{0,\ 1\}$ is binary, the criterion chosen to judge probability forecasts is calibration. In the meteorology literature calibration is referred to as *validity* (Miller, 1962) or reliability (Murphy, 1973). Lichtenstein et al (1982) give a review of the literature on the application of calibration in both meteorology and other fields.

A forecast is said to be well calibrated if, among the times for which a forecaster assigns a probability p for an event occurring, the long-run relative frequency of that event is also p. As discussed in Dawid (1986) and DeGroot and Fienberg (1982), a well calibrated forecasting system does not imply that the forecasts are good. This is because calibration assesses only one aspect of a forecasting system. The assessment of a probability forecast requires the blending of two separate tasks, sorting and labelling (Sanders, 1963, and Dawid, 1986). Sorting is the division of the sequence of events into disjoint subsequences such that all the events in any given subsequence are equally probable. The quality of the sorting process is referred to as resolution.

The second task, labelling, refers to the assigning of a numerical value to the common probability in each subsequence. Calibration only addresses a forecasting system's labelling ability.

In DeGroot and Fienberg (1982, 1983), sorting is referred to as refinement. They address the issue of inadequate well-calibrated forecasters and show how some well-calibrated forecasters can be deemed superior to others by comparing their refinement.

The forecasting assessment criterion of calibration is formalised in Dawid (1982) with the presentation of a general calibration theorem. Supposing that the forecasts arise sequentially from a joint probability distribution $\mathbf{P}$, this criterion requires that, for an arbitrarily selected test set (where the selection process is *admissible*), the difference between the proportion of times in which an event in question occurs and the average forecast probability for those times tends to zero, as the number of forecasts considered in the test set approaches infinity. Any forecast system F which meets this criterion for the selected test set is said to be completely calibrated, thereby deeming F an empirically valid explanation for the sequence of outcomes. A sequence of forecasts which satisfies this criterion of complete calibration has both perfect calibration and maximum attainable resolution. As such, it can be shown that if, by the complete calibration criterion, two forecasting systems, F^1 and F^2, are considered to be valid explanations for a sequence of outcomes, $\mathbf{a}$, with corresponding forecast sequences $\mathbf{p}^1$ and $\mathbf{p}^2$, then $p_i^1 - p_i^2 \longrightarrow 0$ as $i \longrightarrow \infty$ (Dawid, 1985). The calibration criterion satisfies the meta criteria laid out by Dawid (1985) for the selection of an appropriate criterion for the assessment of the empirical validity of a forecasting system.

More recent applications of the calibration criterion can be found in Kling and Bessler (1989) and Bessler and Kling (1991).

1.4 Outline of Thesis

In a situation when point forecasts are made, recursive residuals (Brown et al, 1975) are used as the data-driven apparatus of the forecasting assessment techniques. Hadi and Son (1989) examine some of the distinctive properties of the recursive residual. In Chapter 2. it is shown how the properties of a recursive residual are, in fact, unique to this residual alone, determining its structure.

Although recursive residuals are commonly used as a diagnostic mechanism (Harvey, 1990), Lumsdaine and Ng (1999) have shown that they can also be used to improve the performance of linear models by adding cumulative functions of recursive residuals to the regression equation. A variation of this concept is explored in Chapter 3 where the recursive residuals of a misspecified linear model are used in the formulation of a new model. Examination of the residuals of the new model show that the information lost through misspecification is regained by modelling the residuals in this way which produces the same results as a model that has been correctly specified. Chapter 3 also shows how recursive residuals can be defined and applied in Bayesian scenario.

Recursive residual applications are extended to the scope of HMMs in Chapter 4 where the residuals are defined and analysed for the unobserved state of various hidden Markov models. Using these definitions it is possible to show how a sufficient statistic for such models can be constructed and applied.

Probability forecasts are often assessed using calibration (Dawid, 1982). Dawid (1982, 1985) proposed the criterion of complete calibration for judging the empirical validity of probability forecasts. The basic calibration concepts and Dawid's complete calibration criterion are extended in Chapter 5 to ap-

plications in HMM configurations. Using the real world problem of CpG island detection in human DNA sequences, Chapter 6 illustrates how calibration can be used in the assessment of forecasts generated from HMMs, and presents a test statistic for testing the empirical validity of such forecasts as set out by the complete calibration criterion. The role of estimation in improving forecasts is examined in Chapter 7 using the DNA sequence data. A prequential online estimation method for HMMs is given and the calibration of forecasts constructed from parameter values estimated using both this method and the more common Baum-Welch (Rabiner, 1989) estimation algorithm are scrutinised. The calibration criterion is also examined outside the prequential framework in Chapter 8 using smoothed predictions and cross-validation forecast assessment.

1.5 Basic Concepts

Described below are two concepts that are used frequently throughout the thesis.

1.5.1 Martingales

Let $(X_1, X_2, X_3, \ldots)$ be a sequence with finite mean. The sequence is called a martingale if the conditional expectation of X_{i+1} given the values $X_1, X_2, \ldots, X_i$ is equal to X_i,

$$E\left(X_{i+1}|X_1, X_2, \ldots, X_i\right) = X_i. \tag{1.1}$$

A martingale can also be defined in the following, more general, way. Let β_i be a σ-field such that $\beta_i \subseteq \beta_{i+1}$. Then it is required that X_i be β_i-measurable for all i, and

$$E(X_{i+1}|\beta_i) = X_i. \tag{1.2}$$

In such a case, (X_i) is said to be a martingale adapted to the filtration (β_i). Then (1.1) is recovered when β_i is the σ-field generated by $(X_1, \ldots, X_i)$.

Let $S_1 = X_1$ and let $S_i = X_i - X_{i-1}$ for all $i \geq 2$. Then the constraints (1.1) and (1.2) can be written as

$$E\left(S_{i+1}|S_1, S_2, \ldots, S_i\right) = 0$$

and,

$$E\left(S_{i+1}|\beta_i\right) = 0, \qquad (i = 1, 2, \ldots),$$

and the sequence of variables (S_i), for $i = 2, 3, \ldots$, is said to form a martingale difference sequence with respect to (β_i).

Theorem 1.1 *Let the series (X_i) be a martingale difference sequence, so that $E(X_{i+1}|X_1, X_2, \ldots, X_i) = 0$ $(i \geq 1)$, and define $U_i = X_1 + X_2 + \cdots + X_i$. If c_i, $(i \geq 1)$, is a predictable sequence of random variables such that $c_1 < c_2 < \ldots \longrightarrow \infty$ and*

$$\sum_{k=1}^{\infty} c_k^{-2} E(X_k^2) < \infty \qquad (1.3)$$

hold with probability one, then with probability one

$$c_i^{-1} U_i \longrightarrow 0, \qquad (1.4)$$

and the variables

$$Y_i = \sum_{k=1}^{i} c_k^{-1} X_k \qquad (1.5)$$

converge to zero.

The proof can be found in Feller (1971, pg. 238). The sequence (Y_i) is a martingale sequence and $E(Y_i^2)$ is bounded by the series in (1.3). By the Martingale Convergence Theorem, the sequence (Y_i) converges with probability one and condition (1.3) holds true for each point in the sample space where (Y_i) converges. From Kronecker's lemma (Feller, 1971) the convergence of $\sum_{k=1}^{i} c_k^{-1} X_k$ implies that $c_i^{-1} \sum_{k=1}^{i} X_k \longrightarrow 0$ (i.e $c_i^{-1} S_i \longrightarrow 0$).

1.5.2 Conditional Independence

Let X, Y, and Z denote discrete random variables with a joint distribution P. The conditional distribution of X given $Y = y$, where y is any possible outcome of Y subject to $P(Y = y) \neq 0$, is denoted by $P(X|Y = y)$. Two random variables X and Y are said to be marginally independent, denoted by $X \perp\!\!\!\perp Y$, if

$$P(X|Y = y) = P(X),$$

for all possible values y for Y meaning that the probability of X is independent of the outcome of Y. X is said to be conditionally independent of Y given Z, denoted as $X \perp\!\!\!\perp Y|Z$ if, for any possible values y and z for Y and Z,

$$P(X|Y = y, Z = z) = P(X|Z = z).$$

All the definitions and properties in this section apply to both the discrete and the continuous case, but suppose for simplicity that X, Y, and Z are discrete random variables assuming any possible values x, y and z respectively. Let $a(x, z)$, $b(y, z)$ denote unspecified functions of (x, z) and (y, z) respectively. Then $X \perp\!\!\!\perp Y|Z$ if and only if any of the following equivalent conditions holds:

1. (a) $P(x|y, z) \equiv P(x|z)$ if $P(y, z) > 0$

   (b) $P(x|y, z)$ has the form $a(x, z)$ if $P(y, z) > 0$.

2. (a) $P(x, y|z) \equiv P(x|z)P(y|z)$ if $P(z) > 0$

   (b) $P(x, y|z)$ has the form $a(x, z)b(y, z)$ if $P(z) > 0$.

3. (a) $P(x, y, z) \equiv P(x|z)P(y|z)P(z)$

   (b) $P(x, y, z) \equiv P(x, z)P(y, z)/P(z)$ if $P(z) > 0$

   (c) $P(x, y, z)$ has the form $a(x, z)b(y, z)$.

Chapter 2

The Recursive Residual

2.1 Introduction

Residuals are the core of the forecasting assessment methods used in this thesis. Essentially, the residual is a linear function of the discrepancy, $y - \hat{y}$, between an observed value y and its prediction $\hat{y}$. Depending on the method of formulation of $\hat{y}$, and the linear transformation of $y - \hat{y}$ chosen, any number of different residuals can be produced. This flexibility enables the selection or formulation of residuals with certain desirable properties and characteristics.

The analysis carried out in this thesis is, for the most part, performed within a prequential framework. To remain with the limits of this framework the residual used must also be prequential in nature. The recursive residual (Brown et al, 1975) is one such residual. Described in greater detail in section 2.2, the formulation of the recursive residual is such that it provides a *fair* and sequential assessment of forecasting performance by using only data observed prior to the observation of event y in the formulation of $\hat{y}$. Such a formulation gives the recursive residual a very definite prequential quality making the residual an essential tool in prequential data analysis

(Dawid, 1985).

This chapter examines the various characteristics and properties of recursive residuals. It is shown how the properties of this transformation vector, expressed in terms of a residual transformation matrix, determine its components thereby, proving that the properties possessed by the recursive residual are unique to this residual.

After the recursive residual is briefly introducted in section 2.2, a broader family of residuals, the Linear Unbiased Scalar (LUS) residuals is described in section 2.3. The formulation of the LUS residuals and their corresponding residual transformation matrices paves the way for the introduction of the recursive residual transformation matrix. The structure and properties of this matrix are given in section 2.4, and section 2.5 shows how the properties of the matrix determine its elements.

2.2 Recursive Residuals

Consider the simple linear regression model

$$\mathbf{Y} = \mathbf{X}\theta + \epsilon, \tag{2.1}$$

where $\mathbf{Y}$ is a $n \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ is a $n \times p$ matrix of rank p consisting of observations corresponding to the p independent variables, θ is a $p \times 1$ vector of unknown parameters, and ϵ is the $n \times 1$ vector of unobserved disturbance terms with expectation zero and variance $\sigma^2\mathbf{I}$. Let $\mathbf{x}_i$ denote the $1 \times p$ vector holding the observations in the i^{th} row of $\mathbf{X}$, and $\mathbf{X}_i$ and $\mathbf{Y}_i$ be the leading $i \times p$ submatrix of $\mathbf{X}$ and $i \times 1$ subvector of $\mathbf{Y}$, respectively, containing rows 1 to i. It is assumed that $\mathbf{X}_p$ is of full rank, and it follows that $\mathbf{X}_i$, where $i > p$, is also of rank p. Assuming

that disturbance term is $N(0,\ \sigma^2)$, the recursive residual is defined as

$$w_i = \frac{y_i - \mathbf{x}_i \hat{\theta}_{i-1}}{\sqrt{1 + \mathbf{x}_i \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{x}_i^T}} \tag{2.2}$$

where y_i is the i^{th} observation of $\mathbf{Y}$ and $\hat{\theta}_{i-1} = \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{X}_{i-1}^T \mathbf{Y}_{i-1}$ is the least squares estimate of θ. It is important to note here that $\hat{\theta}_{i-1}$ is evaluated using only the data observed up to and including time $i-1$. The estimate for θ specified in this way gives the residual a prequential quality. The predictions, $\hat{y}_i = x_i \hat{\theta}_{i-1}$, generated using this formulation of $\hat{\theta}$ are also prequential which, in turn, makes the recursive residual a prequential diagnostic mechanism.

Brown, Durbin, and Evans (1975) introduced the recursive residual for the standard linear regression model as an alternative to the ordinary least squares residual which will be discussed in greater detail in Chapter 3. In analytical terms, the recursive residual is merely the prediction error resulting from the difference of y_i from its prequential prediction $\hat{y}_i = \mathbf{x}_i \hat{\theta}_{i-1}$. Consider:

$$\begin{aligned}
var\left[y_i - \hat{y}_i\right] &= var\left[\left(\mathbf{x}_i\theta + \epsilon_\mathbf{i}\right) - \left(\mathbf{x}_i \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{X}_{i-1}^T \mathbf{Y}_{i-1}\right)\right] \\
&= var\left[\epsilon_i\right] + \mathbf{x}_i \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{X}_{i-1}^T var\left[\mathbf{X}_{i-1}\theta + \epsilon_{i-1}\right] \mathbf{X}_{i-1} \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{x}_i^T \\
&= \sigma^2 + \mathbf{x}_i \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{x}_i^T \sigma^2 \\
&= \sigma^2 \left(1 + \mathbf{x}_i \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{x}_i^T\right),
\end{aligned}$$

since this is not constant, $y_i - \hat{y}_i$ is then standardised by dividing it by the square root of the coefficient of the σ^2, $1 + \mathbf{x}_i \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{x}_i^T$, to obtain a constant variance. Because of the standardisation, the recursive residual possess two very desirable properties: homoscedasticity and uncorrelated errors.

2.3 LUS Residuals

The recursive residual belongs to a family of residual known as **L**inear **U**nbiased with **S**calar Covariance Matrix (LUS) residuals. Introduced by Theil (1965, 1968, 1971), residuals within this family are characterised by a residual vector that is linear and unbiased. In addition, the residual vector is subject to the constraint that its covariance matrix be *scalar*, i.e. it can be written in the form $\sigma^2\mathbf{I}$. Using the above, the properties of residual transformation matrix, $\mathbf{C}$ of a LUS residual vector are:

1. The residual transformation matrix $\mathbf{C}$ is an $(n-p) \times n$ matrix not involving $\mathbf{Y}$. The rows of $\mathbf{C}$ are characteristic vectors of the matrix $(\mathbf{I} - \mathbf{H})$ corresponding to unit roots. The rows of $\mathbf{C}$ all have unit length and are also pairwise orthogonal.

2. $\mathbf{CX} = 0$, so that the expectation of the residual vector,

$$\begin{aligned} E[\mathbf{CY}] &= E[\mathbf{C}(\mathbf{X}\theta + \epsilon)] \\ &= 0, \end{aligned}$$

   is equal to the expectation of the disturbance term ensuring that the vector is unbiased.

3. $\mathbf{C}^T\mathbf{C} = (\mathbf{I} - \mathbf{H})$, where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$.

4. $\mathbf{CC}^T = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix so that

$$\begin{aligned} var[\mathbf{CY}] &= var[\mathbf{C}(\mathbf{X}\theta + \epsilon)] \\ &= var[\mathbf{C}\epsilon] \\ &= \sigma^2\mathbf{CC}^T \\ &= \sigma^2\mathbf{I} \end{aligned}$$

which gives the desired scalar covariance matrix.

Theil's notion of an unbiased residual vector is explained below. Consider the standard linear model in (2.1). The least squares residual vector is expressed as

$$\begin{aligned} \mathbf{e} &= \mathbf{Y_{n-p}} - \mathbf{X_{n-p}}\hat{\theta} \qquad (2.3) \\ &= \mathbf{CY}, \end{aligned}$$

where $\mathbf{C}$ is the residual transformation matrix not involving $\mathbf{Y}$, $\mathbf{e}$ is the $(n-p)\times 1$ residual vector, $\mathbf{Y_{n-p}}$ and $\mathbf{X_{n-p}}$ are the $(n-p)\times 1$ and $(n-p)\times p$ submatrices containing the last $(n-p)$ elements of $\mathbf{Y}$ and $\mathbf{X}$ respectively, and $\hat{\theta}$ is the least squares estimate of θ. Recall that the $n \times 1$ vector of disturbances, ϵ is unobserved. Expressed as

$$\epsilon = \mathbf{Y} - \mathbf{X}\theta,$$

it is easy to see that the residual vector offers itself as a natural approximation for at most $n-p$ components of ϵ. Consequently, if the residual vector is regarded as an estimate of ϵ, then it is unbiased if $E[\mathbf{e}] = E[\epsilon]$ which, in this case, is equal to zero.

The conditions of unbiasedness and scalar covariance imposed on $\mathbf{C}$ imply that only $n-p$ residuals can be found. Such conditions require that p of the disturbance terms be discarded. From $\mathbf{CX} = \mathbf{0}$ it is possible to see that the n columns of $\mathbf{C}$ are subject to p linear dependencies. As such p of the disturbance terms can be discarded without any loss of information.

2.4 Recursive Residual Transformation Matrix

Consider the unstandardised recursive residual,

$$\begin{aligned} \grave{r}_i &= y_i - \mathbf{x}_i\hat{\theta}_{i-1} \qquad (2.4) \\ &= y_i - \mathbf{x}_i\left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{X}_{i-1}^T\mathbf{Y}_{i-1}. \qquad (2.5) \end{aligned}$$

$i = p+1, \ldots, n$. The recursive residual vector, being a member of the LUS family of residuals (Hadi and Son, 1989), can be expressed in terms of a residual transformation matrix. Let $\mathbf{B}$ denote the $(n-p) \times n$ recursive residual transformation matrix for the unstandardised recursive residual vector $\grave{\mathbf{r}}$ satisfying

$$\grave{\mathbf{r}} = \mathbf{BY}.$$

Then, $\mathbf{B}$ is of the form

$$\begin{bmatrix}
-\mathbf{x}_{p+1}\left(X_p^T X_p\right)^{-1}\mathbf{X}_p^T & 1 & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\
-\mathbf{x}_{p+2}\left(X_{p+1}^T X_{p+1}\right)^{-1}\mathbf{X}_{p+1}^T & & 1 & 0 & \ldots & \ldots & \ldots & 0 \\
\vdots & & & \ddots & \ddots & & & \vdots \\
-\mathbf{x}_i\left(X_{i-1}^T X_{i-1}\right)^{-1}\mathbf{X}_{i-1}^T & & & & 1 & 0 & \ldots & 0 \\
\vdots & & & & & \ddots & \ddots & \vdots \\
-\mathbf{x}_{n-1}\left(X_{n-2}^T X_{n-2}\right)^{-1}\mathbf{X}_{n-2}^T & & & & & 1 & 0 & \\
-\mathbf{x}_n\left(X_{n-1}^T X_{n-1}\right)^{-1}\mathbf{X}_{n-1}^T & & & & & & 1 &
\end{bmatrix}.$$

Let b_{ij} denote the element of $\mathbf{B}$ on row i and column j where $i = (n-p), \ldots, n$ and $j = 1, \ldots, n$. The transformation matrix $\mathbf{B}$ has a lower triangular structure with all the $i < j$ elements equal to zero and $b_{ij} = 1$ for all $i = j$ elements of $\mathbf{B}$. The remaining $i > j$ elements are defined separately for each of the i rows of $\mathbf{B}$ by the $1 \times (n-i)$ vector $-\mathbf{x}_{n-i+1}\left(\mathbf{X}_{n-i}^T \mathbf{X}_{n-i}\right)^{-1}\mathbf{X}_{n-i}^T$.

The recursive residual transformation matrix possessess all the properties of a LUS transformation matrix discussed in section 2.3. It is, however, mentioned that $\mathbf{B}$ is not the *standardised* recursive residual transformation matrix. The rows of $\mathbf{B}$, although not normalised, remain pairwise orthogonal. Due to this, $\mathbf{BB}^T$ is a diagonal dispersion matrix with i^{th} diagonal element equal to the square of the standardising constant of the i^{th} row of $\mathbf{B}$. In order for $\mathbf{BB}^T = \mathbf{I}$, $\mathbf{B}$ must be standardised i.e. each row of $\mathbf{B}$ must be divided by the sum of the square root of its components. The specification of recursive

residual transformation matrix in its unstandardised form has the advantage of simplifying the algebraic manoeuvres in the next section.

2.5 Results

Let $\bar{\mathbf{B}}$ denote the standardised version of the recursive residual transformation matrix, $\mathbf{B}$, specified in the previous section. Then, in addition to possessing all the properties of a LUS residual transformation matrix listed in section 2.3, $\bar{\mathbf{B}}$ also has the following two properties:

1. $\bar{\mathbf{B}}$ is $(n-p) \times n$ matrix

2. $\bar{b}_{ij} = 0$ for $i < j$

which are specific to the matrix.

Theorem 2.1 *Given the above properties of the recursive residual transformation matrix, the formulation of the recursive residual transformation matrix is unique.*

Proof. Except for $i > j$ elements, the above properties of $\bar{\mathbf{B}}$ define all the characteristics and elements of $\bar{\mathbf{B}}$. The remaining $i > j$ elements can be solved for using a system of equations provided by $\bar{\mathbf{B}}\mathbf{X} = \mathbf{0}$ and $\bar{\mathbf{B}}\bar{\mathbf{B}}^T = \mathbf{I}$. Let $\bar{B}_{ij}$ denote a j-length row vector containing the first j elements of the i^{th} row of $\bar{\mathbf{B}}$. $\bar{B}_{i+1,n}$ is an arbitrary row in $\bar{\mathbf{B}}$. The value of $\bar{B}_{i+1,n}$ is determined by a set of p linearly independent equations from $\bar{B}_{i+1,n}\mathbf{X} = \mathbf{0}$ or equivalently from $\bar{B}_{i+1,i+1}\mathbf{X}_{i+1} = \mathbf{0}$ since $\bar{B}_{i+1,i+1}$ contains the first $i+1$ elements of $\bar{B}_{i+1,n}$ and the remaining elements are equal to zero. The remaining $i-p$ equations come from $\bar{B}_{i+1,i+1}[D_i] = \mathbf{0}$, where $D_i = \left[\bar{B}^T_{p+1,i+1}\ \bar{B}^T_{p+2,i+1}\ \cdots\ \bar{B}^T_{i,i+1}\right]$ is known.

Assume that $rank(X_p) = p$, then it follows that the rank of $\mathbf{X}_{i+1}$, where $(i+1) > p$, is also p. Assume, also, that the diagonal elements of $\bar{\mathbf{B}}$ are positive. Then $\bar{B}_{i+1,i+1}\,[D_i\ \mathbf{X}_{i+1}] = \mathbf{0}$ gives a homogenous set of i equations in i unknowns. For the solution to be unique it is sufficient that D_i be orthogonal to $\mathbf{X}_{i+1}$. D_i is nothing more then a subset of the first i rows of $\bar{\mathbf{B}}$, where only some of the trailing zeros have been left out. With this in mind, the orthogonality of D_i and $\mathbf{X}_{i+1}$ follows from $\bar{\mathbf{B}}\mathbf{X} = \mathbf{0}$.

The concept outlined above will now be implemented to derive the exact formulation of the recursive residual transformation vector. For simplification the matrix $\bar{\mathbf{B}}$ is replaced by $\mathbf{B}$. The first three rows of $\mathbf{B}$ will be solved for initially to establish the recursive structure of the evaluation scheme. Induction will then be used to generalise the results for the remaining rows of $\mathbf{B}$.

Let $\mathbf{B}_i$ denote the i^{th} row of $\mathbf{B}$ and let $B_{i,j}$ denote a j-length row vector containing the first j elements of the i^{th} row of $\mathbf{B}$. The matrix $\mathbf{X}$ is partitioned in the following way

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_p \\ \mathbf{x}_{p+1} \\ \mathbf{x}_{p+2} \\ \vdots \\ \mathbf{x}_n \end{bmatrix}.$$

Solving for $\mathbf{B}_{p+1}$ Expressed in terms of $B_{p+1,p}$, $\mathbf{B}_{p+1}$ has the form

$$\mathbf{B}_{p+1} = (B_{p+1,p},\ 1,\ 0,\ \ldots.\ 0)$$

Multiplying $\mathbf{B}_{p+1}$ by $\mathbf{X}$ yields the first row of $\mathbf{BX}$. Since $\mathbf{BX} = 0$, the result is a $1 \times p$ vector of zeros:

$$\mathbf{B}_{p+1}\mathbf{X} = B_{p+1,p}\mathbf{X}_p + \mathbf{x}_{p+1} = \mathbf{0}.$$

so that

$$B_{p+1,p}\mathbf{X}_p = -\mathbf{x}_{p+1}. \tag{2.6}$$

The $1 \times p$ vector in (2.6) is a system of p equations in p unknowns providing a unique solution for the unknown elements in $B_{p+1,p}$ since $\mathbf{X}_p$ is a $(p \times p)$ matrix of full rank. Using this system of equations solving for $B_{p+1,p}$ is trivial:

$$\begin{aligned} B_{p+1,p} &= -\mathbf{x}_{p+1}\left(\mathbf{X}_p\right)^{-1} \\ &= -\mathbf{x}_{p+1}\left(\mathbf{X}_p^T\mathbf{X}_p\right)^{-1}\mathbf{X}_p^T \end{aligned}$$

which is exactly as desired.

Solving for $\mathbf{B}_{p+2}$ The same procedure used to solve for $\mathbf{B}_{p+1}$ is followed here to solve for $\mathbf{B}_{p+2}$. First, $\mathbf{B}_{p+2}$ is multiplied by $\mathbf{X}$ which (from $B_{p+2}\mathbf{X} = \mathbf{0}$) yields:

$$B_{p+2,p}\mathbf{X}_p + b_{p+2,p+1}\mathbf{x}_{p+1} = -\mathbf{x}_{p+2}, \tag{2.7}$$

a system of p equations in $p+1$ unknowns. It is, however, possible to express one of the unknown component $b_{p+2,p+1}$ in terms of $B_{p+2,p}$ using the matrix $\mathbf{B}\mathbf{B}^T$. $\mathbf{B}\mathbf{B}^T$ is a diagonal matrix which means that any row of $\mathbf{B}$ multiplied by any other row of $\mathbf{B}$, other than itself, is equal to zero. Multiplying $\mathbf{B}_{p+2}$ by $\mathbf{B}_{p+1}^T$ gives

$$\mathbf{B}_{p+2}\mathbf{B}_{p+1}^T = B_{p+2,p}B_{p+1,p}^T + b_{p+2,p+1} = 0$$

which can be rearranged to give

$$b_{p+2,p+1} = -B_{p+2,p}B_{p+1,p}^T. \tag{2.8}$$

Let $C_{p+1} = -B_{p+1,p}^T$, so that

$$b_{p+2,p+1} = B_{p+2,p}C_{p+1}.$$

and note that $\mathbf{X}_p^T C_{p+1} = \mathbf{x}_{p+1}^T$:

$$\begin{aligned}
\mathbf{X}_p^T C_{p+1} &= -\mathbf{X}_p^T B_{p+1,p}^T \\
&= \mathbf{X}_p^T \left[\mathbf{x}_{p+1} \left(\mathbf{X}_p^T \mathbf{X}_p \right)^{-1} \mathbf{X}_p^T \right]^T \\
&= \mathbf{X}_p^T \mathbf{X}_p \left(\mathbf{X}_p^T \mathbf{X}_p \right)^{-1} \mathbf{x}_{p+1}^T \\
&= \mathbf{x}_{p+1}^T.
\end{aligned}$$

Although it is not yet apparent, the specification of C_{p+1} and future C_i's $(i = p+2, \ \ldots, \ n-1)$ plays an important role in clarifying the recursive nature of the algebra used in this proof.

Expressed in terms of $B_{p+2,p}$, $b_{p+2,p+1}$ can be substituted for in equation (2.7) giving a linear system of p equations in p unknowns:

$$\begin{aligned}
B_{p+2,p}\mathbf{X}_p + b_{p+2,p+1}\mathbf{x}_{p+1} &= -\mathbf{x}_{p+2} \\
B_{p+2,p}\mathbf{X}_p + B_{p+2,p}C_{p+1}\mathbf{x}_{p+1} &= -\mathbf{x}_{p+2} \\
B_{p+2,p}\left(\mathbf{X}_p + C_{p+1}\mathbf{x}_{p+1}\right) &= -\mathbf{x}_{p+2}
\end{aligned}$$

It is now possible to solve for $B_{p+2,p}$:

$$\begin{aligned}
B_{p+2,p} &= -\mathbf{x}_{p+2}\left(\mathbf{X}_p + C_{p+1}\mathbf{x}_{p+1}\right)^{-1} \qquad (2.9) \\
&= -\mathbf{x}_{p+2}\left(\mathbf{X}_p + C_{p+1}\mathbf{x}_{p+1}\right)^{-1}\left(\mathbf{X}_p^T\right)^{-1}\mathbf{X}_p^T \\
&= -\mathbf{x}_{p+2}\left(\mathbf{X}_p^T\mathbf{X}_p + \mathbf{X}_p^T C_{p+1}\mathbf{x}_{p+1}\right)^{-1}\mathbf{X}_p^T \\
&= -\mathbf{x}_{p+2}\left(\mathbf{X}_p^T\mathbf{X}_p + \mathbf{x}_{p+1}^T\mathbf{x}_{p+1}\right)^{-1}\mathbf{X}_p^T \\
&= -\mathbf{x}_{p+2}\left(\mathbf{X}_{p+1}^T\mathbf{X}_{p+1}\right)^{-1}\mathbf{X}_p^T.
\end{aligned}$$

Using the value of $B_{p+2,p}$ derived in equation (2.10), $b_{p+2,p+1}$ can be evaluated:

$$\begin{aligned}
b_{p+2,p+1} &= B_{p+2,p}C_{p+1} \qquad (2.10) \\
&= -\mathbf{x}_{p+2}\left(\mathbf{X}_{p+1}^T\mathbf{X}_{p+1}\right)^{-1}\mathbf{X}_p^T C_{p+1} \\
&= -\mathbf{x}_{p+2}\left(\mathbf{X}_{p+1}^T\mathbf{X}_{p+1}\right)^{-1}\mathbf{x}_{p+1}^T.
\end{aligned}$$

Together, $B_{p+2,p}$ and $b_{p+2,p+1}$ make up the vector $B_{p+2,p+1} = [B_{p+2,p}\ b_{p+2,p+1}]$, the unknown components of the row $\mathbf{B}_{p+1}$. Substituting $B_{p+2,p}$ and $b_{p+2,p+1}$ with the values obtained in (2.10) and (2.11) gives the complete composition of $B_{p+2,p+1}$,

$$\begin{aligned} B_{p+2,p+1} &= [B_{p+2,p}\ b_{p+2,p+1}] \\ &= \left[-\mathbf{x}_{p+2}\left(\mathbf{X}_{p+1}^T\mathbf{X}_{p+1}\right)^{-1}\mathbf{X}_p^T\ -\mathbf{x}_{p+2}\left(\mathbf{X}_{p+1}^T\mathbf{X}_{p+1}\right)^{-1}\mathbf{x}_{p+1}^T\right] \\ &= -\mathbf{x}_{p+2}\left(\mathbf{X}_{p+1}^T\mathbf{X}_{p+1}\right)^{-1}\left[\mathbf{X}_p^T\ \mathbf{x}_{p+1}^T\right] \\ &= -\mathbf{x}_{p+2}\left(\mathbf{X}_{p+1}^T\mathbf{X}_{p+1}\right)^{-1}\mathbf{X}_{p+1}^T. \end{aligned}$$

Solving for $\mathbf{B}_{p+3}$ First, p linear equations are obtained from $\mathbf{B}_{p+3}\mathbf{X} = \mathbf{0}$.

$$B_{p+3,p}\mathbf{X}_p + b_{p+3,p+1}\mathbf{x}_{p+1} + b_{p+3,p+2}\mathbf{x}_{p+2} = -\mathbf{x}_{p+3}. \tag{2.11}$$

The computations become more complicated as the difference between p and i, $i = (p+1),\ \ldots,\ n$, becomes larger. As in the case of $\mathbf{B}_{p+2}$, the scalars $b_{p+3,p+1}$ and $b_{p+3,p+2}$ are expressed in terms of $B_{p+3,p}$ using the composition of the matrix $\mathbf{B}\mathbf{B}^T$. The table below shows the expressions obtained for $b_{p+3,p+1}$ and $b_{p+3,p+2}$.

Rows Multiplied	Expression Obtained
$\mathbf{B}_{p+3}\mathbf{B}_{p+2}^T = 0$	$B_{p+3,p}B_{p+1,p}^T + b_{p+3,p+1} = 0$
	$\begin{aligned} b_{p+3,p+1} &= -B_{p+3,p}B_{p+1,p}^T \\ &= B_{p+3,p}C_{p+1} \end{aligned}$
$\mathbf{B}_{p+3}\mathbf{B}_{p+2} = \mathbf{0}$	$B_{p+3,p}B_{p+2,p}^T + b_{p+3,p+1}b_{p+2,p+1} + b_{p+3,p+2} = 0$
	$\begin{aligned} b_{p+3,p+2} &= -B_{p+3}B_{p+2,p}^T - b_{p+3,p+1}b_{p+2,p+1} \\ &= -B_{p+3}B_{p+2,p}^T - B_{p+3,p}B_{p+1,p}^T b_{p+2,p+1} \\ &= B_{p+3,p}\left(-B_{p+2,p}^T - C_{p+1}b_{p+2,p+1}\right) \\ &= B_{p+3,p}C_{p+2} \end{aligned}$

Where $C_{p+2} = -\left(B^T_{p+2,p} - C_{p+1}b_{p+2,p+1}\right)$ and $\mathbf{X}^T_p C_{p+2} = \mathbf{x}^T_{p+2}$:

$$\begin{aligned}
\mathbf{X}^T_p C_{p+2} &= -\mathbf{X}^T_p \left(B^T_{p+2,p} - C_{p+1}b_{p+2,p+1}\right) \\
&= \mathbf{X}^T_p \left(\mathbf{X}_p \left(\mathbf{X}^T_{p+1}\mathbf{X}_{p+1}\right)^{-1} \mathbf{x}^T_{p+2} + C_{p+1}\mathbf{x}_{p+1}\left(\mathbf{X}^T_{p-1}\mathbf{X}_{p-1}\right)^{-1}\mathbf{x}^T_{p-2}\right) \\
&= \mathbf{X}^T_p\mathbf{X}_p \left(\mathbf{X}^T_{p+1}\mathbf{X}_{p+1}\right)^{-1} \mathbf{x}^T_{p+2} + \mathbf{x}^T_{p+1}\mathbf{x}_{p+1}\left(\mathbf{X}^T_{p+1}\mathbf{X}_{p-1}\right)^{-1}\mathbf{x}^T_{p-2} \\
&= \left(\mathbf{X}^T_p\mathbf{X}_p + \mathbf{x}^T_{p+1}\mathbf{x}_{p+1}\right)\left(\mathbf{X}^T_{p+1}\mathbf{X}_{p+1}\right)^{-1}\mathbf{x}^T_{p+2} \\
&= \left(\mathbf{X}^T_{p+1}\mathbf{X}_{p+1}\right)\left(\mathbf{X}^T_{p+1}\mathbf{X}_{p+1}\right)^{-1}\mathbf{x}^T_{p+2} \\
&= \mathbf{x}^T_{p+2}
\end{aligned}$$

Using the above results, it is now possible to solve the system of equations in (2.11). By first substituting for the values of $b_{p+3,p+1}$ and $b_{p+3,p+2}$ in equation (2.11) so that

$$\begin{aligned}
B_{p+3,p}\mathbf{X}_p + B_{p+3,p}C_{p+1}\mathbf{x}_{p+1} + B_{p+3,p}C_{p+2}\mathbf{x}_{p+2} &= -\mathbf{x}_{p+3} \\
B_{p+3,p}\left(\mathbf{X}_p + C_{p+1}\mathbf{x}_{p+1} + C_{p+2}\mathbf{x}_{p+2}\right) &= -\mathbf{x}_{p+3}.
\end{aligned}$$

$B_{p+3,p}$ can be derived:

$$\begin{aligned}
B_{p+3,p} &= -\mathbf{x}_{p+3}\left(\mathbf{X}_p + C_{p+1}\mathbf{x}_{p+1} + C_{p+2}\mathbf{x}_{p+2}\right)^{-1} \\
&= -\mathbf{x}_{p+3}\left(\mathbf{X}_p + C_{p+1}\mathbf{x}_{p+1} + C_{p+2}\mathbf{x}_{p+2}\right)^{-1}\left(\mathbf{X}^T_p\right)^{-1}\mathbf{X}^T_p \\
&= -\mathbf{x}_{p+3}\left(\mathbf{X}^T_p\mathbf{X}_p + \mathbf{X}^T_p C_{p+1}\mathbf{x}_{p+1} + \mathbf{X}^T_p C_{p+2}\mathbf{x}_{p+2}\right)^{-1}\mathbf{X}^T_p \\
&= -\mathbf{x}_{p+3}\left(\mathbf{X}^T_p\mathbf{X}_p + \mathbf{x}^T_{p+1}\mathbf{x}_{p+1} + \mathbf{x}^T_{p+2}\mathbf{x}_{p+2}\right)^{-1}\mathbf{X}^T_p \\
&= -\mathbf{x}_{p+3}\left(\mathbf{X}^T_{p+2}\mathbf{X}_{p+2}\mathbf{x}_{p+2}\right)^{-1}\mathbf{X}^T_p.
\end{aligned}$$

Substituting the above expression for $B_{p+3,p}$ in $b_{p+3,p+1}$ and $b_{p+3,p+2}$ gives their values where

$$\begin{aligned}
b_{p+3,p+1} &= B_{p+3,p}C_{p+1} \\
&= -\mathbf{x}_{p+3}\left(\mathbf{X}^T_{p+2}\mathbf{X}_{p+2}\right)^{-1}\mathbf{X}^T_p C_{p+1} \\
&= -\mathbf{x}_{p+3}\left(\mathbf{X}^T_{p+2}\mathbf{X}_{p+2}\right)^{-1}\mathbf{x}^T_{p+1}
\end{aligned}$$

and

$$\begin{aligned} b_{p+3,p+2} &= B_{p+3,p}C_{p+2} \\ &= -\mathbf{x}_{p+3}\left(\mathbf{X}_{p+2}^T\mathbf{X}_{p+2}\right)^{-1}\mathbf{X}_p^T C_{p+2} \\ &= -\mathbf{x}_{p+3}\left(\mathbf{X}_{p+2}^T\mathbf{X}_{p+2}\right)^{-1}\mathbf{x}_{p+2}^T. \end{aligned}$$

Now that all the unknown elements have been determined, and the vector $B_{p+3,p+2}$ can now be specified as

$$\begin{aligned} B_{p+3,p+2} &= [B_{p+3,p}\ b_{p+3,p+1}\ b_{p+3,p+2}] \\ &= -\mathbf{x}_{p+3}\left(\mathbf{X}_{p+2}^T\mathbf{X}_{p+2}\right)^{-1}\left[\mathbf{X}_p^T\ \mathbf{x}_{p+1}^T\ \mathbf{x}_{p+2}^T\right] \\ &= -\mathbf{x}_{p+3}\left(\mathbf{X}_{p+2}^T\mathbf{X}_{p+2}\right)^{-1}\mathbf{X}_{p+2}^T \end{aligned}$$

Induction To ensure that the recursions displayed in the solution of the first three rows of $\mathbf{B}$ hold throughout, a proof by induction is used to generalise the results for any $i+1$ row of $\mathbf{B}$. It is assumed that the results obtained hold for row i of $\mathbf{B}$, so that

$$\begin{aligned} \mathbf{B}_i &= -\mathbf{x}_i\left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{X}_{i-1}^T, \\ b_{i,i-1} &= B_{i,p}C_{i-1}, \end{aligned}$$

where $C_{i-1} = -B_{i-1,p}^T + C_{p+1}b_{i-1,p+1} + C_{p+2}b_{i-1,p+2} + \cdots,\ + C_{i-2}b_{i-1,i-2}$. and $\mathbf{X}_p^T C_{i-1} = \mathbf{x}_{i-1}^T$.

The above equations are now used to solve for the $\mathbf{B}_{i+1}$. From $\mathbf{BX} = \mathbf{0}$

$$B_{i+1,p}\mathbf{X}_p + b_{i+1,p+1}\mathbf{x}_{p+1} + \cdots + b_{i+1,i}\mathbf{x}_i = -\mathbf{x}_{i+1} \tag{2.12}$$

is obtained and the scalar quantities obtained from $\mathbf{BB}^T$ are summarised in the table below.

Rows Multiplied	Expression Obtained
$\mathbf{B}_{i+1}\mathbf{B}_{p+1}$	$b_{i+1,p+1} = B_{i+1,p}C_{p+1}$
$\mathbf{B}_{i+1}\mathbf{B}_{p+2}$	$b_{i+1,p+2} = B_{i+1,p}C_{p+2}$
$\vdots$	$\vdots$
$\mathbf{B}_{i+1}\mathbf{B}_{i-1}$	$b_{i+1,i-1} = B_{i+1,p}C_{i-1}$
$\mathbf{B}_{i+1}\mathbf{B}_{i}$	$b_{i+1,i} = B_{i+1,p}C_{i}$

C_i in the above table is given by the following formula

$$C_i = -B_{i,p}^T + C_{p+1}b_{i,p+1} + \cdots + C_{i-1}b_{i,i-1},$$

and to simplify the computations $\mathbf{X}_p^T C_i$ is evaluated and found to be equal to $\mathbf{x}_i^T$.

At this point it is now possible to compute $B_{i+1,p}$. The scalar quantities, $b_{i+1,p+1}, \ldots, b_{i+1,i}$, are substituted for in (2.12) which gives

$$\begin{aligned} B_{i+1,p} &= -\mathbf{x}_{i+1}\left(\mathbf{X}_p + C_{p+1}\mathbf{x}_{p+1} + \cdots + C_i\mathbf{x}_i\right)^{-1} \\ &= -\mathbf{x}_{i+1}\left(\mathbf{X}_p + C_{p+1}\mathbf{x}_{p+1} + \cdots + C_i\mathbf{x}_i\right)^{-1}\left(\mathbf{X}_p^T\right)^{-1}\mathbf{X}_p^T \\ &= -\mathbf{x}_{i+1}\left(\mathbf{X}_p^T\mathbf{X}_p + \mathbf{x}_{p+1}^T\mathbf{x}_{p+1} + \cdots + \mathbf{x}_i^T\mathbf{x}_i\right)^{-1}\mathbf{X}_p^T \\ &= -\mathbf{x}_{i+1}\left(\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\mathbf{X}_p^T \end{aligned}$$

The scalar quantities, $b_{i+1,p+1}, \ldots, b_{i+1,i}$, can now be evaluated:

$$\begin{aligned} b_{i+1,p+1} &= B_{i+1,p}C_{p+1} \\ &= -\mathbf{x}_{i+1}\left(\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\mathbf{x}_{p+1}^T \\ b_{i+1,p+2} &= -\mathbf{x}_{i+1}\left(\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\mathbf{x}_{p+2}^T \\ &\vdots \\ b_{i+1,i} &= -\mathbf{x}_{i+1}\left(\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\mathbf{x}_i^T \end{aligned}$$

Putting the components together, the final composition for $B_{i+1,i}$ is $-\mathbf{x}_{i+1}\left(\mathbf{X}_i^T\mathbf{X}_i\right)^{-1}\mathbf{X}_i^T$. This shows that the recursions will hold for all i.

This concludes the proof: the properties of the recursive residual transformation matrix determine its components, making these components unique to these properties.

2.6 Discussion

The recursive residual is given a thorough introduction in this chapter. Belonging to the family of LUS residuals, the recursive residual vector can be expressed in terms of a residual transformation matrix independent of $\mathbf{Y}$. It is found that the properties possessed by the recursive residual transformation matrix determine the matrix's components, and are unique to this residual. The recursive residual has the added advantage of fitting well within the prequential framework making it an ideal data-driven tool for use in the analysis to come.

Chapter 3

Modelling Residuals

3.1 Introduction

This chapter explores methods of modelling residuals to correct model misspecification. Before the topic is examined in detail the purpose of model specification must first be clarified. The process that generates the data is not known. It is not the purpose of model specification to attempt to discover or even describe the generation process. The purpose of the model specified is merely to extract the main features of the data so as to provide reliable predictions.

Analysing residuals after designating a model which generates these predictions provides an accurate indication of the goodness of fit. Any failure the model might have in detecting variation in the data is embodied in the residuals. If the model is adequate it follows that the residuals will exhibit no apparent pattern, making them appear approximately random e.g. the sum of the residuals is approximately equal to zero (Harvey, 1990).

The association between random residuals and satisfactory model specification is discussed further in Dawid (1992). Here the concept of *approxi-*

mately random residuals is given a more precise interpretation using a probability integral transformation. Let R_i be the residual for the observation at time i, and let F_i denote the conditional distribution function of R_i given past observations under the model M, "then $\tilde{U} = (U_1, U_2, \ldots, U_n)$ where $U_i = F_i(R_i)$ should be independent uniform $[0, 1]$" if data arise from M. Different tests can be used to assess the uniformity and independence of $\tilde{U}$.

If the model does not account for a predominant characteristic of the data, then this deficiency on the model's part seeps through and embodies itself in the residuals. This predominant characteristic is now a systematic component of the residual, and, because of this, the residuals no longer fall into the *random* category (Harvey, 1990). In such a situation, the analysis is reassessed and a new model is specified using the insight gained from the residuals.

Once the residuals have detected model misspecification, the action taken need not involve the determination of an entirely new model. Dawid (1992) suggests that the residuals themselves be *massaged* into providing a better fit to the data. This entails looking at the residuals as observations of a new response variable and modelling it accordingly. By modelling the residuals obtained from a misspecified model it is possible to bring the lost information back into the analysis.

The correction of model specification through the modelling of residuals is the primary focus of this chapter. For a standard linear regression model, it will be shown how the residuals of a misspecified model are derived and modelled producing what will be termed the *residual model.* This residual model is also analysed and the residuals for this model, the *secondary residuals*, are obtained. Comparison of the secondary residual of the residual model and the residuals from the correct model will show that the two residuals

are the same, proving that such a modelling strategy can be used to correct misspecification and produce results identical to a correctly specified model.

The method of modelling residuals to correct model misspecification is applied using three different residuals. Section 3.2 looks at a simple linear regression model and how ordinary least square residuals can be used to correct an inadequate model. The resulting residual of the residual model is found to be the same as that of the correctly specified model. The same method is applied in Section 3.3 to recursive residuals which have a slightly more complex structure. Section 3.4 explores how residuals can be defined and used in a Bayesian framework.

3.2 Ordinary Least Squares Residual

For the linear model, one of the more commonly used residuals is the Ordinary Least Squares (OLS) residual. It is easily computed and with it the modeller can determine what important factors have been overlooked. The model under consideration is of the form

$$\mathbf{Y} = \mathbf{X}\theta + \epsilon, \tag{3.1}$$

$\epsilon \sim N(0, \sigma^2\mathbf{I})$, where $\mathbf{Y}$ is the $n \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ is the $n \times p$ matrix of observations corresponding to the p independent variables, θ is a $p \times 1$ vector of parameters and ϵ is a $n \times 1$ vector of error terms. Using

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.2}$$

as the least squares estimator for θ, the predicted value for $\mathbf{Y}$ is $\mathbf{X}\hat{\theta}$ and the OLS residual vector for this model is

$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$= \mathbf{Y} - \mathbf{X}\hat{\theta} \tag{3.3}$$

$$= (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T}$ and $(\mathbf{I} - \mathbf{H})\mathbf{X} = 0$. If the model is correctly specified, then the OLS residual reduces to

$$\mathbf{R} = (\mathbf{I} - \mathbf{H})\epsilon$$

showing that this residual is a linear combination of the true disturbance term, ϵ, with mean zero and variance $\sigma^2(\mathbf{I} - \mathbf{H})$. Under standard regularity conditions, $\mathbf{R}$ will also asymptotically converge in distribution to the true disturbance when n is large in comparison to p.

Despite its optimality, the OLS residual is inappropriate for testing the validity of the assumptions made about the disturbance term of a linear model. This assumption states that the disturbances are independent with mean zero and variance $\sigma^2\mathbf{I}$. In order to check this assumption, it is necessary for the residuals to mirror the properties of the disturbances (Harvey, 1990). Since the OLS residuals are both correlated and heteroscedastic, they are generally not viewed as a valid diagnostic for testing purposes (Harvey, 1993).

3.2.1 The correct model

Assume that the correctly specified model is of the form

$$\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}\phi + \epsilon, \tag{3.4}$$

where $\mathbf{Z}$ is $n \times q$ matrix of observations on q independent variables, ϕ is a $q \times 1$ vector of parameters, and the remaining components are the same as in model (3.1). Let $\hat{\theta}_C$ and $\hat{\phi}$ denote the estimates for θ and ϕ respectively.

under the correct model in (3.4). By writing model (3.4) in matrix notation,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \hat{\theta}_C \\ \hat{\phi} \end{bmatrix},$$

it is possible to obtain a system of two equations with two unknowns:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{bmatrix} \mathbf{Y} = \begin{bmatrix} \hat{\theta}_C \\ \hat{\phi} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}^T\mathbf{Y} \\ \mathbf{Z}^T\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} \end{bmatrix} \begin{bmatrix} \hat{\theta}_C \\ \hat{\phi} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}^T\mathbf{Y} \\ \mathbf{Z}^T\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{X}\hat{\theta}_C + \mathbf{X}^T\mathbf{Z}\hat{\phi} \\ \mathbf{Z}^T\mathbf{X}\hat{\theta}_C + \mathbf{Z}^T\mathbf{Z}\hat{\phi} \end{bmatrix}$$

from which the estimates of $\hat{\theta}$ and $\hat{\phi}$ can be derived:

$$\hat{\theta}_C = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{Z}\hat{\phi}\right) \tag{3.5}$$

$$\begin{aligned} \hat{\phi} &= \left(\mathbf{Z}^T\mathbf{Z} - \mathbf{Z}^T\mathbf{H}\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^T - \mathbf{Z}^T\mathbf{H}\right)\mathbf{Y} \\ &= \left(\mathbf{Z}^T(\mathbf{I} - \mathbf{H})\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}\right). \end{aligned} \tag{3.6}$$

Substituting (3.6) in (3.5) gives the expanded form of $\hat{\theta}_C$,

$$\hat{\theta}_C = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{Z}\left(\mathbf{Z}^T(\mathbf{I} - \mathbf{H})\mathbf{Z}\right)^{-1}\mathbf{Z}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}\right). \tag{3.7}$$

With the prediction for $\mathbf{Y}$, $\hat{\mathbf{Y}}$, equal to $\mathbf{X}\hat{\theta} + \mathbf{Z}\hat{\phi}$, the residual for this model is

$$\begin{aligned} \mathbf{R}_C &= \mathbf{Y} - \mathbf{X}\hat{\theta}_C - \mathbf{Z}\hat{\theta}. \\ &= \mathbf{Y} - \mathbf{X}\hat{\theta} - \mathbf{Z}\hat{\phi} + \mathbf{H}\mathbf{Z}\hat{\phi}, \end{aligned} \tag{3.8}$$

where $\hat{\theta}$ is given in equation (3.2) since from (3.5),

$$\hat{\theta}_C = \hat{\theta} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Z}\hat{\phi},$$

and so

$$\mathbf{X}\hat{\theta}_C = \mathbf{X}\hat{\theta} - \mathbf{H}\mathbf{Z}\hat{\phi}.$$

3.2.2 Correcting misspecified models

Assume that the correctly specified model is of the form given in equation (3.4). The model under consideration, however, is model (3.1). The misspecification takes the form of an omitted variable. The OLS residual in this case is

$$\mathbf{R}_1 = (\mathbf{I} - \mathbf{H})(\mathbf{X}\theta + \mathbf{Z}\phi + \epsilon).$$

Because, in this particular situation, the model is deficient the residual does not have the proper structure of an OLS residual. It is no longer a linear combination of just the disturbances, it is also a linear combination of the omitted variable $\mathbf{Z}$ and parameter ϕ:

$$\mathbf{R}_1 = (\mathbf{I} - \mathbf{H})\mathbf{Z}\phi + (\mathbf{I} - \mathbf{H})\epsilon. \tag{3.9}$$

The model in Equation (3.9) will be referred to as the residual model. Since $\mathbf{R}_1$ clearly has a linear regression structure, finding an adequate fit for $\mathbf{Y}$ can continue by regressing $\mathbf{R}_1$ on the omitted variable $\mathbf{Z}$.

It is worth noting that both $\mathbf{Z}$ and ϵ have been transformed. The distribution of the error term is now $N(0, \sigma^2(\mathbf{I} - \mathbf{H}))$. Later, it will be seen that this transformation compensates for the information not accounted for in the original fit. In the mean time, generalised least squares is used to find an estimate for ϕ.

First the following definitions are needed. Using a $\mathbf{QR}$ decomposition (Schott, 1997),

$$\mathbf{QX} = \begin{bmatrix} \mathbf{R}_* \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{R}_*$ is a $p \times p$ upper triangular matrix, let $\mathbf{Q} = [\mathbf{Q}_1^T \ \mathbf{Q}_2^T]^T$ be a $n \times n$ orthogonal matrix with the $p \times n$ matrix $\mathbf{Q}_1$ denoting the first p rows of $\mathbf{Q}$ and the $(n-p) \times n$ matrix $\mathbf{Q}_2$ denoting the remaining $(n-p)$ rows of $\mathbf{Q}$. Note

that from the orthogonality of $\mathbf{Q}$, $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_n$ and that $\mathbf{Q}_1\mathbf{Q}_1^T = \mathbf{I}_p$. $\mathbf{Q}_2\mathbf{Q}_2^T = \mathbf{I}_{(n-p)}$, and $\mathbf{Q}_1\mathbf{Q}_2^T = \mathbf{0}$. From the above definitions, the following relation is obtained

$$\begin{aligned}\mathbf{X} &= \mathbf{Q}^T\mathbf{R}_* \\ &= \mathbf{Q}_1^T\mathbf{R}_* + \mathbf{Q}_2^T\mathbf{0} \\ &= \mathbf{Q}_1^T\mathbf{R}_*,\end{aligned}$$

and the projection matrix $\mathbf{H}$ can be expressed as

$$\begin{aligned}\mathbf{H} &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ &= \mathbf{Q}_1^T\mathbf{R}_*\mathbf{R}_*^{-1}\left(\mathbf{R}_*^T\right)^{-1}\mathbf{R}_*^T\mathbf{Q}_1 \\ &= \mathbf{Q}_1^T\mathbf{Q}_1.\end{aligned}$$

Similarly, the projection matrix for the orthogonal complement of the vector space of $\mathbf{X}$, $(\mathbf{I} - \mathbf{H})$ can be expressed as $\mathbf{I} - \mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{Q}_2^T\mathbf{Q}_2$ (since $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}_1^T\mathbf{Q}_1 + \mathbf{Q}_2^T\mathbf{Q}_2$) and the following relation is obtained

$$\begin{aligned}(\mathbf{I} - \mathbf{H}) &= \mathbf{Q}_2^T\mathbf{Q}_2 \\ \mathbf{Q}_2(\mathbf{I} - \mathbf{H})\mathbf{Q}_2^T &= \mathbf{Q}_2\mathbf{Q}_2^T\mathbf{Q}_2\mathbf{Q}_2^T \\ \mathbf{Q}_2(\mathbf{I} - \mathbf{H})\mathbf{Q}_2^T &= \mathbf{I}.\end{aligned}$$

Premultipling both sides of the residual model by $\mathbf{Q}_2$ gives

$$\mathbf{Q}_2\mathbf{R}_1 = \mathbf{Q}_2(\mathbf{I} - \mathbf{H})\mathbf{Z}\phi + \mathbf{Q}_2(\mathbf{I} - \mathbf{H})\epsilon,$$

a transformed version of the residual where the covariance structure of the vector of disturbances is constant,

$$\begin{aligned}var\left[\mathbf{Q}_2(\mathbf{I} - \mathbf{H})\epsilon\right] &= \mathbf{Q}_2(\mathbf{I} - \mathbf{H})var\left[\epsilon\right](\mathbf{I} - \mathbf{H})\mathbf{Q}_2^T \\ &= \sigma^2\mathbf{Q}_2(\mathbf{I} - \mathbf{H})\mathbf{Q}_2^T \\ &= \sigma^2\mathbf{I}.\end{aligned}$$

The least squares estimate derived from regressing $\mathbf{Q}_2 R_1$ on $\mathbf{Q}_2(\mathbf{I}-\mathbf{H})\mathbf{Z}$ is the generalised least squares estimate for ϕ:

$$\begin{aligned}
\hat{\phi} &= \left(\mathbf{Z}^T(\mathbf{I}-\mathbf{H})^T\mathbf{Q}_2^T\mathbf{Q}_2(\mathbf{I}-\mathbf{H})\mathbf{Z}\right)^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{H})^T\mathbf{Q}_2^T\mathbf{Q}_2(\mathbf{I}-\mathbf{H})\mathbf{Y} \\
&= \left(\mathbf{Z}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\mathbf{Z}\right)^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\mathbf{Y} \\
&= \left(\mathbf{Z}^T(\mathbf{I}-\mathbf{H})\mathbf{Z}\right)^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}
\end{aligned}$$

For this particular model, using generalised least squares estimation produces an estimate for ϕ that is equivalent to the least squares estimate derived by regressing $\mathbf{R}_1$ on $(\mathbf{I}-\mathbf{H})\mathbf{Z}$:

$$\begin{aligned}
\hat{\phi} &= \left(\mathbf{Z}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})\mathbf{Z}\right)^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})\mathbf{Y} \\
&= \left(\mathbf{Z}^T(\mathbf{I}-\mathbf{H})\mathbf{Z}\right)^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}, \qquad (3.10)
\end{aligned}$$

assuming (incorrectly) that the the errors, $(\mathbf{I}-\mathbf{H})\epsilon$, are independent and identically distributed. Using this estimate to derive the residual for the latter stage regression yields

$$\begin{aligned}
\mathbf{R}_2 &= \mathbf{R}_1 - (\mathbf{I}-\mathbf{H})\mathbf{Z}\hat{\phi} \\
&= \mathbf{Y} - \mathbf{X}\hat{\theta} - (\mathbf{I}-\mathbf{H})\mathbf{Z}\hat{\phi} \\
&= \mathbf{Y} - \mathbf{X}\hat{\theta} - \mathbf{Z}\hat{\phi} + \mathbf{H}\mathbf{Z}\hat{\phi}. \qquad (3.11)
\end{aligned}$$

3.2.3 Results

The estimate for ϕ in (3.10) derived for the residual is identical to that derived from modelling the correct model in (3.4). The same, however, is not true for θ. The misspecified model in (3.1) and the correct model in (3.4) give two different estimates for θ. The estimate for the correct model, $\hat{\theta}_C$, uses information on both $\mathbf{X}$ and $\mathbf{Z}$. The information about $\mathbf{Z}$ comes into $\hat{\theta}_C$ in the form $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Z}\hat{\phi}$ or $\mathbf{H}\mathbf{Z}\hat{\phi}$.

In the misspecified model, however, the analysis, for some reason or another, has not taken $\mathbf{Z}$ into consideration. $\mathbf{Z}$ only comes into the analysis when the residuals for misspecified model pick up on $\mathbf{Z}$ as systematic component in the residual composition. When the residual model is specified, the residuals are regressed on a linear combination of $\mathbf{Z}$, specifically $(\mathbf{I} - \mathbf{H})\,\mathbf{Z}$. This transformation adds a new term, $\mathbf{HZ}$, to the regression which compensates for the information lost in the misspecification of the model. In the misspecified model, $\mathbf{HZ}\hat{\phi}$ is precisely the term missing from $\hat{\theta}$. The incorporation of $\mathbf{HZ}$ in the residual model gives $\mathbf{R}_2$ in equation (3.11) the same composition as $\mathbf{R}_C$ and the final residuals in both cases are the same.

3.3 Recursive residuals

Recursive residuals are introduced in detail in Chapter 2. This section uses the definitions and properties defined in Chapter 2 to reproduce the results in Section 3.2.

As in Section 3.2, the correct model is of the form

$$\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}\phi + \epsilon.$$

The misspecification begins by modelling $\mathbf{Y} = \mathbf{X}\theta + \epsilon$ and the analysis then picks up on the missing component by modelling the residual of the deficient model.

3.3.1 OLS and recursive residual relationships

Let the matrix $\mathbf{B}$ be the recursive residual transformation matrix defined in Chapter 2 so that $\mathbf{BY}$ is the recursive residual vector. The i^{th} row of $\mathbf{B}$ is

$$\mathbf{b}_i = \frac{1}{h_i}\left[\, -\mathbf{x}_i \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{X}_{i-1}^T \quad 1 \quad 0 \quad \dots \quad 0 \,\right]. \tag{3.12}$$

where $\mathbf{x}_i$ is the $1 \times p$ vector containing the values of the p independent variables of $\mathbf{X}$ at time i, and $\mathbf{X}_{i-1}$ is the $(i-1) \times p$ matrix containing the first $(i-1)$ rows of the matrix $\mathbf{X}$. Let h_i be the standardisation constant:

$$h_i = \left(1 + \mathbf{x}_i \left(\mathbf{X}_{i-1}^T \mathbf{X}_{i-1}\right)^{-1} \mathbf{x}_i^T\right)^{\frac{1}{2}}.$$

The properties of $\mathbf{B}$ can be used to relate the OLS residual vector to its recursive residual counterpart. The OLS residual vector for $\mathbf{Y} = \mathbf{X}\theta + \epsilon$ is $\mathbf{R}_1 = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. Using the properties $\mathbf{B}^T\mathbf{B} = (\mathbf{I} - \mathbf{H})$. $\mathbf{BX} = \mathbf{0}$ and $\mathbf{BB}^T = \mathbf{I}$ (Theil, 1975), the recursive residual vector, $\mathbf{R}_{1RR}$, for the above model can be derived by multiplying the recursive residual transformation vector by the OLS residual vector,

$$\begin{aligned} \mathbf{BY} &= \mathbf{BB}^T\mathbf{BY} \\ &= \mathbf{B}\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y} \\ &= \mathbf{BR}_1 \\ &= \mathbf{R}_{1RR}. \end{aligned} \tag{3.13}$$

3.3.2 The correct model

Specifying the recursive residual vector for the correct model first requires the specification of the estimates for the parameters θ, and ϕ. Let $\mathbf{Z}_i$, $\mathbf{B}_i$, and $\mathbf{Y}_i$ be the first i rows of the $\mathbf{Z}$, $\mathbf{B}$, and $\mathbf{Y}$ matrices respectively and let $\mathbf{H}_i$ be defined as $\mathbf{H}_i = \mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T$. By definition, the recursive residual requires that the estimates used in the prediction of y_i, the i^{th} observation $\mathbf{Y}$, contain only information available time $i-1$:

$$\begin{aligned} \hat{\phi}_{i-1} &= \left(\mathbf{Z}_{i-1}^T(\mathbf{I} - \mathbf{H}_{i-1})\mathbf{Z}_{i-1}\right)^{-1} \mathbf{Z}_{i-1}^T(\mathbf{I} - \mathbf{H}_{i-1})\mathbf{Y}_{i-1}. \\ \hat{\theta}_{i-1} &= \left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1} \left(\mathbf{X}_{i-1}^T\mathbf{Y}_{i-1} - \mathbf{X}_{i-1}^T\mathbf{Z}_{i-1}\hat{\phi}_{i-1}\right) \end{aligned}$$

$$\begin{aligned}
\hat{\theta}_{\mathbf{i-1}} &= \left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\left(\mathbf{X}_{i-1}^T\mathbf{Y}_{i-1} - \mathbf{X}_{i-1}^T\mathbf{Z}_{i-1}\left(\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1})\mathbf{Z}_{i-1}\right)^{-1}\right. \\
&\quad \left.\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1})\mathbf{Y}_{i-1}\right) \\
&= \left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{X}_{i-1}^T\mathbf{Y}_{i-1} - \left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{X}_{i-1}^T\mathbf{Z}_{i-1}\left(\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1})\mathbf{Z}_{i-1}\right)^{-1} \\
&\quad \mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1})\mathbf{Y}_{i-1},
\end{aligned}$$

for $i = (s+1), \ldots, n$, $s = max\{p, q\}$. The recursive residual at time i is

$$\begin{aligned}
r_{icRR} &= \frac{1}{g_i}\left(y_i - x_i\hat{\theta}_{i-1} - z_i\hat{\phi}_{i-1}\right) \\
&= \frac{y_i - \mathbf{d}_i\mathbf{Y}_{i-1} - \mathbf{e}_i\mathbf{Y}_{i-1} - \mathbf{f}_i\mathbf{Y}_{i-1}}{g_i} \\
&= \frac{1}{g_i}\left(y_i - \left(\mathbf{d}_i + \mathbf{e}_i + \mathbf{f}_i\right)\mathbf{Y}_{i-1}\right) \\
&= \frac{1}{g_i}\begin{bmatrix} -\left(\mathbf{d}_i + \mathbf{e}_i + \mathbf{f}_i\right) & 1 \end{bmatrix}\mathbf{Y}_i, \qquad (3.14)
\end{aligned}$$

where

$$\mathbf{d}_i = x_i\left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{X}_{i-1}^T \qquad (3.15)$$

$$\mathbf{e}_i = x_i\left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{X}_{i-1}^T\mathbf{Z}_{i-1}\left(\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1})\mathbf{Z}_{i-1}\right)^{-1}\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1}) \qquad (3.16)$$

$$\mathbf{f}_i = z_i\left(\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1})\mathbf{Z}_{i-1}\right)^{-1}\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1}), \qquad (3.17)$$

and g_i is the standardisation constant,

$$\begin{aligned}
g_i &= \left(1 + \mathbf{x}_i\left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{x}_i^T\right. \\
&\quad +\mathbf{x}_i\left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{X}_{i-1}^T\mathbf{Z}_{i-1}\left(\mathbf{Z}_{i-1}^T\left(\mathbf{I}-\mathbf{H}_{i-1}\right)\mathbf{Z}_{i-1}\right)^{-1}\mathbf{Z}_{i-1}^T\left(\mathbf{I}-\mathbf{H}_{i-1}\right) \\
&\quad \left.\mathbf{X}_{i-1}\left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{x}_i^T + \mathbf{z}_i\left(\mathbf{Z}_{i-1}^T\left(\mathbf{I}-\mathbf{H}_{i-1}\right)\mathbf{Z}_{i-1}\right)^{-1}\mathbf{z}_i^T\right)^{\frac{1}{2}}. \qquad (3.18)
\end{aligned}$$

Expressing the results in terms of a recursive residual transformation matrix can be done by specifying $\mathbf{R}_{cRR}$ as $(n-s)$-length vector of the r_{icRR} residuals and the $(n-s) \times n$ matrix $\mathbf{C}$ such that

$$\mathbf{R}_{cRR} = \mathbf{CY}.$$

$\mathbf{C}$ is the recursive residual transformation matrix where the i^{th} row of $\mathbf{C}$ is

$$\mathbf{c}_i = \frac{1}{g_i}\left[\begin{array}{ccccc} -(\mathbf{d}_i + \mathbf{e}_i + \mathbf{f}_i) & 1 & 0 & \cdots & 0 \end{array}\right]. \tag{3.19}$$

for $i = s+1, \ldots, n$.

3.3.3 Modelling recursive residuals

From equation (3.13) the following regression model is derived:

$$\begin{aligned} \mathbf{R}_{1RR} &= \mathbf{BY} \\ &= \mathbf{B}\left(\mathbf{X}\theta + \mathbf{Z}\phi + \epsilon\right) \\ &= \mathbf{BZ}\phi + \mathbf{B}\epsilon \end{aligned} \tag{3.20}$$

Unlike the OLS residual model in (3.9), the disturbance term in model (3.20) has a constant variance since $\mathbf{BB}^T = \mathbf{I}$.

To obtain the recursive residuals for model (3.20) the estimate of ϕ at any arbitrary point in time must contain only information available at that time. Therefore, a prediction made for y_{i+1} will use only the information available at time i. The estimate for ϕ derived by regressing $\mathbf{R}_{iRR}$ on $\mathbf{BZ}$ using only the information upto time i is then

$$\begin{aligned} \hat{\phi}_i &= \left(\mathbf{Z}_i^T\mathbf{B}_i^T\mathbf{B}_i\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i^T\mathbf{B}_i^T\mathbf{B}_i\mathbf{Y}_i \\ &= \left(\mathbf{Z}_i^T(\mathbf{I}-\mathbf{H}_i)\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i^T(\mathbf{I}-\mathbf{H}_i)\mathbf{Y}_i. \end{aligned}$$

Because the estimate for ϕ is different for each $i = q, \ldots, n-1$, the matrix notation of the secondary recursive residual for the model in (3.20) has a complicated structure.

For simplicity, the secondary recursive residual is first expressed as a scalar. Define $\mathbf{b}_i$ as a i-length row vector containing the non-zero elements

of the i^{th} row of $\mathbf{B}$ and the vector $\mathbf{M}_i$ of length i as

$$\mathbf{M}_i = \left[\begin{array}{cc} \left(\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1})\mathbf{Z}_{i-1}\right)^{-1}\mathbf{Z}_{i-1}^T(\mathbf{I}-\mathbf{H}_{i-1}) & \mathbf{0} \end{array} \right].$$

where again $i = s+1, \ldots, n$. The recursive residual at time i is

$$\begin{aligned} r_{i2RR} &= \frac{h_i}{g_i}\left(r_{i1RR} - \mathbf{b}_i\mathbf{Z}_i\hat{\phi}_{i-1}\right) \\ &= \frac{h_i}{g_i}\left(\mathbf{b}_i\mathbf{Y}_i - \mathbf{b}_i\mathbf{Z}_i\mathbf{M}_i\mathbf{Y}_i\right) \\ &= \frac{h_i}{g_i}\left(\mathbf{b}_i - \mathbf{b}_i\mathbf{Z}_i\mathbf{M}_i\right)\mathbf{Y}_i \end{aligned} \tag{3.21}$$

where g_i is the standardisation constant given in equation (3.18). To simplify, equation (3.21) is expressed in terms of the expressions $\mathbf{d}_i$, $\mathbf{e}_i$, and $\mathbf{f}_i$ defined in equations (3.16), (3.17), (3.17) respectively:

$$\begin{aligned} r_{i2RR} &= \frac{h_i}{g_i}\left[\frac{1}{h_i}\left(\begin{array}{cc} -\mathbf{d}_i & 1 \end{array}\right) - \frac{1}{h_i}\left(\begin{array}{cc} \mathbf{e}_i+\mathbf{f}_i & 0 \end{array}\right)\right]\mathbf{Y}_i \\ &= \frac{h_i}{g_i h_i}\left[\begin{array}{cc} -\left(\mathbf{d}_i+\mathbf{e}_i+\mathbf{f}_i\right) & 1 \end{array}\right]\mathbf{Y}_i \\ &= \frac{1}{g_i}\left(y_i - \left(\mathbf{d}_i+\mathbf{e}_i+\mathbf{f}_i\right)\mathbf{Y}_{i-1}\right), \end{aligned} \tag{3.22}$$

where $\mathbf{b}_i = \frac{1}{h_i}\left[\begin{array}{cc} -\mathbf{d}_i & 1 \end{array}\right]$ and $\mathbf{b}_i\mathbf{Z}_i\mathbf{M}_i = \frac{1}{h_i}\left[\begin{array}{cc} \mathbf{e}_i+\mathbf{f}_i & 0 \end{array}\right]$.

To present the recursive residuals in terms of a recursive residual transformation matrix the following matrices are used: let $\bar{\mathbf{B}}$ be a block diagonal matrix with $\mathbf{b}_i$ denoting the nonzero components of the i^{th} row of the matrix $\mathbf{B}$. The matrix has the following structure

$$\left[\begin{array}{cccc} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{b}_{p+1} & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{b}_n \end{array}\right],$$

and likewise the matrices $\bar{\mathbf{Z}}$, $\bar{\mathbf{M}}$, $\bar{\mathbf{Y}}$, and the matrix of recursive residuals. $\bar{\mathbf{R}}_{2RR}$ are defined as

$$\bar{\mathbf{Z}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_{q+1} & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_n \end{bmatrix}, \quad \bar{\mathbf{M}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_{q+1} & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M}_n \end{bmatrix},$$

$$\bar{\mathbf{Y}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{Y}_{s+1} & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Y}_n \end{bmatrix}$$

where $s = max\{q, p\}$. With matrix $\bar{\mathbf{\Phi}} = \bar{\mathbf{M}}\bar{\mathbf{Y}}$, a diagonal matrix with $\hat{\phi}_i$ as the i^{th} diagonal component, the recursive residual matrix can be written as

$$\begin{aligned} \bar{\mathbf{R}}_{2RR} &= \bar{\mathbf{B}}\bar{\mathbf{Y}} - \bar{\mathbf{B}}\bar{\mathbf{Z}}\bar{\mathbf{\Phi}} \\ &= \left(\bar{\mathbf{B}} - \bar{\mathbf{Z}}\bar{\mathbf{M}}\right)\bar{\mathbf{Y}}. \end{aligned}$$

Such a formulation produces the recursive residual matrix, $\bar{\mathbf{R}}_{2RR}$. as a diagonal matrix

$$\bar{\mathbf{R}}_{2RR} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 0 & r_{(s+1)2RR} & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & r_{n2RR} \end{bmatrix}$$

where r_{i2RR}, $i = (s+1), \ldots, n$, is given in equation (3.21). Multiplying $\bar{\mathbf{R}}_{2RR}$ by a $n \times 1$ unit vector will produce $\mathbf{R}_{2RR}$, the $n \times 1$ vector of recursive residuals for the residual model in (3.20).

3.3.4 Results

Although the residuals transformation structure for the appropriate model and the corrected misspecified model are not the same, the scalar results clearly show that the recursive residual itself is the same in both cases. The resulting scalar form of the residual in (3.14) is identical to that resulting from the residual model in (3.22).

3.4 Residual Analysis in Bayesian Models

This section examines how residuals can be defined and exploited in a Bayesian framework. For the purposes of this analysis, the definition of a recursive residual is generalised. Without regard to the exact form of recursive residual defined in Chapter 2, a recursive residual is any prediction error $y_i - \hat{y}_i$ (where $\hat{y}_i$, the prediction for y_i, contains only past information) standardised to acquire a constant variance. With this definition, the predictive distribution taken in a Bayesian context of a random variable Y is also a recursive residual.

3.4.1 The predictive distribution

Consider first the simple linear regression model of the form

$$Y_i = \phi X_i + \epsilon_i,$$

where Y_i is the i^{th} observation of the response variable Y, X_i is the i^{th} observation of the explanatory variable X, and ϵ_i is the independent and identically distributed disturbance term with mean zero and precision $\frac{1}{\sigma^2} = h$. Since the conditional distribution of Y_i is Normal, a Normal-gamma conjugate prior is used to perform Bayesian analysis on Y_i. Hence, the problem is

formulated in the following way:

$$\begin{aligned} Y_1, \ldots, Y_n | H,\ \Phi &\sim N\left(X_i\phi,\ h^{-1}\right) \\ \Phi | H &\sim N\left(0,\ (\lambda_0 h)^{-1}\right) \\ H &\sim \frac{\chi^2_{\nu_0}}{\nu_0 \tau_0^2} \end{aligned}$$

where $Y_1, \ldots, Y_n$ is an observation sequence from a Normal population. The resulting posterior distributions of Φ and H are

$$\begin{aligned} \Phi | H,\ data &\sim N\left(\mu_n,\ (\lambda_n h)^{-1}\right) \\ H |\ data &\sim \frac{\chi^2_{\nu_n}}{\nu_n \tau_n^2}. \end{aligned}$$

The hyperparameters μ_n, λ_n, ν_n, and τ_n^2 are given by

$$\mu_n = \frac{\sum_{i=1}^n Y_i X_i}{\lambda_0 + \sum_{i=1}^n X_i^2} \qquad \lambda_n = \lambda_0 + \sum_{i=1}^n X_i^2$$

$$\nu_n = \nu_0 + n \qquad \nu_n \tau_n^2 = \nu_0 \tau_0^2 + \sum_{i=1}^n Y_n^2 - \frac{\left(\sum_{i=1}^n Y_i X_i\right)^2}{\lambda_0 + \sum_{i=1}^n X_i^2}.$$

These results are straightforward and can be found in a number of data analysis books (i.e. Bromelling 1986, Zellner 1971).

To find the predictive distribution of a future observation Y_{n+1} a variable transformation technique is applied. Three variables U, V, and W are defined by:

$$\begin{aligned} U &= \left(Y_{n+1} - \phi X_{n+1}\right) h^{1/2} \\ V &= \left(\lambda_n h\right)^{1/2} \left(\Phi - \mu_n\right) \\ W &= \tau_n H^{1/2}, \end{aligned}$$

where

$$\begin{aligned} U | \Phi = \phi, H = h &\sim \sim N\left(0,\ 1\right) \\ V | H = h &\sim N\left(0,\ 1\right) \\ W &\sim \sqrt{\frac{\chi^2_{\nu_n}}{\nu_n}}. \end{aligned}$$

Since the distribution of U is unaffected by conditioning on Φ and H, U is independent of both W and V, and likewise since the distribution of V is unaffected by conditioning of H, V is independent of W. The following series of computations detail the evolution of U, V, and W through to the predictive distribution of Y_{n+1}. The computations begin with

$$U + \lambda_n^{-1/2} X_{n+1} V = H^{1/2} \left(Y_{n+1} - \mu_n X_{n+1} \right). \tag{3.23}$$

By the properties of U and V, and since X_{n+1} is given, the above equation gives a Normal distribution with mean zero and variance $\left(1 + \lambda_n^{-1} X_{n+1}^2\right)$. Dividing the left-hand side of (3.23) by its standard deviation gives

$$S = \frac{U + \lambda_n^{-1/2} X_{n+1} V}{\left(1 + \lambda_n^{-1} X_{n+1}^2\right)^{1/2}} = \frac{H^{1/2}(Y_{n+1} - \mu_n X_{n+1})}{\left(1 + \lambda_n^{-1} X_{n+1}^2\right)^{1/2}} \sim N(0,\ 1) \tag{3.24}$$

$$\left(1 + \lambda_n^{-1} X_{n+1}^2\right)^{1/2} S = H^{1/2} \left(Y_{n+1} - \mu_n X_{n+1} \right).$$

Note that the distribution of S is unaffected by conditioning on H and is, therefore, independent of W. Some further manipulation gives

$$\begin{aligned} Y_{n+1} &= \frac{\left(1 + \lambda_n^{-1} X_{n+1}^2\right)^{1/2} S \tau_n}{W} + \mu_n X_{n+1} \\ &\sim \frac{t_{\nu_n} \tau_n}{\left(1 + \lambda_n^{-1} X_{n+1}^2\right)^{-1/2}} + \mu_n X_{n+1}, \end{aligned} \tag{3.25}$$

since the distribution of $\frac{S}{W}$, is a t-distribution with ν_n degrees of freedom (DeGroot, 1989). If equation (3.25) is rearranged, then under the full Bayesian model the predictive distribution of Y_{n+1} given past observations of Y's can be written in the form of a residual,

$$\frac{Y_{n+1} - \mu_n X_{n+1}}{\left(1 + \lambda_n^{-1} X_{n+1}^2\right)^{1/2}} \sim t_{\nu_n}.$$

Substituting for the hyperparameters, the final form of this residual is

$$\frac{Y_{n+1} - \left(\frac{\sum_{i=1}^n Y_i X_i}{\lambda_0 + \sum_{i=1}^n X_i^2} \right) X_{n+1}}{\left[\frac{1}{\nu_0 + n} \left(\nu_0 \tau_0^2 + \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i X_i\right)^2}{\lambda_0 + \sum_{i=1}^n X_i^2} \right) \left(1 + (\lambda_0 + \sum_{i=1}^n X_i^2)^{-1} X_{n+1}^2 \right) \right]^{1/2}}.$$

This residual will be referred to as V_{n+1}.

3.4.2 The two stage approach

V_{n+1} is obtained by eliminating the dependence on both parameters, Φ, and H, from Y_{n+1} simultaneously. If, however, this method of derivation is divided into two stages, so that the parameters are eliminated one at a time, would the result be the same? What ensues is a thorough investigation of this question.

To begin the analysis, the problem is formulated in much the same way as in the previous situation. The three variables U, S and Y described in subsection 3.4.1 are used to derive equation (3.23) and equation (3.24). From equation (3.24) a new variable Z_{n+1} is formed:

$$\begin{aligned}
\left(1+\lambda_n^{-1}X_{n+1}^2\right)^{1/2} S &= H^{1/2}\left(Y_{n+1}-\mu_n X_{n+1}\right) \\
\frac{S}{H^{1/2}} &= \frac{Y_{n+1}-\mu_n X_{n+1}}{\left(1+\lambda_n^{-1}X_{n+1}^2\right)^{1/2}} = Z_{n+1} \\
Z_{n+1}|H &\sim N\left(0,\ h^{-1}\right).
\end{aligned}$$

The variable Z_{n+1} is a linear transformation of Y_{n+1} depending on past observations of Y, where through this transformation the parameter ϕ is eliminated. If the precision, H, is known then Z_{n+1} is the recursive residual for Y_{n+1}. This concludes the first stage in deriving the residual.

The next stage focuses on transforming Z_{n+1} into a variable which is independent of the precision H. This procedure requires the posterior distribution of $H|Z_1, \ldots, Z_n$. Consequently a new set of random variables, $Z_1, \ldots, Z_n$, are introduced to the analysis taking the place of the Y sequence.

The new sample $Z_1, \ldots, Z_n$ is created by transforming the original sequence of observations $Y_1, \ldots, Y_n$ by using stage one of the analysis described in the beginning of this section. In essence, this is the replacement of original

3.4.3 Confirmation

The two methods give two different formulations for the residual of Y_{n+1}, W_{n+1}, and V_{n+1}. Conditional on past observations, both Y_{n+1} and W_{n+1} are linear functions of Y_{n+1} and, independent of past observations of Y, both Y_{n+1} and W_{n+1} have a t-distribution with ν_n degrees of freedom. The question that still remains is whether W_{n+1} and V_{n+1} are equivalent. A visual examination of the two quantities shows that they are almost identical except for the summation terms in the denominators of each. Therefore, in order to prove that they are equal, it must be shown that the quantity

$$\sum_{i=1}^{n} \left(\frac{Y_i - \left(\frac{\sum_{k=1}^{i-1} Y_k X_k}{\lambda_0 + \sum_{k=1}^{i-1} X_k^2} \right) X_i}{\left(1 + \left(\lambda_0 + \sum_{k=1}^{i-1} X_k^2 \right)^{-1} X_i^2 \right)} \right)^2 \tag{3.26}$$

in W_{n+1} is equivalent to

$$\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i X_i \right)^2}{\lambda_0 + \sum_{i=1}^{n} X_i^2} \tag{3.27}$$

in V_{n+1}.

The nested summations in both (3.26) and (3.27) make this an exceedingly difficult task. Therefore, to avoid the cumbersome algebra involved, an algebraic technique is devised to prove that the two quantities are the same. This technique makes use of the fact that any variable, Y, can be represented by only one probability density function. Hence, if the pdf of Y is found by two different methods, then those two different methods produce the same pdf.

The pdf of interest here, is $f(Y_1, \ldots, Y_n|H)$ where $Y_i = \phi X_i + \epsilon_i$, and H is the unknown precision. This pdf can be found by either of the following two methods. In the first method, $f(Y_1, \ldots, Y_n|H)$ can be found directly by multiplying the likelihood of Y by the prior density $\pi(\phi|\ h)$ and integrating

observations by their residuals had the precision been known. This makes it possible to perform a posterior analysis on the unknown precision H. Given the joint probability distribution and the prior,

$$\begin{aligned} Z_1, \ldots, Z_n|H &\sim N\left(0,\ H^{-1/2}\right) \\ H &\sim \frac{\chi^2_{\nu_0}}{\nu_0\tau_0^2}, \end{aligned}$$

the posterior distribution of $H|Z_1, \ldots, Z_n$ then becomes

$$H|Z_1, \ldots, Z_n \sim \frac{\chi^2_{\nu_n}}{\nu_n\tau^{*2}}$$

where

$$\nu_n = n + \nu_0 \quad \nu_n\tau_n^{*2} = \nu_0\tau_0^2 + \textstyle\sum_{i=1}^n Z_i^2.$$

It is now possible to formulate the predictive distribution of Z_{n+1}. Let

$$\begin{aligned} U &= H^{1/2}\left(Z_{n+1}\right) \\ V &= \tau_n^* H^{1/2} \end{aligned}$$

where

$$\begin{aligned} U|H &\sim N\left(0,\ 1\right) \\ V &\sim \sqrt{\frac{\chi^2_{\nu_n}}{\nu_n}}. \end{aligned}$$

Note that since the distribution of U does not involve H then $U \perp\!\!\!\perp V$. Then

$$\frac{U}{V} = \frac{Z_{n+1}}{\tau_n^*} \sim t_{\nu_n}.$$

This residual will be referred to as W_{n+1}. Substituting for Z_{n+1} and τ_n^* gives

$$\frac{Y_{n+1} - \left(\frac{\sum_{i=1}^n Y_i X_i}{\lambda_0 + \sum_{i=1}^n X_i^2}\right) X_{n+1}}{\left[\frac{1}{\nu_0+n}\left(\nu_0\tau_0^2 + \sum_{i=1}^n \left(\frac{Y_i - \left(\frac{\sum_{k=1}^{i-1} Y_k X_k}{\lambda_0 + \sum_{k=1}^{i-1} X_k^2}\right) X_i}{\left(1 + \left(\lambda_0 + \sum_{k=1}^{i-1} X_k^2\right)^{-1} X_i^2\right)}\right)^2\right)\left(1 + \left(\lambda_0 + \sum_{i=1}^n X_i^2\right)^{-1} X_{n+1}^2\right)\right]^{1/2}}.$$

W_{n+1} is the recursive residual brought about by breaking up the process into two separate stages.

with respect to ϕ producing the following results:

$$f\left(Y_1,\ \ldots,\ Y_n|h\right)$$

$$= \int f\left(Y_1,\ \ldots,\ Y_n|\phi,\ H\right)\pi\left(\phi,|\ h\right)d\phi$$

$$= \frac{1}{(2\pi)^{n/2}}\left[\frac{h^n\lambda_0}{\lambda_0+\sum_{i=1}^n X_i^2}\right]^{1/2} exp\left\{-\frac{h}{2}\left(\sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_iX_i\right)^2}{\lambda_0+\sum_{i=1}^n X_i^2}\right)\right\} \quad (3.28)$$

Otherwise, it is possible to obtain $f(Y_1,\ \ldots,\ Y_n|H)$ recursively from

$$f\left(Y_1|H\right)f\left(Y_2|Y_1,\ H\right)\cdots f\left(Y_n|Y_1,\ \ldots,\ Y_{n-1},h\right).$$

This gives the pdf in a different form,

$$f\left(Y_1,\ \ldots.\ Y_n|\ h\right)$$

$$= \frac{1}{(2\pi)^{n/2}}\left[\frac{h^n\lambda_0}{\lambda_0+\sum_{i=1}^n X_i^2}\right]^{1/2} exp\left\{-\frac{h}{2}\sum_{i=1}^n\left(\frac{Y_i-\left(\frac{\sum_{k=1}^{i-1}Y_kX_k}{\lambda_0+\sum_{k=1}^{i-1}X_k^2}\right)X_i}{\left(1+\left(\lambda_0+\sum_{k=1}^{i-1}X_k^2\right)^{-1}X_i^2\right)}\right)^2\right\} \quad (3.29)$$

Expressed in terms of (3.28) or (3.29), the two forms of the pdf are different only in the summation quantities found in the exponential terms. These same quantities in (3.28) and (3.29) are equal to (3.26) and (3.27) respectively. In this case, however, (3.28) and (3.29) correspond to the same pdf and therefore must be equivalent. Hence, it follows that that (3.26) and (3.27) are two equal quantities and consequently, V_{n+1} and W_{n+1} must also be equivalent.

3.4.4 Alternative justification

It can also be shown that V_{n+1} and W_{n+1} are equivalent by using a more general argument.

Let X and Y be two random variables such that $Y = g(X)$, where g is an increasing linear function. Assume that the two variables X and Y have the same distribution function F. Then if F is strictly increasing, the following two statements hold:

1. if $X > Y$ then, $F(X) > F(Y)$

2. if $X < Y$ then, $F(X) < F(Y)$

These two statements lead to the conclusion that $F(X) = F(Y)$ *if and only if* $X = Y$.

This argument readily applies to V_{n+1} and W_{n+1}. The one step method described in subsection 3.4.1 takes the distribution of $Y_{n+1}|\{Y_1, \ldots, Y_n, \Phi, H\}$ and uses the posterior distribution of Φ, and H to find the predictive distribution of Y_{n+1}. This marginal distribution conditional on its past is merely a linear transformation of Y_{n+1} conditional on its past and the parameters Φ, and H. From this linear transformation, the residual $V_{n+1} \perp\!\!\!\perp \{Y_1, \ldots, Y_n\} \sim t_{\nu_n}$ is formulated such that V_{n+1} is a positive linear function of Y_{n+1}.

The two step approach takes the conditional distribution of Y_{n+1} and clears away the mean. What this produces is a new variable Z_{n+1}, where Z_{n+1}, dependent only on its precision and its past, is a positive linear function of Y_{n+1}. The dependence on the precision is removed using another linear transformation giving the residual W_{n+1}, where $W_{n+1} \perp\!\!\!\perp \{Y_1, \ldots, Y_n\} \sim t_{\nu_n}$.

V_{n+1} and W_{n+1} have the same properties. Independent of the past V_{n+1} and W_{n+1} have the same t-distribution. And dependent on the past, V_{n+1} and W_{n+1} are both linear transformations of Y_{n+1}. W_{n+1} is then a linear function of V_{n+1} and consequently, by the argument above, $V_{n+1} \equiv W_{n+1}$.

3.5 Discussion

Once an initial model is put forth, the residuals can be examined to ensure its adequacy. This can be done either graphically, to ensure that there is no apparent trend, or by using a probability integral transformation to ensure that the probability integeral transformations of the residuals are idependent and Uniformly distributed. This chapter presents various methods of treating the problem of model inadequacy

In the Bayesian setting, the predictive distribution of a variable is given a new interpretation. The structure of the predictive distribution as illustrated in section 3.4 is the same as that of a recursive residual, and is, therefore, treated as such in the analysis. For the Normal linear model with unknown mean and precision it is shown how the residual can be derived using two alternative methods. The first method is a direct approach where the desired distribution is found by eliminating the model parameters simultaneously.

In the second method, the residual is derived by eliminating one parameter at a time. First, the mean is eliminated, which results in the specification of a new variable. This new variable can be interpreted as the residual for the misspecified Normal linear model with unknown mean and known precision. The misspecification in the Bayesian case comes in the form a misspecified prior distribution where the precision is wrongly assumed to be known. By eliminating the precision from the new variable the desired predictive distribution is obtained.

In both cases, the predictive distribution is the same. This assertion is proved both algebraically and with a more general justification argument. The results obtained demonstrate how the analysis can be divided into evolutionary stages without loss of information.

For the standard linear regression model, the result of modelling the resid-

uals of a misspecified model is the same as having had initially specified the correct model. In this case, misspecification often takes the form of an omitted exogenous variable. The correction method applied by the modelling of the misspecified model's residuals repossesses any information in the dependent variable of interest not accounted for in the model. The residuals are used to reintroduce any lost or neglected information back into the study to improve predictions.

Recall that the residual is a linear function of the variable of interest. This characteristic enables the accommodation of the techniques presented here to a wide range of linear models (i.e. time series models and more complex regression models).

The modelling of residuals redirects the analysis from an analysis that was essentially exogenous in nature to an endogenous analysis of the dependent variable. The analysis, stated in this way, requires no further knowledge of the independent variables affecting the variable of interest. The modelling of residuals in this way is a data driven approach that allows flexibility in terms of model specification. In the case when no obvious alternative to the misspecified model is present, such a correction scheme can be highly useful.

Chapter 4

Hidden Markov Models and Recursive Residuals

4.1 Introduction

The previous chapter presented different ways in which recursive residuals can be used and applied. It was shown how a data-driven utility such as the recursive residual could be used in multistage model development, and could also be used in the correction of model misspecification.

This chapter focuses on extending the concepts in Chapter 3 to more complex model formulations represented by a broad class of models referred to here as hidden Markov models.

Hidden Markov models (HMMs) provide an interesting scenario for the application of recursive residuals. A residual is generally regarded as a safety net that catches the information provided in the data but neglected by the model. In HMMs, however, the model accounts for information known to exist, but simply not observed. This, to some extent, reverses the typical use of a residual. The question hence changes from "What more can the data

tell me about my model" to "Can the model guide me in any way as to the nature of the *unobserved observations* unaccounted for in those observations available." Hence the aim of this chapter is to gain knowledge and insight about the nature of the data using the residuals.

To delve deeper into this question it is first necessary to define some of the basic structures used in this study. Section 4.2 will give a brief overview of the structure of HMMs. Section 4.3 defines the general residual structures used in the HMM applications. The residuals are then calculated for various models and the results of the applications are listed in section 4.4 and generalised in section 4.5. A model due to Hamilton (1989) is also examined in section 4.6. Section 4.7 presents a data compression technique for the observed series that results in no loss of information. The results in section 4.7 are analysed in greater detail in section 4.8 where a more general explanation is given.

4.2 Hidden Markov Models

A Markov model is stochastic process typically used to describe a system which at any time t is in one of a set of N distinct states. At any given time, the current state of the system, X_t, depends only on the outcome of the previous state, X_{t-1}, i.e. $P(X_t|X_{t-1}, X_{t-2}, \ldots, X_1) = P(X_t|X_{t-1})$. A HMM is an extension of the concept of Markov models. In a HMM the state sequence of the Markov Model is hidden and at each time point a noisy signal is emitted depending on the state of the system. In such a system it is this series of noisy signals that constitute the observation sequence. Rabiner (1989) describes the HMM as "a doubly embedded stochastic process with an underlying stochastic process that is not observed but hidden". The hidden process can only be observed indirectly through the noisy emissions

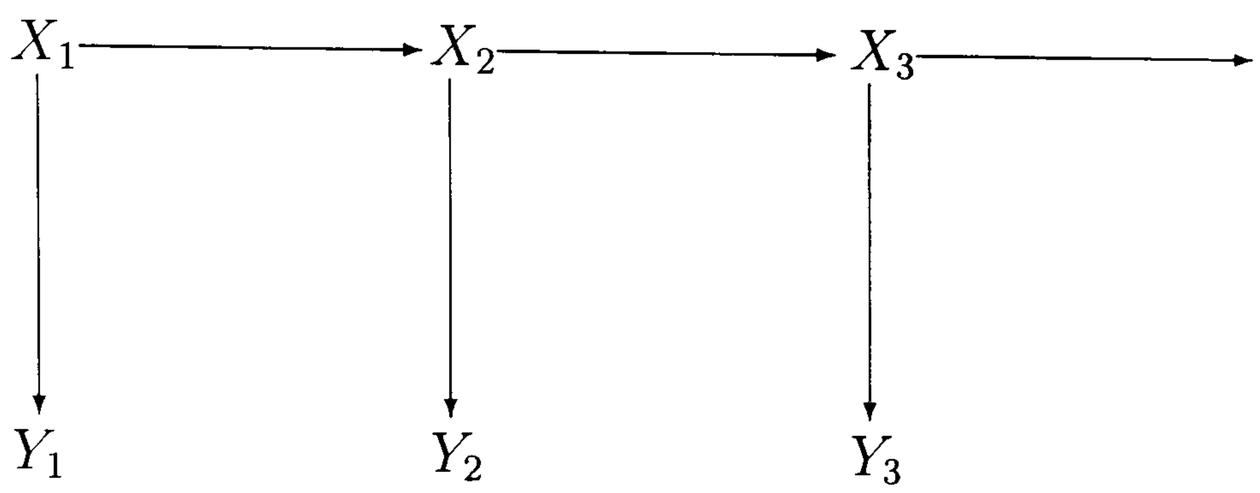

Figure 4.1: The causal structure of the HMM.

it generates that together constitute the sequence of outcomes.

Figure 4.1 displays a causal diagram typical of the structure of a first order HMM. Here, the X's represent the hidden system states where the state of the system is governed by a Markovian evolution process. The Y's are then the observed values which are noisy emissions of the state. The X's can only be observed indirectly through the observation of Y's. The diagram in Figure 4.1 describes a causal system conveying the conditional independence properties of a HMM:

$$X_i \perp\!\!\!\perp \{Y_1, \ldots, Y_{i-1}, X_1, \ldots, X_{i-2}\} \mid X_{i-1}$$
$$Y_i \perp\!\!\!\perp \{Y_1, \ldots, Y_{i-1}, X_1, \ldots, X_{i-1}\} \mid X_i.$$

The causality rules define a system where, given the value X_{i-1}, X_i is independent of past observations of Y and past X's. Likewise, given the value of the state X_i, Y_i is independent of past observations of Y and past X's. In short, given the present the future is independent of the past.

It is the analysis of the hidden state, X_i, that is the main concern of this study. If knowledge about the states can be gained, then this knowledge will provide much more meaningful insight about the data generating system then the raw observations.

The term HMM usually denotes a model with discrete state space and

with either a discrete or continuous observation sequence. Here, however, the term HMM is used to embody all variations of models with a hidden Markov structure emitting a noisy observation sequence. These models include factor analysis, principal component analysis, linear dynamic systems (state-space models), mixtures of Gaussian clusters and the like. Roweis and Ghahramani (1998) give a review of all these models and show how all these models are variations of single basic generative model which exploits the conditional independence structure of the data. The maximum likelihood parameter estimation algorithms for this class of models are also closely linked, and can all be expressed as modifications of the Expectation-Maximisation (EM) algorithm. Various authors have shown the basic equivalence of the different estimation techniques (Hinton, 1995, Roweis and Ghahramani, 1998).

Probabilistic independence networks (PINs) also present a very efficient framework for the representation of HMMs. The graphical representation of PINS allows for more flexibility and provides for a framework that permits the representation of more complex data formations. Smythe et al (1997) show how the forward-backward and Viterbi algorithms commonly used in HMMs are special cases of the more general propagation algorithms used in PINs. The graphical representation of the first order HMM is shown in Figure 4.2. For the first order HMM, the representation of the causal diagram in Figure 4.1 is the same as the graphical representation.

4.3 Recursive Residual Applications in HMMs

In the next section, a recursive residual (refer to Chapter 2 and Chapter 3) will be defined for the state variable X. Up to now, there has not been a formal definition for the state prediction error in the HMM literature. The

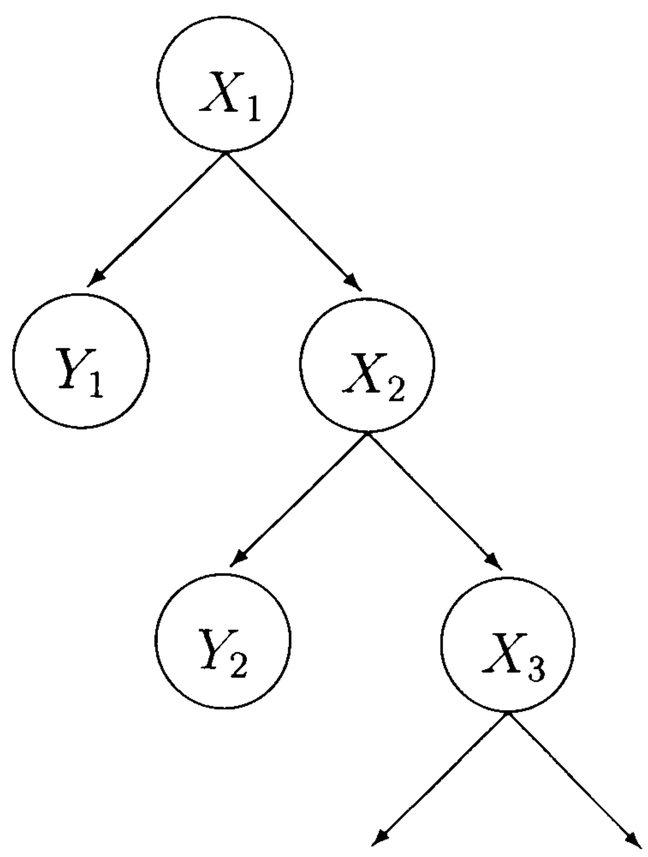

Figure 4.2: The DAG representation of an HMM.

specification of residuals in HMMs is limited to the observation sequence. Elliott et al (1995) describes how these residuals are prominent features of various online estimation schemes which attempt to reduce the observation prediction error.

In order to define a recursive residual for a HMM, further elaboration on the concepts of "realised value" and "prediction" are needed. In a data generating system such as an HMM, a realised value for the state as it is normally understood is unobservable. And because the state is never observed a prediction based on past observations of states is also infeasible since its evaluation would require the values of past observations of states. Therefore, what is suggested here is an approximation of both the predicted and the observed values based on what information is available, namely the sequence of Y's.

Consider, for example, a model of the form

$$X_{i+1} = \theta_i X_i + \epsilon_i, \quad \epsilon_i \sim N(0, W_i) \tag{4.1}$$

$$Y_i = X_i + \eta_i, \quad \eta_i \sim N(0, V_i) \tag{4.2}$$

where (4.1) is the system equation and (4.2) is the observation equation. The information available at time i is the sequence $D_i = (Y_1, Y_2, \ldots, Y_i)$. Given the known values of θ_i and V_i it is possible to find the distribution of $X_{i+1}|Y_1, \ldots, Y_i$. From this distribution, a point forecast can be defined as $\xi_{i+1} = E(X_{i+1}|D_i)$. Following the same line of logic, the closest possible "estimate" for the observation is the point forecast, $\zeta_{i+1} = E(X_{i+1}|D_{i+1})$, of the updated or filtered distribution for $X_{i+1}|Y_1, Y_2, \ldots, Y_{i+1}$, where $D_{i+1} = \{D_i, Y_{i+1}\}$. The recursive residual can now be defined as $Res_{i+1} = \zeta_{i+1} - \xi_{i+1}$.

If the recursive residuals as defined above are computed over a period of time to form a sequence, then this is a martingale difference sequence. This is proved by examining the conditional expectation of Res_{i+1}, $E(Res_{i+1}|Y_1, Y_2, \ldots, Y_i)$ which is equal to

$$\begin{aligned}
&E[E(X_{i+1}|Y_1, Y_2, \ldots, Y_{i+1}) - E(X_{i+1}|Y_1, Y_2, \ldots, Y_i)\,|Y_1, Y_2, \ldots, Y_i] \\
&= E(X_{i+1}|Y_1, Y_2, \ldots, Y_i) - E(X_{i+1}|Y_1, Y_2, \ldots, Y_i) \\
&= 0.
\end{aligned}$$

Hence, like all recursive residuals, the *Res* residuals defined above are uncorrelated. Homoscedacity can be achieved easily by standardising the residual by dividing by its standard deviation.

4.4 Results

All the results in this section were obtained using Bayesian updating formulas (West and Harrison, 1997) under the assumption that at time $i = 0$ the initial information for $X_0|D_0$ is $N(m_0, C_0)$.

The residual *Res* is first derived for the model described in (4.1) and (4.2). Note that, although the model is univariate, the results will remain

valid for the multivariate case. The results are summarised below:

Prior at time $i-1$

$$X_{i-1}|D_{i-1} \sim N(m_{i-1}, C_{i-1})$$

Forecast at time i

$$X_i|D_{i-1} \sim N(a_i, R_i),$$

where $a_i = \theta_i m_{i-1}$ and $R_i = \theta_i C_{i-1}\theta_i + W_i$

Posterior at time i

$$X_i|D_i \sim N(m_i, C_i),$$

where

$$\begin{aligned} m_i &= a_i + A_i e_i, & C_i &= R_i - A_i Q_i A_i^T, \\ e_i &= Y_i - a_i, & Q_i &= R_i + V_i, \\ A_i &= R_i (Q_i)^{-1}. \end{aligned}$$

Forecast for Y_i **at time** i

$$Y_i|D_{i-1} \sim N(a_i, R_i + V_i),$$

Residual at time i

$$\begin{aligned} Res_i &= m_i - a_i \\ &= A_i (Y_i - a_i). \end{aligned}$$

Essentially, *Res* is no more than a multiple of the residual based on the observed Y's, and, after standardisation, for the univariate case, would be identical to it. Although the aim of the analysis is to gain knowledge about the expression of the hidden series, the results basically reiterate that the information in the observed series contains all information available to make a statement about the hidden series.

4.5 Generalisation

The model analysed above is linear and Normally distributed. In this section, a nonlinear system with Normal disturbance terms is analysed to see if the results in the previous section also hold true for the recursive residual in this generalised case.

The model is defined as follows. Let $F_i(\cdot)$ and $g_i(\cdot)$ be the nonlinear regression function for the observation equation, and the nonlinear vector evolution function for the system equation respectively. Then the observation and system equations are of the form

$$X_{i+1} = g_i(X_i) + \epsilon_i, \qquad \epsilon_i \sim N(0, W_i) \tag{4.3}$$

$$Y_i = F_i(X_i) + \eta_i, \qquad \eta_i \sim N(0, V_i) \tag{4.4}$$

To define the residual, the distribution of $X_i|D_i$ must be specified. For nonlinear systems, however, this conditional distribution is most probably not Normal, and can be complex. Standard analysis of models such as these is essentially based on linearisation of the nonlinear structures, and approximating the non-Normal distribution by a Normal.

The easiest and most widely used linearisation method is the Extended Kalman Filter. The linearisation techniques described in this section are fairly common and can be found in a number of publications (West and Harrison , 1997, and Ghahramani, 1998).

The EKF uses a first order Taylor series expansion of the functions $g_i(\cdot)$ and $F_i(\cdot)$ to linearise (4.3) and (4.4). It must be assumed that the two functions are at least once differentiable. The EKF uses inductive reasoning to approximate the posterior distribution. At time $i-1$ it is assumed that historical information about the state, X_{i-1}, is approximated by Normal

posterior distribution:

$$(X_{i-1}|D_{i-1}) \sim N(m_{i-1}, C_{i-1}).$$

Using m_{i-1} as a point estimate for X_{i-1}, g_i can be expressed as a first order Taylor series expansion,

$$g_i \approx g_i(m_{i-1}) + G_i(X_{i-1} - m_{i-1}),$$

where

$$G_i = \left[\frac{dg_i(X_{i-1})}{dX}\right]_{X=m_{i-1}}$$

is known. The system equation can be approximated by a linear model.

$$X_i \approx g_i(m_{i-1}) + G_i(X_{i-1} - m_{i-1}) + \epsilon_i = h_i + G_i(X_{i-1}) + \epsilon_i.$$

where $h_i = g_i(m_{i-1}) - G_i m_{i-1}$. The distribution of the state is now $X_{i-1}|D_{i-1} \sim N(a_i, R_i)$ where $a_i = h_i + G_i m_{i-1} = g_i(m_{i-1})$ and $R_i = G_i C_{i-1} G_i^T + W_i$.

The observation equation can be dealt with in a similar fashion, giving the following approximation:

$$Y_i \approx f_i + \mathbf{F}_i(X_i - a_i) + \eta_i = (f_i - \mathbf{F}_i a_i) + \mathbf{F}_i X_i + \eta_i,$$

where $f_i = F_i(a_i)$ and

$$\mathbf{F}_i = \left[\frac{dF_i(X)}{dX}\right]_{X=a_i}.$$

The function $F_i(\cdot)$ is linearised about the point estimate for X_i, a_i.

It is now possible to approximate the predictive density of $Y_i|D_{i-1}$. By using this density, the desired posterior distribution for $X_i|D_i$ can be obtained. This distribution is Normal, $X_i|D_i \sim N(m_i, C_i)$, where the parameters statisfy

$$m_i = a_i + A_i e_i, \qquad C_i = R_i - A_i Q_i A_i^T,$$
$$A_i = R_i F_i Q_i^{-1}, \qquad e_i = Y_i - f_i.$$

The residual, $Res_i = m_i - a_i = A_i e_i$, remains a linear expression of the residual for the observed series. Hence the results obtained section 4.4 hold true for the more general class of HMMs.

4.6 The Hamilton Model

A second model is considered to see if a change in model structure will lend a different interpretation of the residual. In the previous model, the Y's merely represented the additive noise of the X process. We now consider a model similar to that used by Hamilton (1989) where the Y's are determined by a more complex procession of X's and Y's:

$$Y_i = \alpha X_i + Z_i \tag{4.5}$$

$$X_i = \theta_i X_{i-1} + \eta_i \qquad \eta_i \sim N(0, W_i) \tag{4.6}$$

$$Z_i = \phi_i Z_{i-1} + \epsilon_i \qquad \epsilon_i \sim N(0, V_i) \tag{4.7}$$

where Y_i, X_i, Z_i are scalar, α, θ_i, and ϕ_i are known and η_i and ϵ_i are independent and with known variances. Substituting (4.7) into (4.5) gives

$$Y_i = \alpha X_i + \phi_i (Y_{i-1} - \alpha X_{i-1}) + \epsilon_i. \tag{4.8}$$

This model is used in econometrics to model the dynamics of a time series characterised by episodes in which there is marked shift in regime. Figure 4.3 shows the causal diagram for the Hamilton model. Figure 4.4 shows the causal structure when equations (4.5) and (4.7) are replaced by equation (4.8).

The conditional independence structure that characterised the previous example does not hold for this model. Y_i is no longer conditionally independent of past observations and past states given X_i, but rather

$$Y_i \perp\!\!\!\perp \{Y_1, \ldots, Y_{i-2}, X_1, \ldots, X_{i-2}\} | \{X_i, X_{i-1}, Y_{i-1}\}.$$

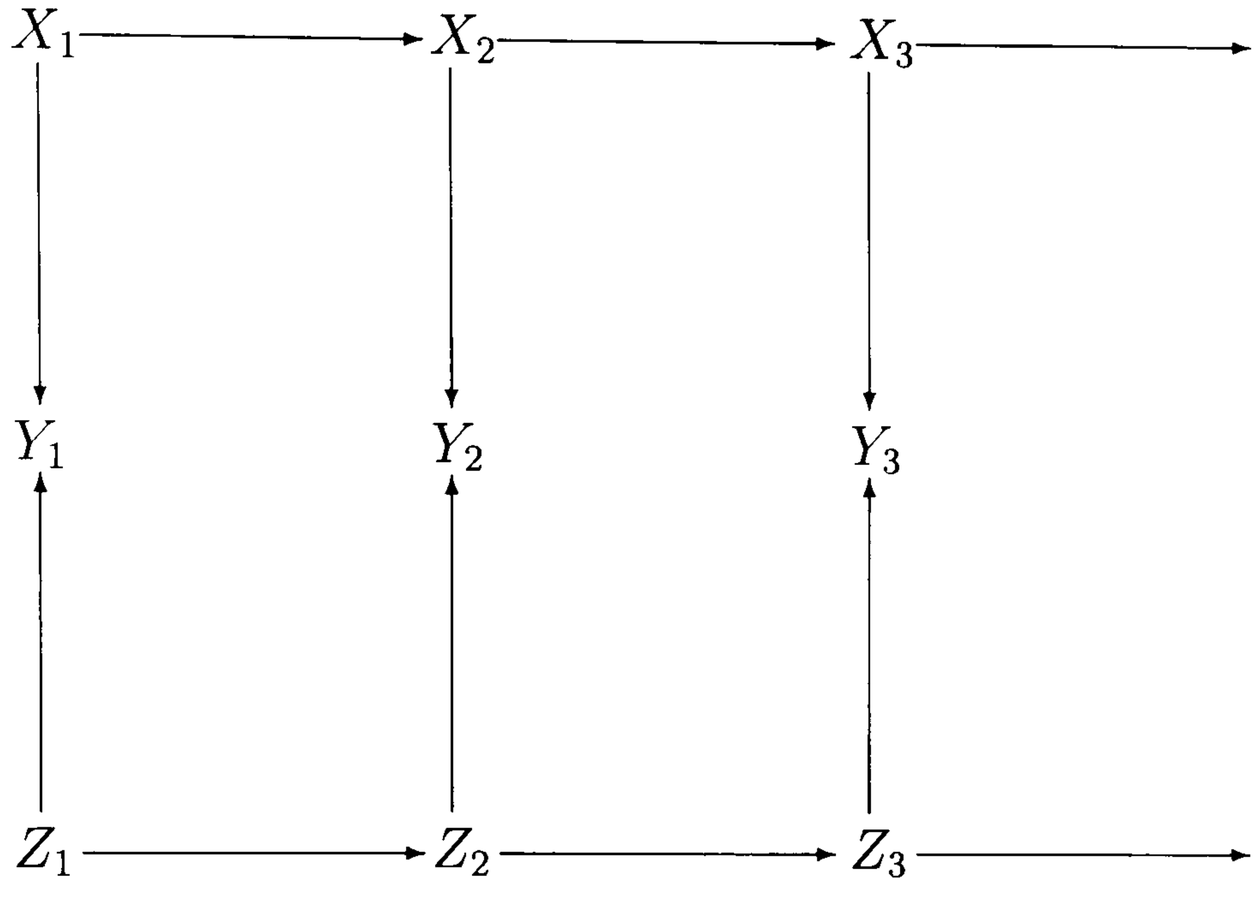

Figure 4.3: The causal diagram for the model as described in equations (4.5) - (4.7).

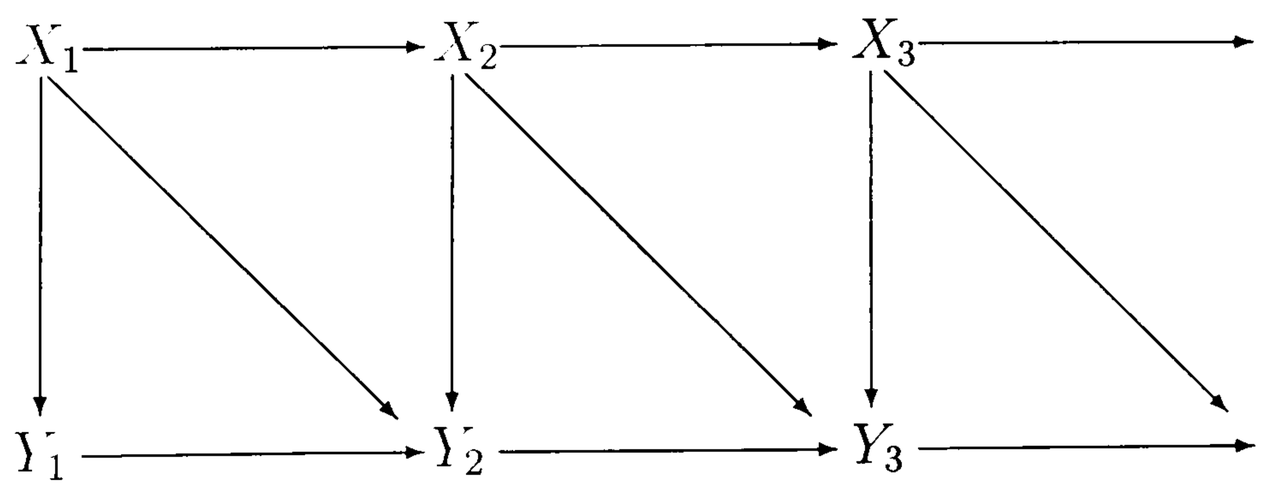

Figure 4.4: The causal structure of the HMM as described in equations (4.6) and (4.8).

The residual in this case remains a multiple of the residual for the observed sequence. The results are summarised below. Assuming at time $i = 0$ $X_0|D_0 \sim N(m_0, C_0)$:

Prior at time $i-1$

$$X_{i-1}|D_{i-1} \sim N(m_{i-1}, C_{i-1})$$

Forecast at time i

$$X_i|D_{i-1} \sim N(a_i, R_i),$$

where $a_i = \theta_i m_{i-1}$ and $R_i = \theta_i C_{i-1}\theta_i + W_i$

Forecast for Y_i

$$Y_i|D_{i-1} \sim N(f_i, Q_i)$$

where $f_i = \alpha a_i + \phi_i(Y_{i-1} - \alpha m_{i-1})$, and $Q_i = \alpha^2 R_i + \phi_i^2 \alpha^2 C_{i-1} + V_i$.

Posterior at time i

$$X_i|D_i \sim N(m_i, C_i),$$

where

$$m_i = a_i + A_i Q_i^{-1} e_i \qquad C_i = R_i - A_i Q_i^{-1} A_i^T$$
$$e_i = Y_i - \phi_i Y_{i-1} + \phi_i \alpha m_{i-1} - \alpha a_i \qquad A_i = \alpha R_i - \phi_i \alpha \theta_i C_{i-1}$$

Residual at time i

$$\begin{aligned} Res_i &= m_i - a_i \\ &= A_i(Y_i - \alpha a_i - \phi(\alpha Y_{i-1} - \alpha m_{i-1})) \\ &= A_i(Y_i - f_i). \end{aligned}$$

Res_i for the Hamilton model is also a function of the observation prediction error and once standardised would be equivalent to it.

4.7 Data Compression

Presented below is the analysis of a special case of the multivariate model where the Y observations are multivariate of order r, but the state variable X remains univariate. It is shown that by using generalised least squares to estimate X, it is possible to reduce the dimensionality of Y from r to one. Further analysis also shows that residuals derived from such a model are the same as those derived from the multivariate case, indicating no loss of information.

4.7.1 Compression of Y_i

Consider a model with an observation equation of the form

$$\mathbf{Y}_i = \mathbf{L}X_i + \eta_i \ \sim \ N\left(\mathbf{L}X_i,\ \mathbf{V}_i\right),$$

where $\mathbf{Y}_i$ is an $r \times 1$ vector of observations, $\mathbf{L}$ is an $r \times 1$ vector of some predefined constants, X_i is the univariate state of the system, and η_i is the $r \times 1$ vector of Normally distributed random errors with mean zero and covariance matrix $\mathbf{V}_i$. The stystem equation is the same as that in the univariate case (4.1).

The generalised least squares estimate of X_i, $\left(\mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{Y}_i$. can be regarded as a transformation that compresses $\mathbf{Y}_i$ into a one-dimensional form. This value will be denoted by Y_i^*. Substituting $\mathbf{Y}_i$ in Y_i^* gives

$$\begin{aligned} Y_i^* &= \left(\mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{L}\right)^{-1}\mathbf{L}\mathbf{V}_i^{-1}\left(\mathbf{L}^T X_i + \eta_i\right) \\ &= X_i + \left(\mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{V}_i^{-1}\eta_i, \end{aligned}$$

where Y_i^* decomposes into a univariate linear system equation of the form

$$Y_i^* = X_i + \gamma_i, \quad \gamma_i \ \sim \ N\left(0, \left(\mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{L}\right)\right).$$

The Bayesian updating formulas are again used to derive the residual for X_i. The table below summarises the hyperparameters for the posterior distribution based on the multivariate $\mathbf{Y}_i$ and its compressed counterpart Y_i^*.

Multivariate $r \times 1$ vector $\mathbf{Y}_i$	Univariate Y_i^*
Posterior at time $i-1$ for some mean m_{i-1} and variance c_{i-1}, $X_{i-1}\|D_{i-1} \sim N(m_{i-1},\ c_{i-1})$.	Posterior at time $i-1$ for some mean m_{i-1}^*, and variance c_{i-1}^*, $X_{i-1}\|D_{i-1} \sim N\left(m_{i-1}^*,\ c_{i-1}^*\right)$.
prior at time i, $X_i\|D_{i-1} \sim N(a_i,\ r_i)$ where $a_i = \theta_i m_{i-1}$ and $r_i = \theta_i^2 c_{i-1} + w_i$.	prior at time i $X_i\|D_{i-1} \sim N(a_i^*,\ r_i^*)$ where $a_i^* = \theta_i m_{i-1}^*$ and $r_i^* = \theta_i^2 c_{i-1}^* + w_i$.
Forecast for Y_i at time i, $Y_i\|D_{i-1} \sim N(a_i\mathbf{L},\ q_i)$, where $q_i = \mathbf{L}r_i\mathbf{L}^T + \mathbf{V}_i$.	Forecast for Y_i^* at time i, $Y_i^*\|D_{i-1} \sim N(a_i^*\mathbf{L},\ q_i^*)$, where $q_i^* = r^* + \left(\mathbf{L}^T\mathbf{V}_i^T\mathbf{L}\right)^{-1}$.
Update, $X_i\|D_i \sim N(m_i,\ c_i)$, where $m_i = a_i + A_i e_i$, $A_i = r_i\mathbf{L}^T q_i^{-1}$, $e_i = Y_i - a_i\mathbf{L}$, and $c_i = r_i - r_i\mathbf{L}^T q_i^{-1}\mathbf{L}r_i$.	Update, $X_i\|D_i \sim N(m_i^*,\ c_i^*)$, where $m_i^* = a_i^* + A_i^* e_i^*$, $A_i^* = r_i^* q_{-1}^{*-1}$ $e_i^* = \left(\mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{V}_i^{-1}Y_i - a_i^*$, and $c_i^* = r_i^* - r_i^* q_i^{*-1} r_{-1}^*$.
Residual, $Res_i = A_i e_i$	Residual $Res_i^* = A_i^* e_i^*$

$Res_i^* = Res_i$ can be proved using induction.

4.7.2 Proof

The state equation, $X_i = \theta_i X_{i-1} + \epsilon_i$ where ϵ_i is $N(0,\ w_i)$, is the same for both cases. Assume at time $i-1$ $X_{i-1}|D_{i-1} \sim N(m_{i-1},\ c_{i-1})$, and that

$m_{i-1} = m^*_{i-1}$, and $c_{i-1} = c^*_{i-1}$. From this, the forecast mean and variance for $X_i|D_i$ in the univariate cases, a^*_i, r^*_i are also the same as a_i. r_i. and hence $Res^*_{i-1} = Res_{i-1}$.

It now remains to be seen if the hyperparameters at time i. m_i. c_i. are equivalent to the their transformed counterparts m^*_i, c^*_i.

First the variance,

$$c_i = r_i - r_i \mathbf{L}^T \left(r_i \mathbf{L}\mathbf{L}^T + \mathbf{V}_i \right)^{-1} \mathbf{L} r_i,$$

is rewritten by expanding the inverse term (the formula can be found in Schott (1997)) and placing $\mathbf{L}^T$ and $\mathbf{L}$ inside the expansion producing

$$= r_i - r_i \left[\mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} - \mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} \left(r_i^{-1} + \mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} \right)^{-1} \mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} \right] r_i.$$

The value inside the brackets in the above equation is the expansion of $\left(r_i + \left(\mathbf{L}\mathbf{V}_i^{-1}\mathbf{L} \right)^{-1} \right)^{-1}$, and therefore c_i can be written as

$$c_i = r_i - r_i \left(r_i + \left(\mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} \right)^{-1} \right)^{-1} r_i,$$

which is equivalent to

$$c^*_i = r^*_i - r^*_i \left(r^*_i + \left(\mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} \right)^{-1} \right)^{-1} r^*_i.$$

The same technique is used to show that $m_i = m^*_i$,

$$m_i = a_i + \mathbf{L}^T \left(\mathbf{L} r_i \mathbf{L}^T + \mathbf{V}_i \right)^{-1} (Y_i - a_i \mathbf{L}).$$

The inverse is expanded and the $\mathbf{L}$ vector is multiplied in to give

$$m_i = a_i + r_i \left[\mathbf{L}^T V_i^{-1} - \mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} \left(r_i^{-1} + \mathbf{L}^T \mathbf{V}_i \mathbf{L} \right) \mathbf{L}^T \mathbf{V}_i^{-1} \right] (Y_i - a_i \mathbf{L}).$$

$\mathbf{L}^T \mathbf{V}_i^{-1}$ is then factored out and multiplied into $(Y_i - a_i \mathbf{L})$ so that m_i can be expressed in the following form

$$m_i = a_i + r_i \left[\mathbf{I} - \mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} \left(r_i^{-1} + \mathbf{L}^T \mathbf{V}_i \mathbf{L} \right)^{-1} \right] \left(\mathbf{L}^T \mathbf{V}_i^{-1} Y_i - a_i \mathbf{L}^T \mathbf{V}_i^{-1} \mathbf{L} \right).$$

m_i^* can also be expressed in the above form. Again the inverse of q_i^* is expanded to give

$$m_i^* = a_i + r_i \left[\mathbf{L}\mathbf{V}_i^{-1}\mathbf{L} - \mathbf{L}\mathbf{V}_i^{-1}\mathbf{L}\left(r_i^{-1} + \mathbf{L}\mathbf{V}_i^{-1}\mathbf{L}\right)^{-1}\mathbf{L}\mathbf{V}_i^{-1}\mathbf{L}\right]\left[\left(\mathbf{L}\mathbf{V}_i^{-1}\mathbf{L}\right)^{-1}\mathbf{L}\mathbf{V}_i^{-1}\mathbf{L} - a_i\right].$$

$\mathbf{L}\mathbf{V}_i^{-1}\mathbf{L}$ is then factored out of the expansion and multiplied into

$$\left[\left(\mathbf{L}\mathbf{V}_i^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{V}_i^{-1}Y_i - a_i\right]$$

giving an expression of m_i^* equivalent to m_i:

$$m_i^* = a_i + r_i \left[\mathbf{I} - \mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{L}\left(r_i^{-1} + \mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{L}\right)^{-1}\right]\left(\mathbf{L}^T\mathbf{V}_i^{-1}Y_i - \mathbf{L}^T\mathbf{V}_i^{-1}\mathbf{L}a_i\right).$$

It can now be deduced from $m_i^* = a_i^* + A_i^* e_i^*$, and $m_i = a_i + A_i e_i$ that the residuals $Res_i^* = A_i^* e_i^*$, and $Res_i = A_i e_i$ are also equivalent.

4.8 Sufficiency

The results for the special case of the linear dynamic model presented in section 4.7 can be given a much broader interpretation. The problem consists of a noisy array of multivariate observations originating from a hidden univariate source. The results show that the multivariate data, $\mathbf{Y}_i$, can be summarised using a univariate statistic, T_i, where T_i is a function of $\mathbf{Y}_i$, without loss of information. The statistic, T_i, summarises all the information in $\{Y_1, \ldots, Y_i\}$ about the hidden variable, X_i, such that the knowledge of the individual values of the Y's becomes irrelevant. T_i is thus a sufficient statistic for X_i with the following property:

$$Y_i \perp\!\!\!\perp X_i | T_i$$

The diagram in Figure 4.5 shows the causal structure of the model prior to the data compression. Figure 4.6 shows the causal diagram of the model

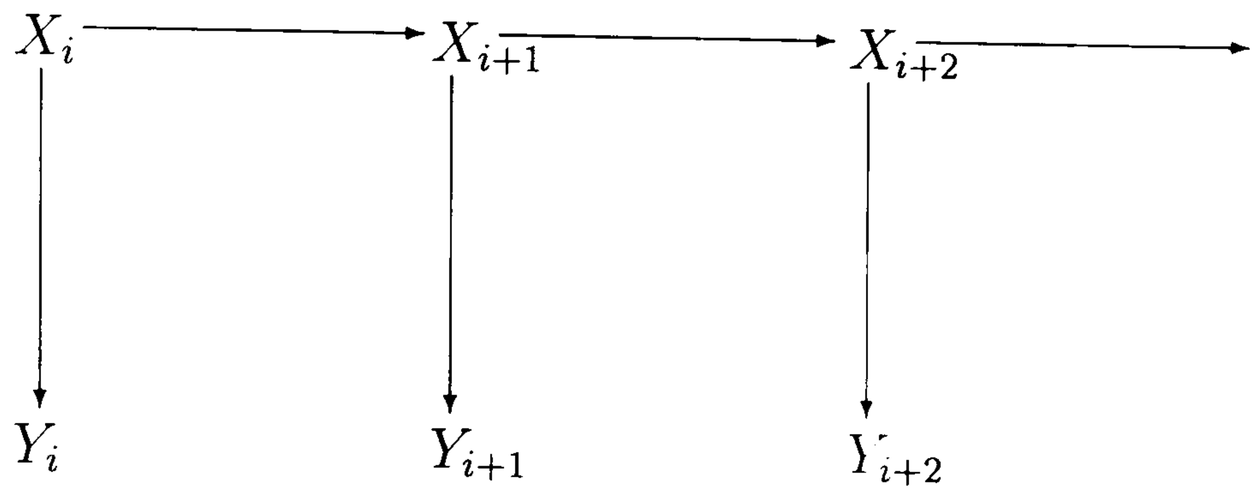

Figure 4.5: The causal structure prior to compression.

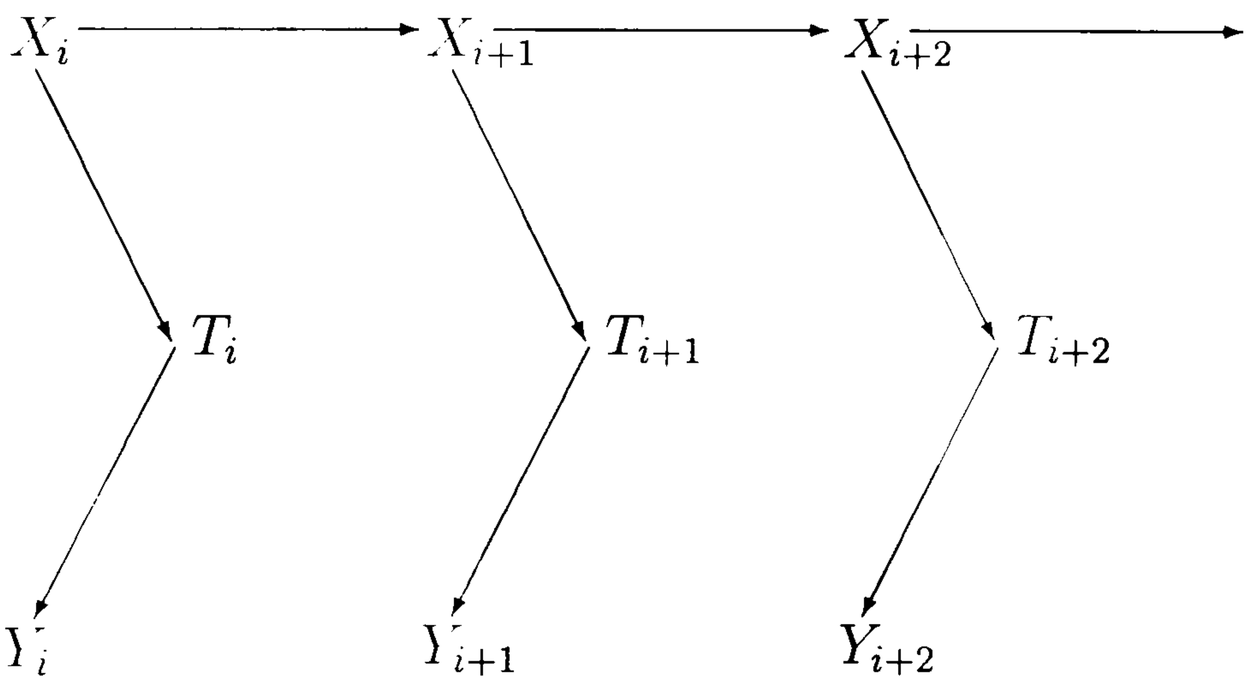

Figure 4.6: The causal structure of the model with the sufficient statistic T.

after the multivariate data has been compressed. The diagram in Figure 4.6 makes it clear that Y_i is conditionally independent of X_i given T_i. By moralising the graph it is also possible to see the implications T_i has on the predictions, ζ_{i+1} and ξ_{i+1}. X_{i+1} is conditionally independent of $\{Y_1, \ldots, Y_i\}$ given $\{T_1, \ldots, T_i\}$ and X_{i+1} also is conditionally independent of $\{Y_1, \ldots, Y_{i+1}\}$ given $\{T_1, \ldots, T_{i+1}\}$.

The forecast distributions for X_{i+1} conditional on D_i, and X_{i+1} conditional on D_{i+1} are equivalent to the forecast distributions for X_{i+1} conditional on $\{T_1, \ldots, T_i\}$ and $\{T_1, \ldots, T_{i+1}\}$ respectively. Therefore the forecasts, ζ_{i+1} and ξ_{i+1}, are the same whether they are based on the sufficient statistics or the observation sequence.

The results of section 4.7 show how the concept of sufficiency can be applied in an HMM structure. Y_i^* summarises the matrix $\mathbf{Y}_i$ in a univariate statistic with the following conditional independence property

$$Y_i \perp\!\!\!\perp \{X_1, \ldots, X_i\} | Y_i^*.$$

The residual analysis in section 4.7 reveal that the residuals obtained using Y_i are equivalent to those obtained using Y_i^* showing that there is no loss of information, and also that the predictions and updates in both cases are also equivalent. Y_i^*, therefore, is a sufficient statistic for X_i.

4.9 Discussion

The aim of this chapter is to use residuals to compensate for the lack of information available to the forecaster when dealing with a Hidden Markov Model. Residuals for the HMM are defined and it is hoped that through the analysis of the structure of the residuals for the hidden state sequence more information about the sequence can be gained. The results, however, show

that residuals, in the univariate case, for the hidden state are equivalent to the residuals of the observation sequence. These results hold even for the nonlinear case.

The Hamilton model presented an example of HMM with a more complex causal structure. In the Hamilton example a noisy signal at time i is a function of both the current and previous state of the system and past values of the observation sequence. The residuals for the state of system still show that they are a multiple of the residual for the observation sequence. The complex causal structure of the Hamilton model does not provide any further understanding of the hidden sequence.

All the models considered in this chapter have a system equation in the form (4.1) where the current state is a function of the previous state. All the residuals evaluated for this system equation show that the only information available for the state at time i is the corresponding observation at time i. It would, then, seem logical to consider a system equation in which the current state at time i is a function of the previous state at time $i-1$ and the observations at time i and $i-1$ making full use of the information.

The comparison of residuals gives a good basis for exploring statistical methods and making judgements on them. Section 4.7 looks at a Linear Dynamic Model with a univariate state variable and a multivariate observation sequence. For this special case, it is found that the observable series can be compressed in the form of a one dimensional transformation. The residuals for the state variable are computed using both the original and the minimised formulation and are found to be equivalent showing that the compression of the data results in no loss of information. The equivalence of the two residuals is proven algebraically in section 4.7. Section 4.8 also proves their equivalence using a more general argument which states that the one

dimensional compression being a function of the data is a sufficient statistic for the state.

A data compression technique such as this can be very practical when dealing with problems in speech recognition where there are many variables involved and also in other complex multivariate HMM applications. The reduction in the dimesionality simplifies the modelling scheme without loss of information and allows for a concentrated focus on the analysis of the variable of interest and the predictions made ffor them.

Chapter 5

Calibration for Hidden Markov Models

5.1 Introduction

From this chapter onwards, the analysis shifts from the assessment and analysis of point forecasts to the assessment and analysis of probability forecasts. The focus of the applications remains the hidden Markov model described in Chapter 4. Empirical calibration is the selected method of assessment for the probability forecasts of HMMs.

Calibration is a term commonly used to describe the discrepancy between the assessed probability of a sequence of events and the observed frequency of the occurance of those events. The formulation of probability forecasting assessment by calibration was largely established for the use of meteorologists. On the evening before any given day a weather forecaster is expected to make a forecast or a statement regarding his *degree of belief* about the outcome of rain on that particular day based on information currently available. At the end of the day, the event under speculation becomes known

and a value (rain $= 1$, no rain $= 0$) is observed. Over a long period, the forecaster develops a sort of history or forecasting portfolio. Using this portfolio a meteorologist's ability as a forecaster can be evaluated by empirical calibration.

For a forecaster to be well calibrated, it is expected that out of those days to which the forecaster assigns the probability p of rain, rain occurs a p proportion of the time. What results is a comparison between the probability forecast and the relative frequency of the outcome. Calibration can be evaluated graphically by plotting the relative frequency of the outcomes ω, against p, in what is called a calibration curve. The calibration curve of a well calibrated forecaster should lie on (or close to) the diagonal. A forecaster is said to be well calibrated if $p \equiv \omega$. In order to calibrate forecasts all that is required is a set of sequentially generated forecasts and its corresponding set of realised outcomes.

For probability forecasts made about the state of a hidden Markov model the corresponding set of realised outcomes is not available. Nevertheless, it is still necessary to evaluate the forecasts' performance. The outcomes are replaced with a one step ahead filtered prediction. With this in mind, assessment by calibration determines the forecasts' success in explaining the forecast's updated prediction.

This chapter attempts to judge the performance of probability forecasts for a sequence of unobserved outcomes, the state of an HMM, by defining a new calibration criterion. All of the work presented in this chapter is an extension of the calibration concepts introduced by Dawid (1982). Section 5.2 gives an introduction to these calibration techniques. How these concepts can be extended for use in HMM forecasting assessment is explained in section 5.3. The application of empirical calibration and the concepts developed

in section 5.3 are illustrated in section 5.4 through the use of two examples.

5.2 The Calibration Criterion

Before discussing the calibration criterion, it is first necessary to define a forecasting system and to clarify the basis behind the selection of calibration as the validation method for probability forecasts. Let $a = (a_1, a_2, \ldots)$ denote an infinite series of observed outcomes of an uncertain binary event observed sequentially over time. After observing a series of i outcomes, $a^i = (a_1, \ldots, a_i)$, it is possible to assign a probability P_{i+1} to the occurrence of the next event. Constructing such a probability forecast for each i and a^i constitutes a *forecasting system* (FS) (Dawid, 1986). A *prequential forecasting system* is defined as a forecasting system that is defined by a rule which associates a choice of P_{i+1} for every i and with any possible set of outcomes of $a^i = (a_1, \ldots, a_i)$.

Suppose the forecasts are generated sequentially from a fixed probability distribution Π. The purpose of forecasting assessment is to determine the overall adequacy of Π as a probabilistic explanation for a. The assessment should depend on Π only through the sequence of forecasts that it, in fact, made (Dawid, 1984). This requirement is referred to as the *Prequential Principle*. In application, an assessment method would require only the sequence of outcomes and their corresponding forecasts to be in accordance with the prequential principle. The criterion of *complete calibration* introduced by Dawid (1982) is one such method.

Given a sequence of forecasts, P, a subsequence is selected from it arbitrarily using an *admissible* selection process whereby the inclusion of any particular event i is determined by previous outcomes only, $(a_1, a_2, \ldots, a_{i-1})$,

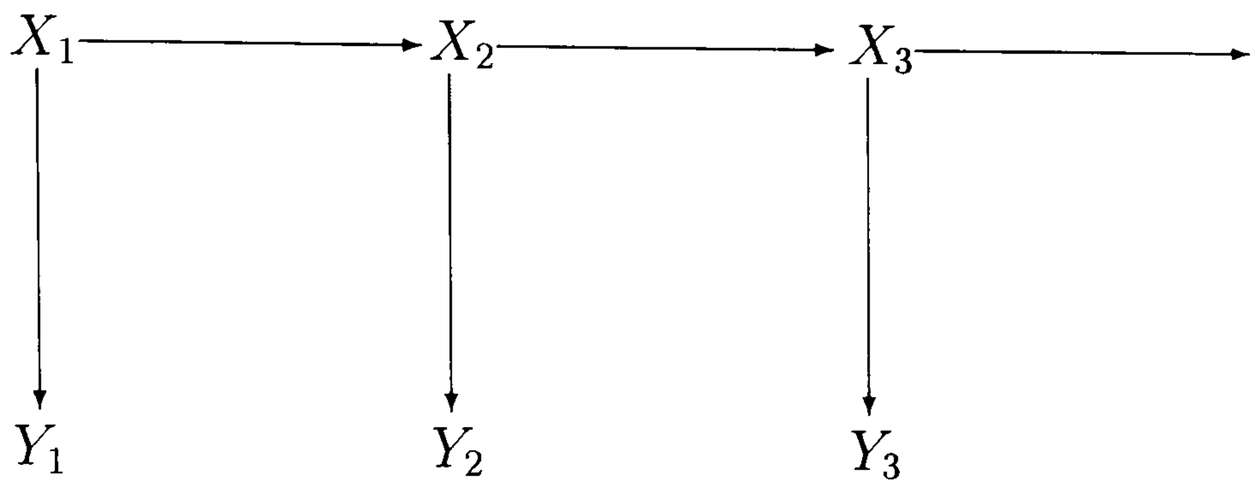

Figure 5.1: The causal structure of the HMM.

and not by a_i or any later outcome. If $\bar{P}^n$ is the average forecast probability for the first n events in such a subsequence, and $\bar{a}_n$ is the corresponding empirical relative frequency of those events, then the calibration criterion requires, inorder for the sequence P to be valid, that $(\bar{a}_n - \bar{P}^n) \longrightarrow 0$ as $n \longrightarrow \infty$.

5.3 The Calibration Criterion for HMMs

The previous chapter looked at the Normal linear state-space model with an unobservable state. A causal diagram for such a data generating system is given in Figure 5.1. Without reference to any particular model, a definition for both the forecast and observed value were given for any data generated as function of an underlying hidden state. In this chapter, the HMM models examined are restricted to HMMs with a discrete state space. Specifically, the state space of the HMM is binary $\{0,\ 1\}$ where the state of the system is either in a particular state $(X_i = 1)$ or not $(X_i = 0)$. To examine this class of HMMs, the p_i's and q_i's must be redefined as probability forecasts. Suppose that P is the joint distribution for $(X_1, X_2, \ldots)$. As such, the forecast for the state is now $p_i = P(X_i = 1|Y_i, \ldots,\ Y_{i-1})$, the conditional distribution for X_i given D_{i-1} and its outcome or update is $q_i = P(X_i = 1|Y_1,\ \ldots, Y_i)$.

Algorithm 5.3.1 describes the computation of p_i and q_i.

Algorithm 5.3.1

1. Initialisation:

$$\begin{aligned} p_1 &= P(X_1 = 1) \\ P(Y_1) &= P(Y_1|X_1 = 1)p_1 + P(Y_1|X_1 = 0)(1 - p_1) \\ q_1 &= \frac{p_1 P(Y_1|X_1 = 1)}{P(Y_1)}. \end{aligned}$$

2. Iterate: $i = 2, \ldots N$

$$\begin{aligned} p_i &= q_{i-1}P(X_i|X_{i-1} = 1) + (1 - q_{i-1})P(X_i|X_{i-1} = 0) \\ P(Y_i|D_{i-1}) &= P(Y_i|X_i = 1)p_i + P(Y_i|X_i = 0)(1 - p_i) \\ q_i &= \frac{p_i P(Y_i|X_i = 1)}{P(Y_i|D_{i-1})}. \end{aligned}$$

3. Stop. Note that the expectation is:

$$\begin{aligned} E(q_i|D_{i-1}) &= \sum_{Y_i} \frac{p_i P(Y_i|X_i)P(Y_i|Y_1, \ldots, Y_{i-1})}{P(Y_i|Y_1, \ldots, Y_{i-1})} \\ &= p_i \sum^{Y_i} P(Y_i|X_i) \\ &= p_i. \end{aligned}$$

These probability forecasts are used in this section to define a calibration criterion. This new criterion assess the p_i's performance in predicting the q_i's which adheres strongly to the calibration criterion of Dawid (1982) as well as running parallel with the proof of the former.

Assume that the forecasts are made sequentially from a fixed joint probability distribution P. To create an arbitrary test set of time points, an indicator variable, $U_i = \{0, 1\}$, is D_{i-1}-measurable and is used to denote

the exclusion or inclusion respectively of any time point i in the set, where $i = 1, \ldots, k, \ldots, n$. Let

$$\nu_k = \sum_{i=1}^{k} U_i, \quad \rho_k = \nu_k^{-1} \sum_{i=1}^{k} U_i q_i, \quad \pi_k = \nu_k^{-1} \sum_{i=1}^{k} U_i p_i.$$

The forecasts made are said to be completely calibrated if $\pi_k \doteq \rho_k$ is attained for a large enough collection of subsequences of admissible forecasts.

Theorem 5.1 *Let the selection process be admissible. Then, with P-probability* 1, $\pi_k - \rho_k \to 0$ *as* $\nu_k \longrightarrow \infty$.

Proof. Theorem 5.1 is a variation of Theorem 1.1 given in Chapter 1. Let $B_i = \nu_i^{-1}$, $X_i = B_i U_i (q_i - p_i)$, and $S_k = \sum_{i=1}^{k} X_i$. Since $E(q_i | D_{i-1}) = p_i$, and the conditional expectation of X_i, $E(X_i | D_{i-1})$ is equal to 0:

$$\begin{aligned} E(X_i | D_{i-1}) &= B_i U_i \left[E\left(q_i | D_{i-1}\right) - E\left(p_i | D_{i-1}\right) \right] \\ &= B_i U_i \left[p_i - p_i \right] \\ &= 0 \end{aligned}$$

and (S_k) is a martingale adapted to D_{k-1}. For any realisation of U_i's, the successive nonzero terms of the sequence $(B_1 U_1)^2$, $(B_2 U_2)^2$, ... are

$$1, \ \frac{1}{2^2}, \ \frac{1}{3^2}, \ \ldots,$$

so that,

$$\sum_{i=1}^{k} \left(B_i U_i\right)^2 \leq \sum_{n=1}^{\infty} n^{-2} = \frac{\pi^2}{6}.$$

Using the above result, and

$$E\left(X_i^2\right) = E\left(B_i U_i\right)^2 var\left(q_i | D_{i-1}\right) \leq \frac{1}{4} E\left(B_i U_i\right)^2$$

it can be shown that $E(S_k^2)$ is bounded:

$$E\left(S_k^2\right) \ = \sum_{i=1}^{k} E\left(X_i^2\right) \leq \tfrac{1}{4} E\left[\sum_{i=1}^{k} \left(B_i U_i\right)^2\right] \ \leq \frac{\pi^2}{24}.$$

Then, by the martingale convergence theorem the sequence

$$S_k = \sum_{i=1}^{k} B_i U_i \left(q_i - p_i\right)$$

converges with P-probability 1, and by Kronecker's lemma (Feller. 1971 pg. 238)

$$\rho_k - \pi_k = B_k \sum_{i=1}^{k} U_i \left(q_i - p_i\right) \longrightarrow 0,$$

as $\nu_k \longrightarrow \infty$. The only assumption made is that the p_i's are evaluated sequentially according to a fixed probability distribution P.$\diamondsuit$

5.4 Applications

In theory, if the model is appropriate, the calibration criterion should hold for any sequence of p_i's and q_i's since $E(q_i|D_{i-1}) = p_i$. It remains to be seen how q_i and p_i will behave in practice. The calibration criterion is illustrated by defining q_i and p_i for the examples below.

5.4.1 Example 1

This example is taken from the book *Biological Sequence Analysis* (Durbin. et al 1998) and is used here to demonstrate the application of the calibration criterion on a basic HMM configuration.

Consider a casino which occasionally switches from using a fair die. denoted by F, to a loaded die, denoted by L. At any given time, the die in use is determined by a Markov chain forming a series of hidden states. denoted by X, which are unknown to the gambler. The transition probabilities governing the switch between the fair and loaded die are summarised in the matrix below:

X_{i+1}

X_i		F	L
	F	0.95	0.05
	L	0.10	0.90

All that the gambler observes is the result of the toss, a value from 1 to 6. which make up the series of observed Y's. When a fair die is used there is equal probability of observing any of the six possible outcomes. The loaded die, however, favours an outcome of 6. The probability of the observed outcome is summarised below:

$$P(Y_i|X_i = F) = \begin{cases} \frac{1}{6} & if \quad Y_i = 1 \\ \frac{1}{6} & if \quad Y_i = 2 \\ \frac{1}{6} & if \quad Y_i = 3 \\ \frac{1}{6} & if \quad Y_i = 4 \\ \frac{1}{6} & if \quad Y_i = 5 \\ \frac{1}{6} & if \quad Y_i = 6 \end{cases}, \quad P(Y_i|X_i = L) = \begin{cases} \frac{1}{10} & if \quad Y_i = 1 \\ \frac{1}{10} & if \quad Y_i = 2 \\ \frac{1}{10} & if \quad Y_i = 3 \\ \frac{1}{10} & if \quad Y_i = 4 \\ \frac{1}{10} & if \quad Y_i = 5 \\ \frac{1}{2} & if \quad Y_i = 6 \end{cases}.$$

The observed series (shown in Figure 5.2) consists of 300 consecutive tosses of a die at the casino described above.

The forecasts and updates are computed using Algorithm 5.3.1 with the above model. Calibration applications on the data show that the forecasts are well calibrated. Setting $U_i = 1$ for all i, the calculated ρ_{300} and π_{300} were found to be 0.3566 and 0.3538 respectively.

Graphical analysis of the forecasts gives some additional insight into their behaviour. The plot in Figure 5.3 is a very crude summary of the relation between $(p_i,\ q_i)$. Let S_i be the set of prediction and update at time i, (p_i, q_i), computed from the observed series and let S^n denote the series $(S_1, S_2, \ldots, , S_n)$ of such pairs. For an infinite realisation of (p_i, q_i), S^∞, a joint distribution for (p_i, q_i), Π, is determined on the unit square. Π is a

1	3	1	5	1	1	6	2	4	6	4	4	6	6	4	4	2	4	5
19	3	1	1	3	2	1	6	6	1	1	6	4	1	5	2	1	3	3
37	6	2	5	1	4	4	5	4	3	6	3	1	6	5	6	6	2	6
55	5	6	6	6	6	6	6	5	1	1	6	6	4	5	3	1	3	2
73	6	5	1	2	4	5	6	3	6	6	6	4	6	3	1	6	3	6
91	6	6	3	1	6	2	3	2	6	4	5	5	2	3	6	2	6	6
109	6	6	6	6	2	5	1	5	1	6	3	1	2	2	2	5	5	5
127	4	4	1	6	6	6	5	6	6	5	6	3	5	6	4	3	2	4
145	3	6	4	1	3	1	5	1	3	4	6	5	1	4	6	3	5	3
163	4	1	1	1	2	6	4	1	4	6	2	6	2	5	3	3	5	6
181	3	6	6	1	6	3	6	6	6	4	6	6	2	3	2	5	3	4
199	4	1	3	6	6	1	6	6	1	1	6	3	2	5	2	5	6	2
217	4	6	2	2	5	5	2	6	5	2	5	2	2	6	6	4	3	5
235	3	5	3	3	3	6	2	3	3	1	2	1	6	2	5	3	6	4
253	4	1	4	4	3	2	3	3	5	1	6	3	2	4	3	6	3	3
271	6	6	5	5	6	2	4	6	6	6	6	2	6	3	2	6	6	6
289	6	1	2	3	5	5	2	4	5	2	4	2						

Figure 5.2: The data series for Example 1. The first column gives the starting value of i for each row.

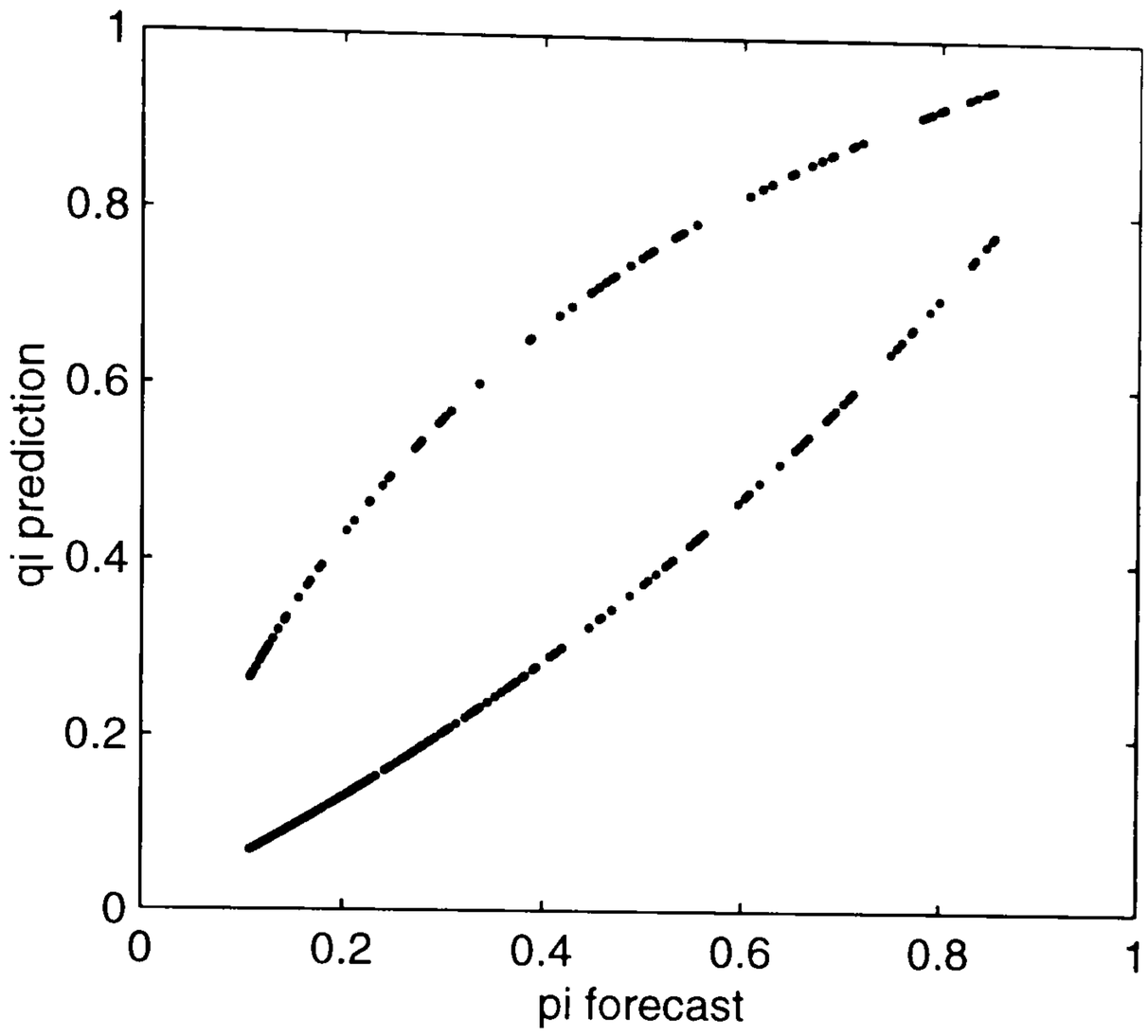

Figure 5.3: Plot of S_i for the dice example.

limiting empirical distribution for (p_i, q_i), and is random depending on the sequence S. The distribution, Π, contains less information then the actual sequence of p's, and q's. The order of the observations once S is plotted is no longer known.

The shape of the q_i versus p_i plot is largely determined by the value of Y_i. For this example, the outcome probability is one of three possibilities: $\frac{1}{6}$, $\frac{1}{10}$, and $\frac{1}{2}$ depending on the value of Y_i, and the current state of the system. The prediction of the state at time i as determined by the model depends primarily on the value of Y_i and whether or not it is equal to 6 or $\{1, 2, 3, 4, 5\}$, therefore, the sequence of Y_i's are used as binary sequence where each $Y_i = \{\{1, 2, 3, 4, 5\}, 6\}$. It is for this reason that the plot in Figure 5.3 takes the form of two separate curves around the diagonal. Figure 5.4 below is the plot of predictions against their updates for all values

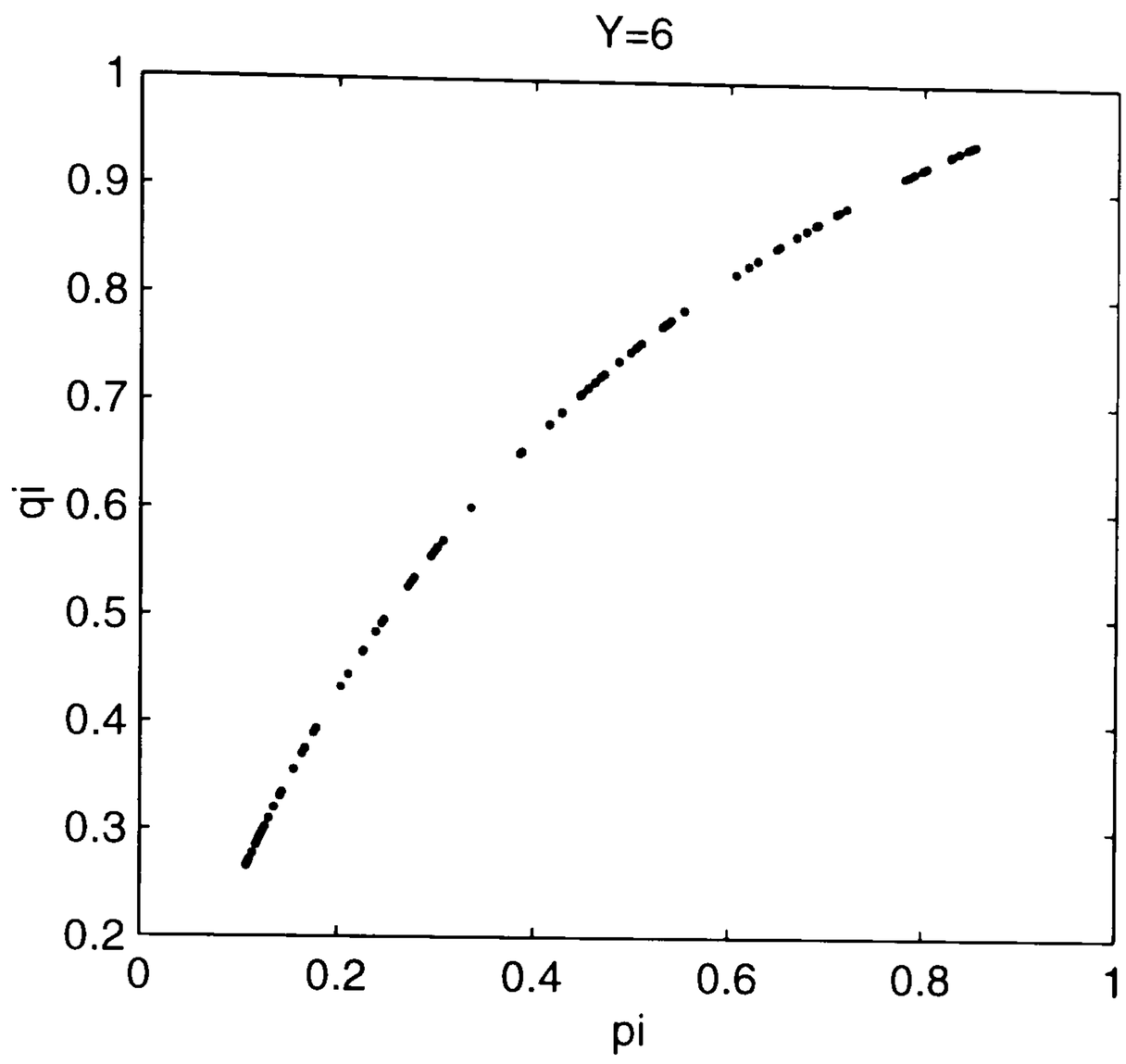

Figure 5.4: The plot of (p_i, q_i) pairs for $Y_i = 6$.

of i where $Y_i = 6$ and shows the top curve of Figure 5.3. The bottom curve of Figure 5.3 is composed of all pairs of p_i and q_i for which $Y_i = \{1, 2, 3, 4, 5\}$. This is shown in Figure 5.5.

According to the calibration criterion, the plot should show that the overall average of p_i's should be approximately equal to the overall average of q_i's for general calibration to hold. The plot of q_i versus p_i shown in Figure 5.3 shows an almost symmetric distribution of points around the diagonal. For an HMM, this is an indication of good calibration. Such a distribution of points indicates that the average q_i's for a fixed value of p_i are close in value p_i which is the desired attribute of well calibrated forecasts.

This is illustrated more clearly using a calibration plot. A calibration plot, such as the one in Figure 5.6, provides a venue for a more accurate assessment of calibration by means of the calibration criterion. The p_i forecasts

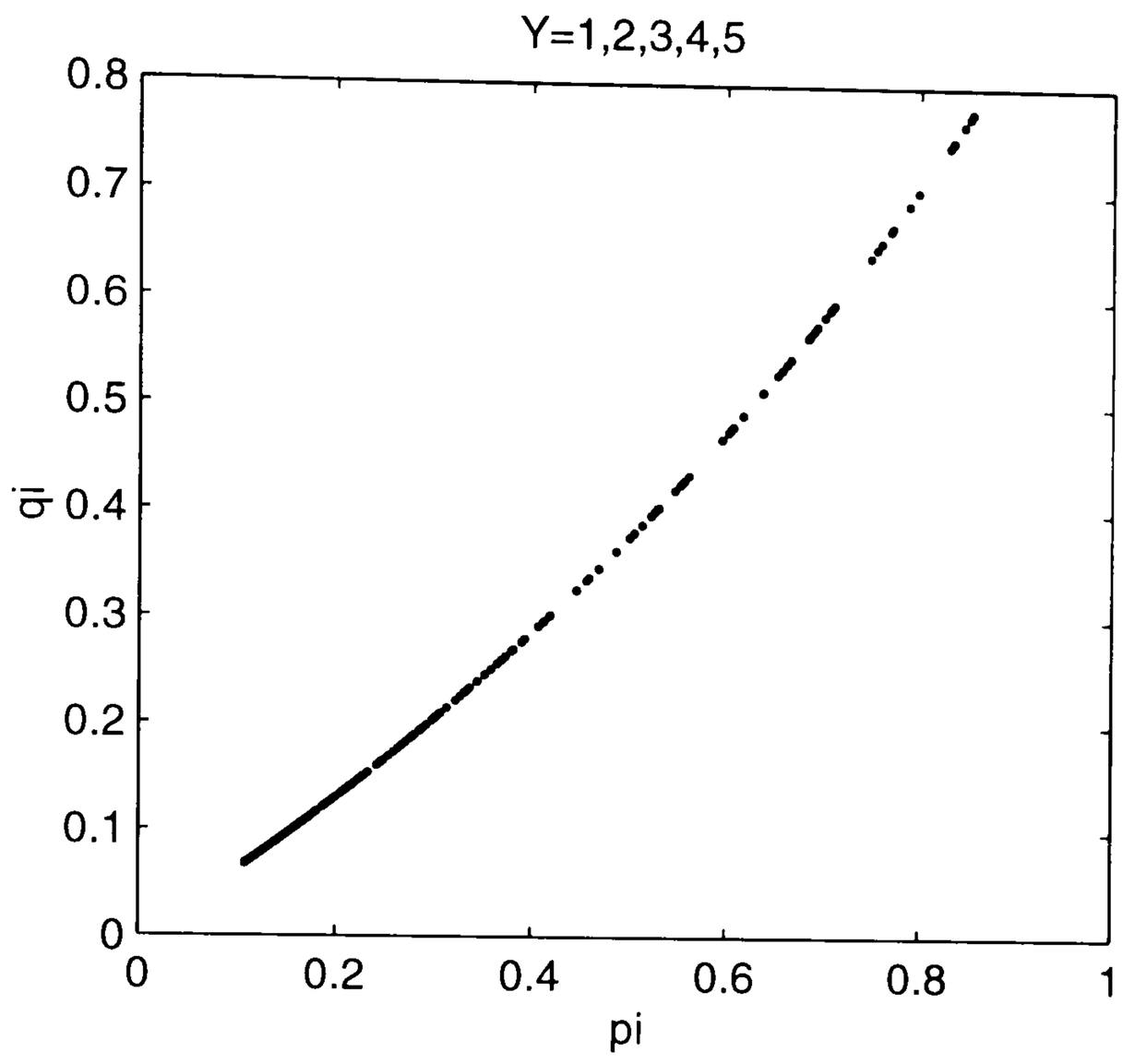

Figure 5.5: The plot of (p_i, q_i) pairs for $Y_i = \{1, 2, 3, 4, 5\}$.

are divided into prespecified intervals. For this example, and the remaining examples throughout this thesis, there are eleven such intervals:

$$\begin{aligned} 0 \leq p_i &< 0.05 \\ 0.05 \leq p_i &< 0.15 \\ 0.15 \leq p_i &< 0.25 \\ &\vdots \\ 0.85 \leq p_i &< 0.95 \\ 0.95 \leq p_i &\leq 1 \end{aligned} \quad .$$

For each of the eleven intervals, the average q_i is computed, $\bar{q}_j$, for those q_i's who's corresponding p_i lies within interval j, $j = 1, \ldots, 11$. This is plotted against $\bar{p}_j$, the average of the p_i's that lie within interval j. As specified by the calibration criterion, good calibration is indicated by a straight line through the diagonal of a $\{\bar{p}_j, \bar{q}_j\}$ plot. Examination of the calibration plot

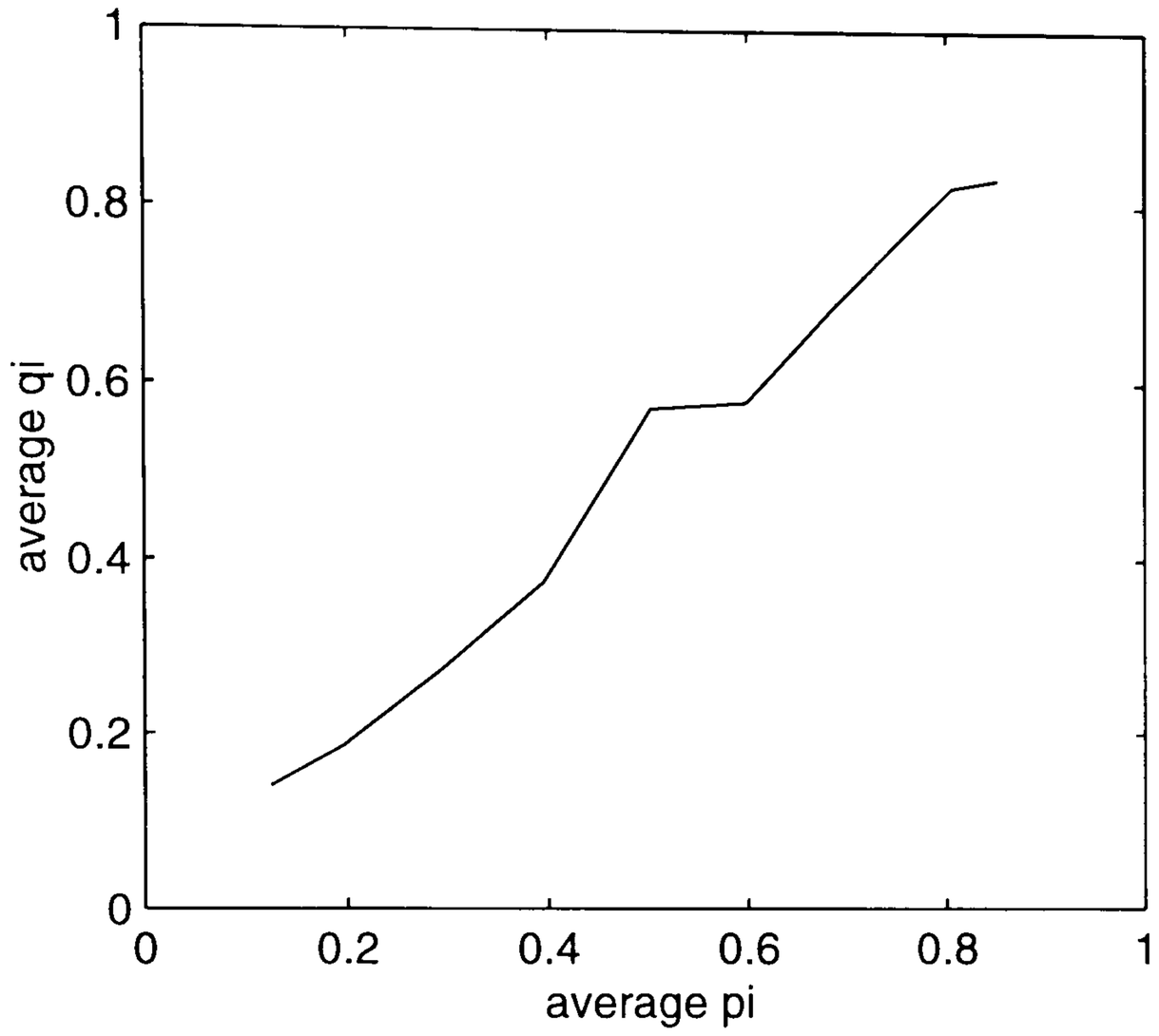

Figure 5.6: The plot of average q_i's versus average p_i's taken at fixed p_i intervals.

in Figure 5.6 shows that except for a slight bump near the $j = 6$ interval where $\bar{p}_6 = 0.5$, the points lie on the diagonal indicating that the forecasts are well calibrated.

5.4.2 Example 2a

Consider a Bernoulli distributed state variable X with success probability θ. Let Y be the observed binary variable where $Y|X$ is defined by the following transition matrix:

Y \ X	0	1
0	s	f
1	f	s

.

Then $P(Y = 1|\theta) = \theta s + (1 - \theta)f$. Note that in this example, the X's are independent of each other. Therefore, $\{X_i, Y_i\}$ pairs are also independent across i.

For a given value of θ,

$$\begin{aligned} p_i &= P(X_i = 1|Y_1, \ldots, Y_{i-1}) \\ &= P(X_i = 1) \\ &= \theta \end{aligned}$$

and

$$q_i = P(X_i = 1|Y_i) = \begin{cases} \frac{\theta f}{\theta f + (1-\theta)s}, & if \quad Y_i = 0 \\ \frac{\theta s}{\theta s + (1-\theta)f}, & if \quad Y_i = 1 \end{cases}.$$

It is easy to see in this example that the calibration criterion holds in theory. The conditional expectation of q_i is equal to

$$\begin{aligned} E\left(q_i|D_{i-1}\right) &= P(X_i = 1|Y_i = 0)P(Y_i = 0) + P(X_i = 1|Y_i = 1)P(Y_i = 1) \\ &= \frac{\theta f}{\theta f + (1-\theta)s}\left(\theta f + (1-\theta)s\right) + \frac{\theta s}{\theta s + (1-\theta)f}\left(\theta s + (1-\theta)f\right) \\ &= \theta \end{aligned}$$

Values for θ and s were generated randomly. Using $\theta = 0.7413$ and $s = 0.6009$, a series of 500 observations are simulated and used to compute p_i and q_i. With the values of s and θ known, q_i can take one of two values: $q_i = 0.655$ if $Y_i = 0$ or $q_i = 0.8119$ if $Y_i = 1$. The value of p_i remains constant at 0.7413, the value of θ. The statistics ρ_k and π_k are 0.7393 and 0.7413 respectively showing that the predictions are well calibrated overall.

5.4.3 Example 2b

The analysis has now changed slightly to take into consideration calibration of forecasts generated from a Bayesian model. Consider the same model in

section 5.4.2, but now assume that θ is uniformly distributed over the interval $[0, 1]$. First the value of p_i is computed:

$$\begin{aligned} p_i &= P(X_i = 1|Y_1, \ldots, Y_{i-1}) \\ &= E(P(X_i|Y_1, \ldots, Y_{i-1}, \theta)|Y_1, \ldots, Y_{i-1}) \\ &= E(\theta|Y_1, \ldots, Y_{i-1}). \end{aligned}$$

Further evaluation of this result requires the posterior distribution of $\theta|Y_1, \ldots, Y_{i-1}$. This distribution, however, cannot be determined directly since Y_i has a Bernoulli distribution with probability of success $\theta^* = (s - f)\theta + f$. To determine the posterior distribution, θ is expressed in terms of θ^*. To simplify the task of evaluating p_i, the posterior distribution of $\theta^*|Y_1, \ldots, Y_{i-1}$ is first determined and from that the result

$$p_i = E\left[\frac{\theta^* - f}{s - f}|Y_1, \ldots, Y_{i-1}\right]$$

can be obtained (note that the distribution of θ^* is Uniformly distributed over $[f, s]$ assuming $s \geq f$). This leads to a truncated Beta distribution with density of the form

$$f(\theta^*|Y_1, \ldots, Y_{i-1}) = a\left(\sum_{k=1}^{i-1} Y_k + 1,\ i - \sum_{k=1}^{i-1} Y_k\right) \theta^{*\sum_k^{i-1} Y_k} (1 - \theta^*)^{i-1-\sum_k^{i-1} Y_k},$$

where $f \leq \theta^* \leq s$, and the normalising constant of the pdf, $a(\cdot)$, is a function of the exponents of θ^*, and $(1 - \theta^{*)}$ such that,

$$a^{-1}(\alpha,\ \beta) = \int_f^s \theta^{*\alpha}(1 - \theta^*)^{\beta-1} d\theta.$$

The conditional expectation of θ is then

$$\begin{aligned} p_i &= E\left[\frac{\theta^* - f}{s - f}|Y_1, \ldots, Y_{i-1}\right] \\ &= \frac{a\left(\sum_{k=1}^{i-1} Y_k + 1,\ i - \sum_{k=1}^{i-1} Y_k\right) - a\left(\sum_{k=1}^{i-1} Y_k + 2,\ i - \sum_{k=1}^{i-1} Y_k\right) f}{(s - f)\, a\left(\sum_{k=1}^{i-1} Y_k + 2,\ i - \sum_{k=1}^{i-1} Y_k\right)}. \end{aligned}$$

q_i can be evaluated in a similar manner. The value of q_i. as in the case when θ is known, takes on two values depending on the value of $Y_i = \{0.\ 1\}$. For $Y_i = 0$, q_i takes the form

$$\begin{aligned} q_i &= P(X_i = 1 | Y_1,\ \ldots,\ Y_i = 0,) \\ &= E\left(P(X_i | Y_1,\ \ldots,\ Y_i = 0, \theta) | Y_i,\ \ldots,\ Y_i = 0\right) \\ &= E\left(\theta | Y_1,\ \ldots,\ Y_i = 0\right). \end{aligned}$$

Expressed in terms of θ^* the expectation becomes:

$$E\left[\frac{(\theta^* - f)\, f}{(s - f)\,(1 - \theta^*)} | Y_i,\ \ldots,\ Y_i = 0,\right],$$

the evaluation of which is

$$q_i = \frac{a\left(Y_k^* + 1, i - 1 - Y_k^*\right) f}{(s - f)} \left(\frac{1}{a\left(Y_k^* + 2, i - Y_k^*\right)} - \frac{f}{a\left(Y_k^* + 1, i - Y_k^*\right)}\right).$$

where $Y_k^* = \sum_{k=1}^{i-1} Y_k$ and for $Y_i = 1$,

$$E\left[\left(\frac{\theta^* - f}{s - f}\right) \frac{s}{\theta^*} | Y_1, \ldots, Y_i = 1\right] = \frac{s}{s - f}\left[\frac{a\left(Y_k^* + 1, i - Y_k^*\right) - a\left(Y_k^* + 2, i - Y_k^*\right)^2 f}{a\left(Y_k^* + 1, i - Y_k^*\right)}\right].$$

Further calculations show that $E(q_i | D_{i-1}) = p_i$.

The p_i and q_i were evaluated using the simulated series of section 5.4.2. Figure 5.7 is a plot of these predictions. Examination of the plot shows a high concentration of points around the value of θ. The calibration statistics. $\pi_k = 0.7181$ and $\rho_k = 0.7185$, not only indicate good calibration. but show that the statistics also seem to be approaching the unknown value of θ. the probability of observing $Y_i = 1$. The calibration plot in Figure 5.8 further emphasises that the forecasts are empirically well calibrated.

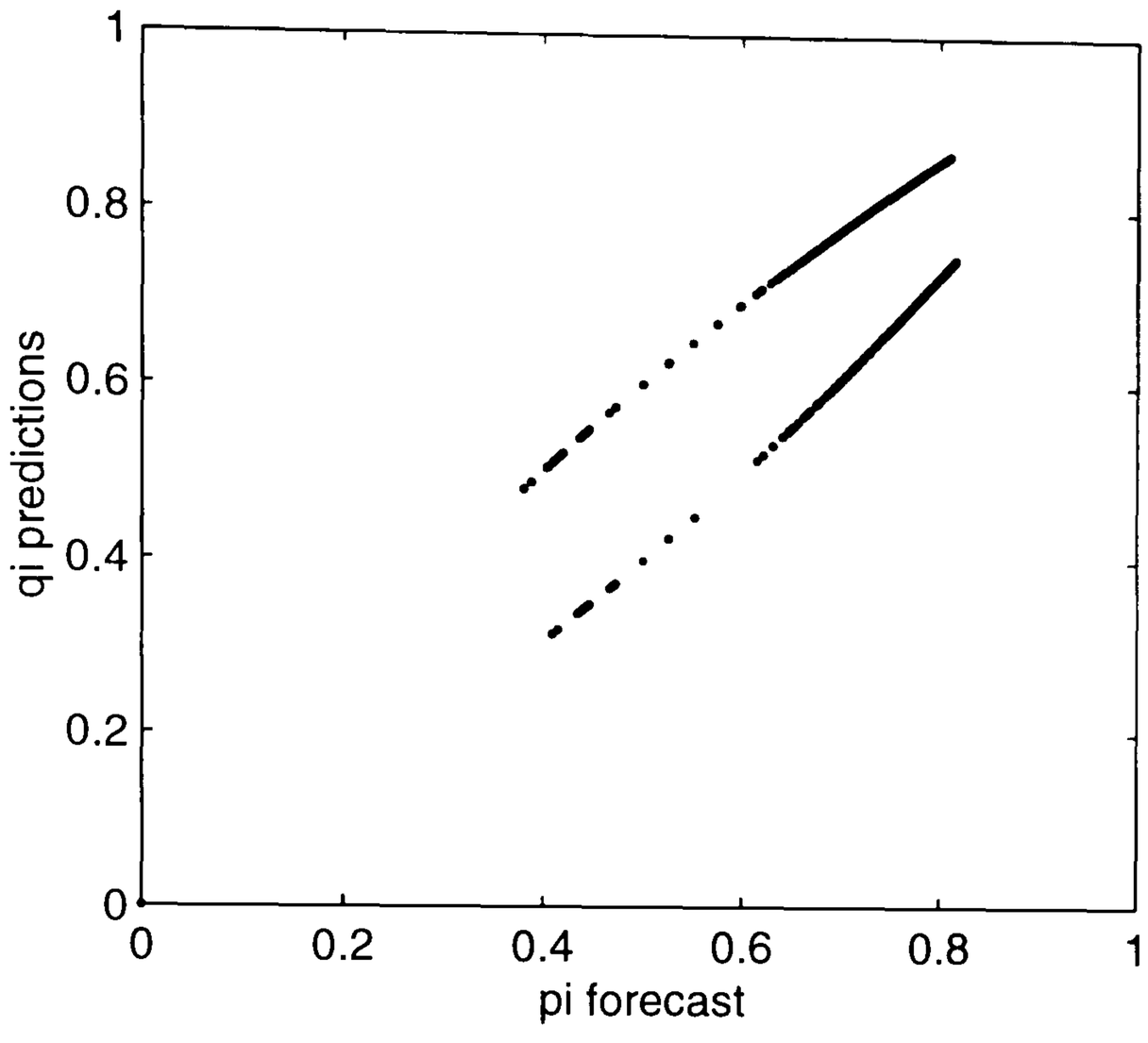

Figure 5.7: The plot q_i's versus p_i's for unknown θ.

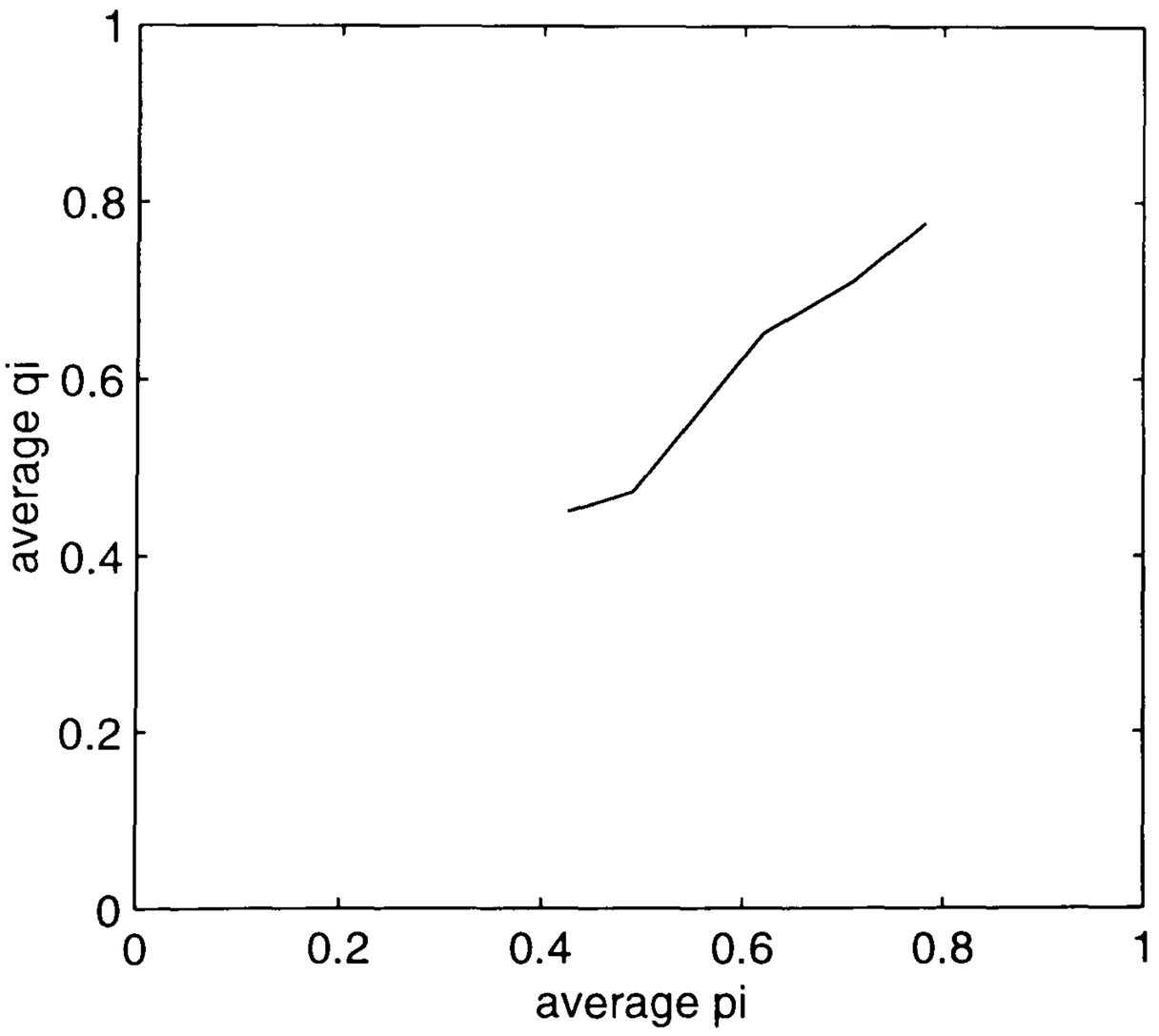

Figure 5.8: The plot of average q_i's versus average p_i's taken at fixed p_i intervals.

5.5 Discussion

Calibration, in its adherence to the prequential principle, provides a practical means of evaluating the validity of probability forecasts and provides a rule for judging the success of a forecasting system in explaining the outcomes. In the HMM structure the realised sequence of outcomes is replaced by the sequence of one step ahead predictions, q_i.

A theorem, a variation of the calibration criterion, is given which specifies the asymptotic behaviour of well calibrated forecasts for HMMs. The forecasts are said to be well calibrated if the difference between the average of the forecasts and the average of the update is zero. Two examples are used to investigate the calibration of the forecasts in application. In both examples, it is observed that the calibration criterion holds.

Chapter 6

CpG Island Example

6.1 Introduction

To illustrate the performance of the forecasts for the state of an HMM and their calibration, the calibration criterion is implemented on a real world problem. The data analysed for this purpose is the Xq28 DNA sequence, a human DNA sequence obtained from GenBank (accession number U82695). The sequence is approximately 80,000 base pairs long.

In the human genome, a CpG island is a strand in the DNA sequence characterised by the occurrences of CG dinucleotides (written as CpG to distinguish it from a C-G base pair across the two strands). Typically, wherever a CpG dinucleotide occurs in the sequence, the C nucleotide (cytosine) is chemically modified by a process called methylation. The methylation process often acts as a catalyst for mutation of the nucleotide from C to T making the CpG dinucleotide a rare incident. There are, however, short instances in the genome (a few hundred to a few thousand base pairs long) where the methylation process is suppressed. In these regions there are more CpG dinucleotides than elsewhere in the sequence and, in general, there are

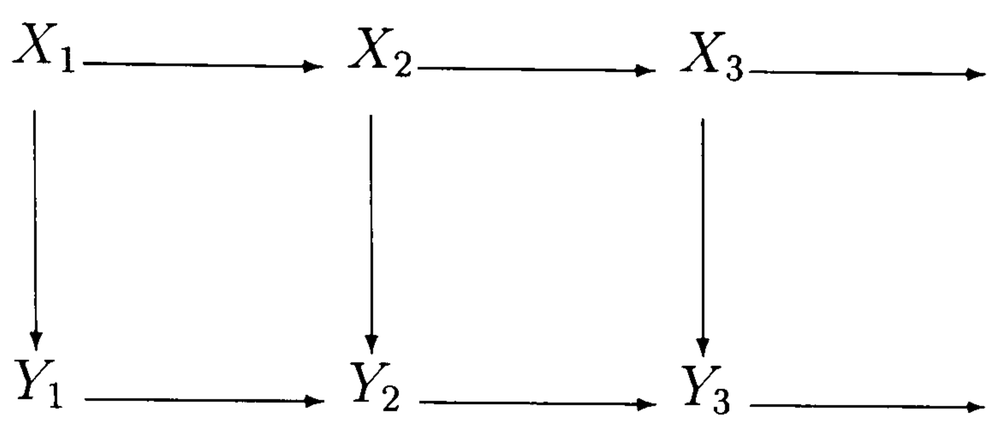

Figure 6.1: The causal diagram of the CpG Island model.

also more occurrences of C and G nucleotides. These regions are known as CpG islands, and they act as identifiable landmarks for locating genes.

For this example a short segment of 2500 base pairs of the DNA sequence is used. The segment begins at the 8000^{th} nucleotide and ends at 10500. This particular stretch of the DNA sequence has been chosen because previous studies have shown that it is a strand rich in CpG islands (Benson et al, 2000).

6.2 The Hidden Markov Model

The formulation of the HMM model for this problem includes two hidden states, the CpG island state, and the non-CpG island state corresponding to $X_i = 1$ and $X_i = 0$, respectively. Any DNA sequence consists of a long chain of four possible nucleotides, A, T, G, C, which together make up the set of observed outcomes for this example, $Y = \{A, T, G, C\}$. The outcomes are governed by a Markov process $P(Y_i|Y_{i-1}, X_i)$. Figure 6.1 illustrates the causal diagram.

The matrix of transition probabilities in the CpG island state is different from that of the non-CpG island state. The values of both the hidden state

X_{i+1}

X_i		0	1
	0	0.9997	0.000245
	1	0.001171	0.998829

non-CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	0.300	0.210	0.205	0.285
T	0.177	0.292	0.239	0.292
C	0.322	0.302	0.298	0.078
G	0.248	0.208	0.246	0.298

CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	0.180	0.120	0.274	0.426
T	0.079	0.182	0.355	0.384
C	0.170	0.188	0.368	0.274
G	0.161	0.125	0.339	0.375

Table 6.1: The transition probabilities of the CpG island model.

transition probabilities, $P(X_i|X_{i-1})$, and the transition probabilities for the observed outcomes were obtained from the literature (Churchill, 1992. Durbin et al, 1998). The transition probabilities are displayed in Table 6.1. The main distinction between the two states can be observed by comparing the transition probability $P(Y_i = G|Y_{i-1} = C)$ for both states. This probability is higher in the CpG island state at $P(Y_i = G|Y_{i-1} = C, X_i = 1) = 0.274$ than it is in the non-CpG island state where it is equal to $P(Y_i = G|Y_{i-1} = C, X_i = 0) = 0.078$.

The forecasts, p_i, and the update, q_i, are computed based on the definitions given in the previous chapter: $p_i = P(X_i = 1|Y_1, \ldots Y_{i-1})$ and $q_i = P(X_i = 1|Y_1, \ldots, Y_i)$. The plot shown in Figure 6.2 is the plot of the p_i's against q_i's.

It can be seen that the plot consists of sixteen very definite arcs on each side of the diagonal giving it its unique shape. Like the previous two exam-

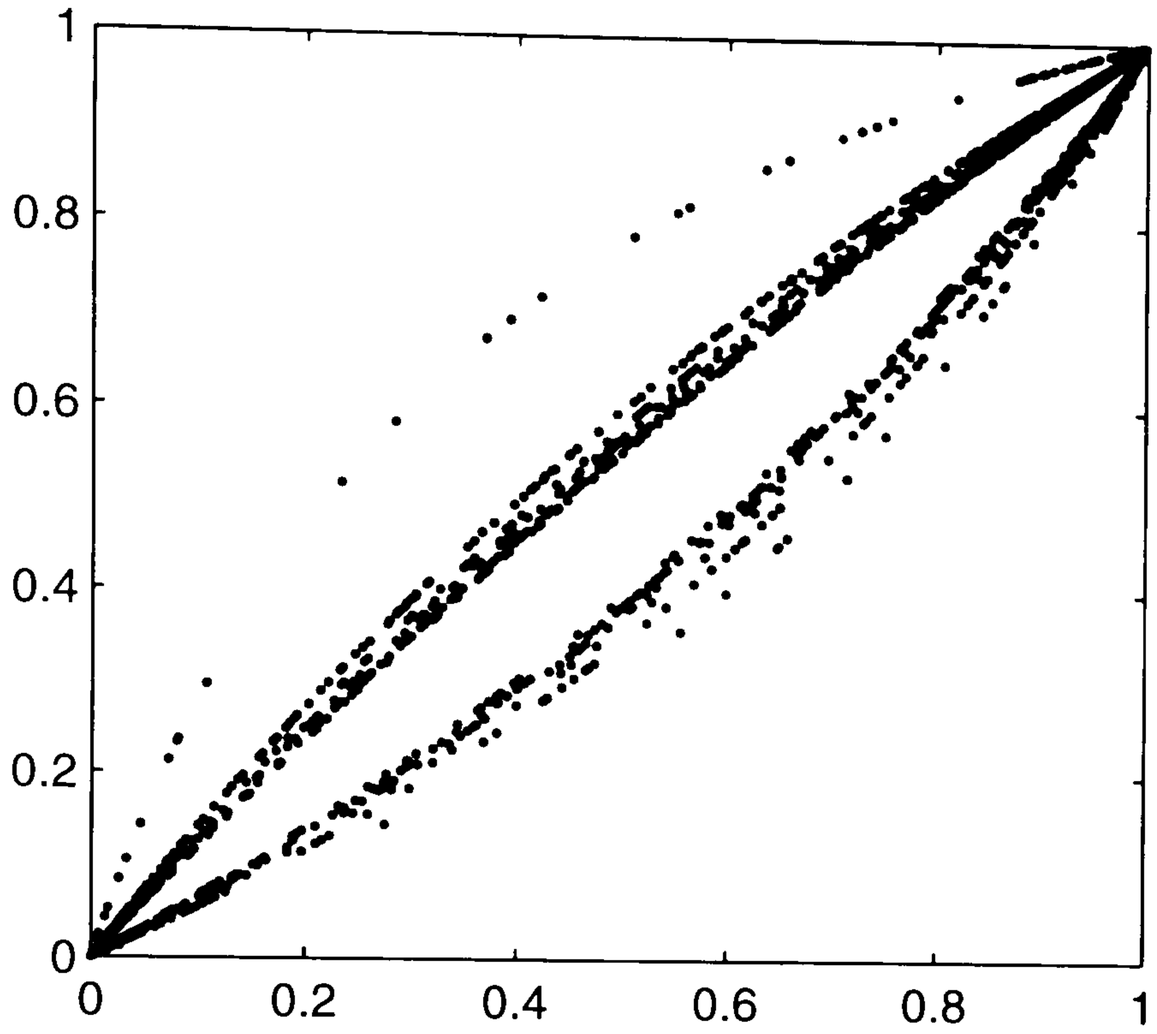

Figure 6.2: Plot of q_i updates against the p_i forecasts.

ples in Chapter 5, the distribution of the (p_i, q_i) pairs is for the most part determined by the value of Y_i. For this example, Y_i can take one of four possible values. The probability of an outcome, Y_i, however, is dependent not only on the state of the system at time i, but also on the value of Y_{i-1} since the series of outcomes also form a Markov Chain. Therefore, the number of possible transitions or outcomes are sixteen. Analysis of the forecast probabilities shows that each of these arcs pertains to one of the sixteen possible outcome transitions (i.e. a transition from A to A, A to T, A to C, A to G etc...). The arc on the outer boundary of the remaining points towards the top of the figure shows p_i against q_i when the transition of the observed values is from C to G.

Interval	Average p	Average q	n
$0 \leq p_i \leq 0.05$	0.0165	0.0180	382
$0.05 \leq p_i \leq 0.15$	0.0876	0.0914	133
$0.15 \leq p_i \leq 0.25$	0.2026	0.2210	76
$0.25 \leq p_i \leq 0.35$	0.2984	0.3161	92
$0.35 \leq p_i \leq 0.45$	0.3986	0.4180	96
$0.45 \leq p_i \leq 0.55$	0.4978	0.4956	104
$0.55 \leq p_i \leq 0.65$	0.6000	0.5943	107
$0.65 \leq p_i \leq 0.75$	0.7026	0.7139	125
$0.75 \leq p_i \leq 0.85$	0.8037	0.7931	184
$0.85 \leq p_i \leq 0.95$	0.9075	0.9031	346
$0.95 \leq p_i \leq 1.00$	0.9803	0.9796	856
overall	0.6413	0.6423	2501

Table 6.2: Average $q_i's$ for fixed values of p_i.

6.3 Assessing Calibration

For this working example the calibration criterion seems to hold well. For different values of p_i taken within a given interval, the mean value of q_i is computed and compared with the value of the former. Table 6.2 shows the results obtained.

Figure 6.3 is a plot of the average q_i for fixed values of p_i. The straight line across the diagonal is the sought after characteristic of calibrated forecasts.

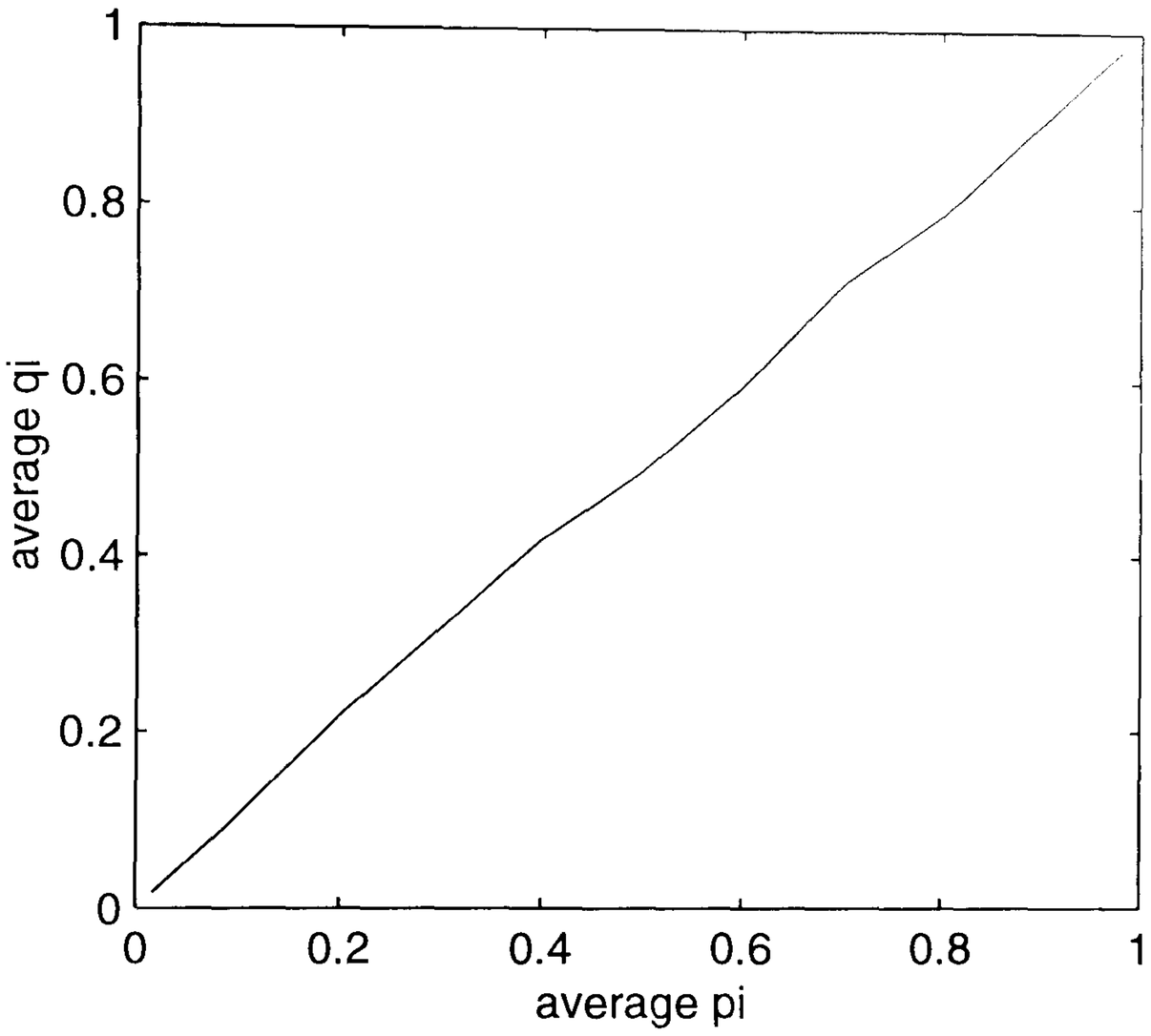

Figure 6.3: Calibration plot of the forecasts: for fixed values of p_i, the plot shows $\overline{q}_i$ against $\overline{p}_i$.

6.3.1 The test statistic

Figure 6.3 above only shows the empirical calibration of the forecasts. In order to assess accurately the calibration of the forecasts, a test statistic is used. Sellier-Moiseiwitsch and Dawid (1993) developed a test statistic in the framework of probability forecasting to test the discrepancy between the number of times an event of interest occurred and the number of times that same event is *expected* to occur. With minor adjustments the same test statistic can be adapted for use in this analysis.

The test statistic used to test the overall calibration of the forecasts, Z_0, is defined as

$$Z_0 = \frac{\sum (q_i - p_i)}{[\sum var(q_i|D_{i-1})]^{1/2}}. \tag{6.1}$$

A variation on this test statistic can also be defined as

$$Z_k = \frac{\sum U_i(q_i - p_i)}{[\sum U_i var(q_i|D_{i-1})]^{1/2}}, \tag{6.2}$$

where $U_i = 0$, or 1 is D_{i-1}-measurable meaning that the value of U_i is assigned using only the information available at time *i-1*. The sequence of U's allows the assessment of a predetermined subset of points such as testing the calibration of the forecasts for fixed values of p_i, or prespecified intervals of p_i. For the remainder of this chapter, Z_k $(k = 1, \ldots, 11)$ will be used to denote the test statistic for the k subsequences of (p_i, q_i) pairs where the pair is included in the k^{th}-subsequence if its corresponding p_i lies within the prespecified k^{th}-interval (refer to Chapter 5). In the case where $U_i = 1$ for all i, Z_k becomes the overall test statistic Z_0

The null hypothesis of this test statistic maintains that the p_i forecasts and the q_i updates fulfil the requirements of the validity criterion of complete calibration as discussed in Chapter 5. For the hidden Markov model, the calibration criterion does not draw upon any particular probability model,

but rather makes a single assumption that the p_i forecasts for q_i are generated sequentially as conditional probability forecasts from a fixed distribution P. The null hypothesis for complete calibration assessment can then be stated as the assertion that the p_i forecasts are constructed sequentially from the same probability distribution P as that from which the q_i updates are generated.

6.3.2 Derivation of the test statistic

Suppose that, as stated in the null hypothesis, the (p_i, q_i) pairs are generated from the same distribution P and that the σ - fields, D_i, containing all the information available at time $i = 1, \ldots, n$ are nested, so that $D_i \subseteq D_{i+1}$.

Let $S_i = U_i(q_i - p_i)$, where $U_i = 0, 1$ is D_{i-1}-measurable and let $\tilde{S} = \sum_i^n S_i$. Then, under the distribution P, $(\tilde{S})$ is a martingale adapted to D_n. The cumulative conditional variance of (S_i), $\sum_{i=1}^n U_i var(q_i|D_{i-1})$, is denoted by W^n. The calibration test statistic can then be written as $Y_n = \frac{\tilde{S}}{(W^n)^{1/2}}$. If $U_i = 1$ for all i, then Y_n is merely Z_0.

Theorem 6.1 *Suppose that c_n is a sequence of constants such that $c_1 < c_2 < \cdots \to \infty$, and η is a strictly positive finite random variable for which under P, the following conditions hold:*

$$\sum c_n^{-2} E[S_n^2] < \infty \tag{6.3}$$

$$\sum P(|S_n c_n^{-1}| > \varepsilon) \xrightarrow{p} 0, \text{ for all } \varepsilon > 0 \tag{6.4}$$

$$\frac{W^n}{c_n^2} \xrightarrow{p} \eta^2 \tag{6.5}$$

$$D_i \subseteq D_{i+1}. \tag{6.6}$$

Then $Y_n \xrightarrow{L} N(0,1)$.

Proof: Theorem 6.1 is a variation on the Central Limit Theorem for Martingales (Hall and Heyde, 1980, pg. 64) with $\frac{S_i}{c_i}$ above, replacing X_{ni} and $D_i = \mathcal{F}_{ni}$.$\diamondsuit$

Seillier-Moiseiwitsch and Dawid (1992) used this variation of the Central Limit Theorem for Martingales to derive the distribution of the their test statistic. The same theorem can be applied here to derive the distribution of Y_n. Condition (6.4) will hold if the cumulative expectation in (6.3) is bounded which (as shown in the proof of Theorem 5.1) is the case if $c_i = \sum_{j=1}^{i} j^2$. For condition (6.5) to hold, W^n must essentially tend to infinity at an identical rate for each sequence, regardless of the data generated by P. Also, η is a random variable since its value need not be the same for different realisations of the forecasts and updates. Condition (6.5) is justifiably upheld if it can be assumed that, for the data at hand, W^n will approach infinity as more and more data become available.

6.3.3 Generalisations and results

The results of Theorem 6.1 can also be extended to test statistics evaluated for several subsequences of (p_i, q_i) as long as the length of the subsequences is infinite. The subsequences calibration test statistic can be formulated as a multivariate generalisation of the overall calibration test. Let $S_{ik} = U_{ik}(q_i - p_i)$, where U_{ik} is D_{i-1}-measurable indicator variable indicating the inclusion or exclusion of the forecast and update pair at time i in the subsequence k. Each p_i forecast can only be assigned to one of K subsequences, so that $S_i = (S_{i1}, \ldots, S_{iK})$ is a K-length vector with one $(q_i - p_i)$ element and $K - 1$ zero-elements, and $S_{ih}S_{ik} = 0 \ \forall \ k \neq h$ so that all the subsequences are disjoint. The cumulative conditional covariance matrix of S_i is a diagonal matrix with elements of the form

$$W_k^n = \sum_{i \in k}^{n} var(q_i | D_{i-1}). \quad (k = 1, \ldots, K).$$

Interval	n	Test Statistic Z	p-value	Z^2	p-value (χ^2)
$0 \leq p_i \leq 0.05$	382	2.5671	0.0103	6.5900	0.00142
$0.05 \leq p_i \leq 0.15$	133	1.0023	0.3162	1.0046	
$0.15 \leq p_i \leq 0.25$	76	1.9904	0.0465	3.9617	
$0.25 \leq p_i \leq 0.35$	92	1.8778	0.0604	3.5261	
$0.35 \leq p_i \leq 0.45$	96	1.6991	0.0893	2.8869	
$0.45 \leq p_i \leq 0.55$	104	-0.1900	0.8493	0.0361	
$0.55 \leq p_i \leq 0.65$	107	-0.5120	0.6087	0.2621	
$0.65 \leq p_i \leq 0.75$	125	1.2842	0.1991	1.6491	
$0.75 \leq p_i \leq 0.85$	184	-1.8649	0.0622	3.4779	
$0.85 \leq p_i \leq 0.95$	346	-1.9557	0.0505	3.8247	
$0.95 \leq p_i \leq 1.00$	856	-1.7382	0.0822	3.0213	
overall	2501	0.8408	0.4005	$\sum Z_k^2 = 30.24$	

Table 6.3: Results of the calibration test statistic.

Using this formulation, Helland(1982) shows how both the asymptotic Normality and asymptotic independence of the K test statistics, for the departure from calibration within each subsequence, follows from the multivariate generalisation of Theorem 6.1.

Since the test statistics are asymptotically Normally distributed and independent, it is possible to infer that $\sum_k Z_k^2$ has a χ_K^2 distribution and can be used as a as a portmanteau statistic to test the validity of calibration performance combining the test statistics derived from each subsequence in an overall measure of discrepancy. This allows the assessment of calibration over all the probability ranges simultaneously.

Table 6.3 above shows the value of the test statistic and the corresponding p-value for the overall test of calibration and for calibration of various subsets.

Each of the subsets includes only those values of $(p_i,\ q_i)$ such that p_i lies within a prespecified interval. The overall test indicates that the predictions are well calibrated. The χ^2 test, shown in the far right column of the table is equivalent to 30.24 with 11 degrees of freedom giving a significance level of just over 0.1%. In general, the results for the subgroups show that the p_i's and the q_i's are calibrated, however, in some subsets (especially when the predictions near the zero-one extremes) the p-values are small. The χ^2 test statistic also shows a high level of significance. Since the event under scrutiny, the state of the system, is a binary event then it would seem reasonable to assume that the test statistic would be more sensitive in the subsets nearer to, or containing, zero and one. Although none of the p-values, except for the χ^2, show a high level of significance, the results raise some concern about the validity of the model.

6.4 The Test Statistic Distribution

To investigate the test statistic distribution further, the analysis resorts to Fisherian inferential methods as described by Dawid (1995). The underlying concept behind Fisher's inference techniques is that of *inductive inference*, a method of extracting information solely from the data at hand. This requires the selection of the appropriate *frame of reference* which will supply the inferential model used to analyse the given data. The frame of reference is specifically designed for the data observed and hence the inferential model depends largely on the data.

Fisher's approach to statistical inference is in stark contrast to the more commonly used inferential methods based on Neyman's concept of *inductive behaviour*, an investigation of the long run performance of different inference

making rules and a comparison of the performance of these rules. Instead of formulating the analysis around the observed data, as in the Fisherian method, Neyman's primary focus revolved around the production model, the probability process either known or assumed to have generated the data at hand. It is through the repeated sampling of the production model that the long term performance of different inference making methods can be analysed and compared.

The test statistic distribution is examined using both the prequential frame of reference and the production frame of reference. For both frames, an empirical distribution for the test statistic is built using 1000 simulated values of Z_0 and Z_k. The prequential frame of reference uses the available observations as a foundation for the simulation. The production frame of reference, on the other hand, uses the model described in Table 6.1 to generate new samples of data. Both techniques are discussed in greater detail below.

6.4.1 The prequential frame of reference

The probability forecasting techniques and assessments used throughout this study all adhere to the prequential principle (Dawid, 1985), as discussed in Chapter 1, Chapter 4, and Chapter 5. It is, therefore, consequential that the frame of reference used to make inferences about the test statistic distribution is the prequential frame of reference. The prequential principle and the many criterion it fulfils in assessing empirical probability statements (Dawid, 1985) make the prequential method a favourable choice as a frame of reference in its own right.

An empirical distribution function for each of the test statistics is built using 1000 simulated values of Z_0 and Z_k, $k = 1, \ldots, 11$. Each test statistic is computed from the original 2500 long segment of the DNA sequence

consisting of the 8000^{th} to the 10500^{th} nucleotide. At each point i in the observation sequence, a prediction, p_i, is computed using only past values of this original sequence. A new q_i is then generated using $\tilde{y}_i$ in place of Y_i, where $\tilde{y}_i$ is a simulated value for Y_i simulated from the distribution of $Y_i|D_{i-1}$. The information base $D_{i-1} = \{Y_i, \ldots, Y_{i-1}\}$ is the set of past observations from the original data sequence. At each time i, $\tilde{y}_i$ is simulated and used only to compute q_i after which it is discarded. This procedure is repeated for each observation in the sequence, a total of 2500 repetitions. The p's and q's are then used to evaluate $Z_0, \ldots, Z_k$, giving one value for each test statistic. This process is repeated 1000 times, generating 1000 values of the test statistics. Note that throughout the 1000 iterations the value of the p_i's remains the same since they are always computed from the original sequence. The flowchart in Figure 6.4 illustrates the simulation procedure.

The Normal probability plots for the simulated test statistics in Figure 6.5 show that they are all reliably Normally distributed. The results show that the *p-value* of observing Z_0, evaluated empirically from the simulated distribution, is 0.3900 which is very close to that of the Normal distribution. Table 6.4 summarises the results obtained.

It is interesting to note that although the estimates for the transition probabilities are not estimated based on information from this particular sequence (but rather taken from the literature), this does not seem to affect the calibration of the forecasts. Based on these results, the Z calibration test statistics are evidently Normally distributed. Therefore, the p-values computed for the test statistics are correct and it also follows that the p_i are validated by the complete calibration criterion as perfectly calibrated forecasts of their q_i updates.

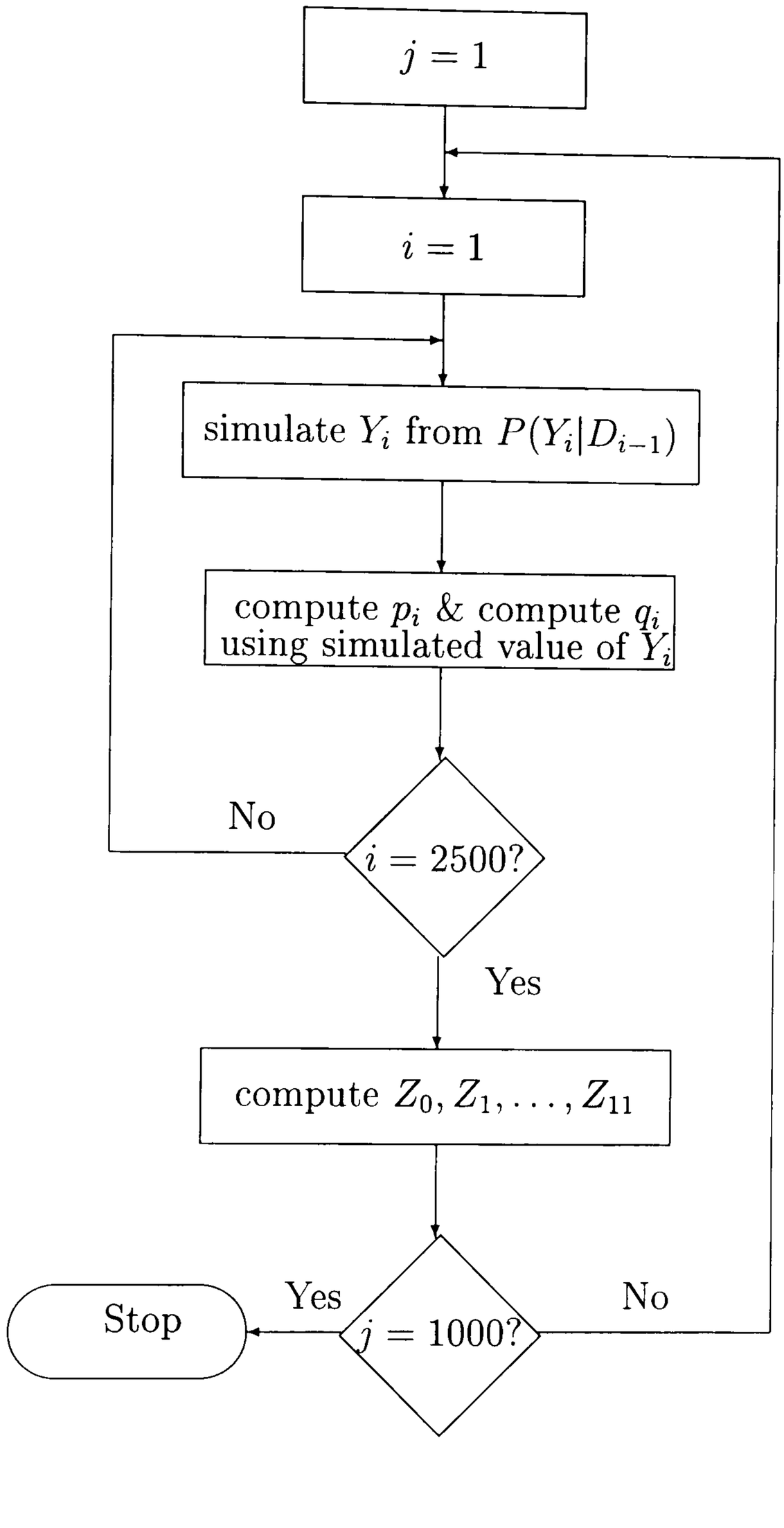

Figure 6.4: Flowchart of the test statistic simulation using the prequential frame of reference. Note that since the p_i's are fixed, for computational efficiency, they can be computed prior to the simulation procedure.

Figure 6.5: The normal probability plots, simulated Z versus n-score, for the prequentially simulated test statistics.

Interval	n	Test Statistic Z	Two-tailed Normal p-value	Simulated p-value *Prequential*	Simulated p-value *Production*
$0 \leq p_i \leq 0.05$	382	2.5671	0.0103	0.0100	0.0190
$0.05 \leq p_i \leq 0.15$	133	1.0023	0.3162	0.3300	0.3610
$0.15 \leq p_i \leq 0.25$	76	1.9904	0.0465	0.0500	0.0490
$0.25 \leq p_i \leq 0.35$	92	1.8778	0.0604	0.0490	0.0390
$0.35 \leq p_i \leq 0.45$	96	1.6991	0.0893	0.0820	0.0730
$0.45 \leq p_i \leq 0.55$	104	-0.1900	0.8493	0.8320	0.6700
$0.55 \leq p_i \leq 0.65$	107	-0.5120	0.6087	0.6070	0.4340
$0.65 \leq p_i \leq 0.75$	125	1.2842	0.1991	0.2220	0.1520
$0.75 \leq p_i \leq 0.85$	184	-1.8649	0.0622	0.0550	0.0370
$0.85 \leq p_i \leq 0.95$	346	-1.9557	0.0505	0.0460	0.0380
$0.95 \leq p_i \leq 1.00$	856	-1.7382	0.0822	0.0780	0.0530
overall	2501	0.8408	0.4005	0.3900	0.5690

Table 6.4: Results of the prequential and the production test statistic simulations. The n column refers the number of p_i forecasts, computed for the original data sequence, that lie within the specified interval.

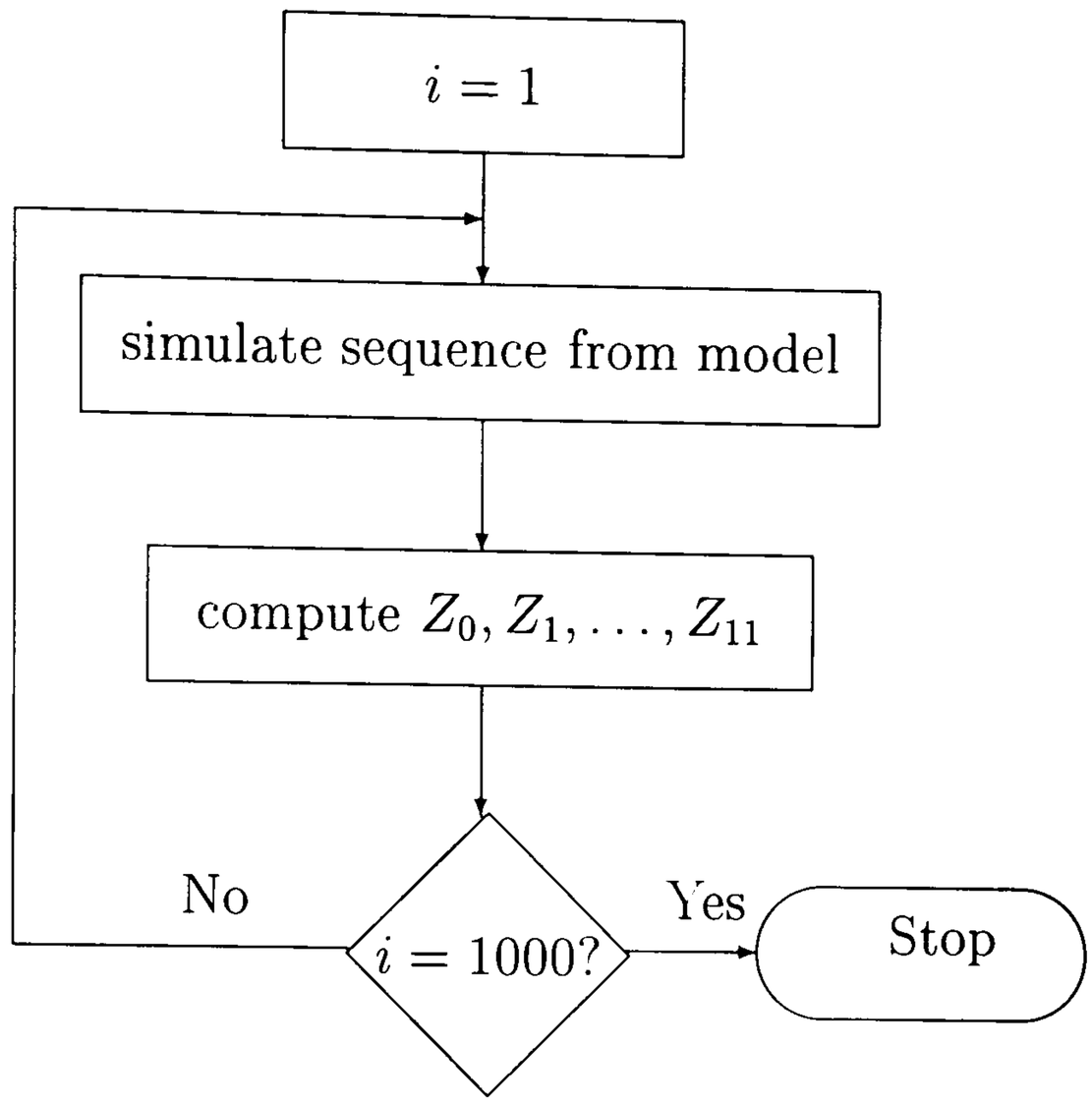

Figure 6.6: Flowchart of the test statistic simulation using the production

frame of reference.

6.4.2 The production frame of reference

In order to validate the results obtained in section 6.4.1 above, the *production principle* must be adhered to. The production principle is the term given for the minimal validity requirement on any inferential procedure. This principle states that the overall probabilistic properties of the production experiment should be compatible with the conclusions drawn from the selected method of inference. The production experiment examines the long run behaviour of data produced from what is presumed to be the correct model.

For this example, the samples are simulated from the model described in Table 6.1. By repeatedly sampling from this model, 1000 data sequences of length 2500 are generated and the test statistics are calculated for each sequence. This process constitutes the production experiment. The flowchart in Figure 6.6 describes the process.

In the production frame of reference predictions for the state are evaluated from the correct model. The state transition probabilities give a very low probability of a transition between states. There is also little fluctuation between states in the simulated sequences and the probability forecasts computed for the simulated sequence reflect this, often leaning towards the 0/1 extremes. Hence, for many of the samples there may not be probability forecasts made that lie within specific forecast ranges defined for the test statistics of the subsequences. Due to this, in some iterations, not all of the Z_k's can be computed.

Table 6.4 shows that the empirical p-values for the test statistics evaluated using the production model are not close in value to their Normal, or prequential counterparts. Even the production *p-value* for the overall calibration test statistic, Z_0, evaluated at 0.5690 is far from equal to the Normal *p-value* of 0.4005. Again, this could be attributed to the forecasts generated under the production frame of reference and the forecasts' values close proximity either to 0 or 1 and the small sample sizes of the Z_k's simulated distributions.

The Normal probability plots of the production simulated test statistics are given in Figure 6.7. The plots show that the test statistics $Z_4, \ldots, Z_9$ are slightly misaligned towards the tails of their distributions, but the plots still lie within the limits of a "straight line" interpretation. Otherwise, the remaining plots show that the distribution for the test statistics is reliably Normal. Hence, by the *production principle*, the inferences drawn using the prequential frame of reference uphold the probability statements of the production model. The test statistics are Normally distributed.

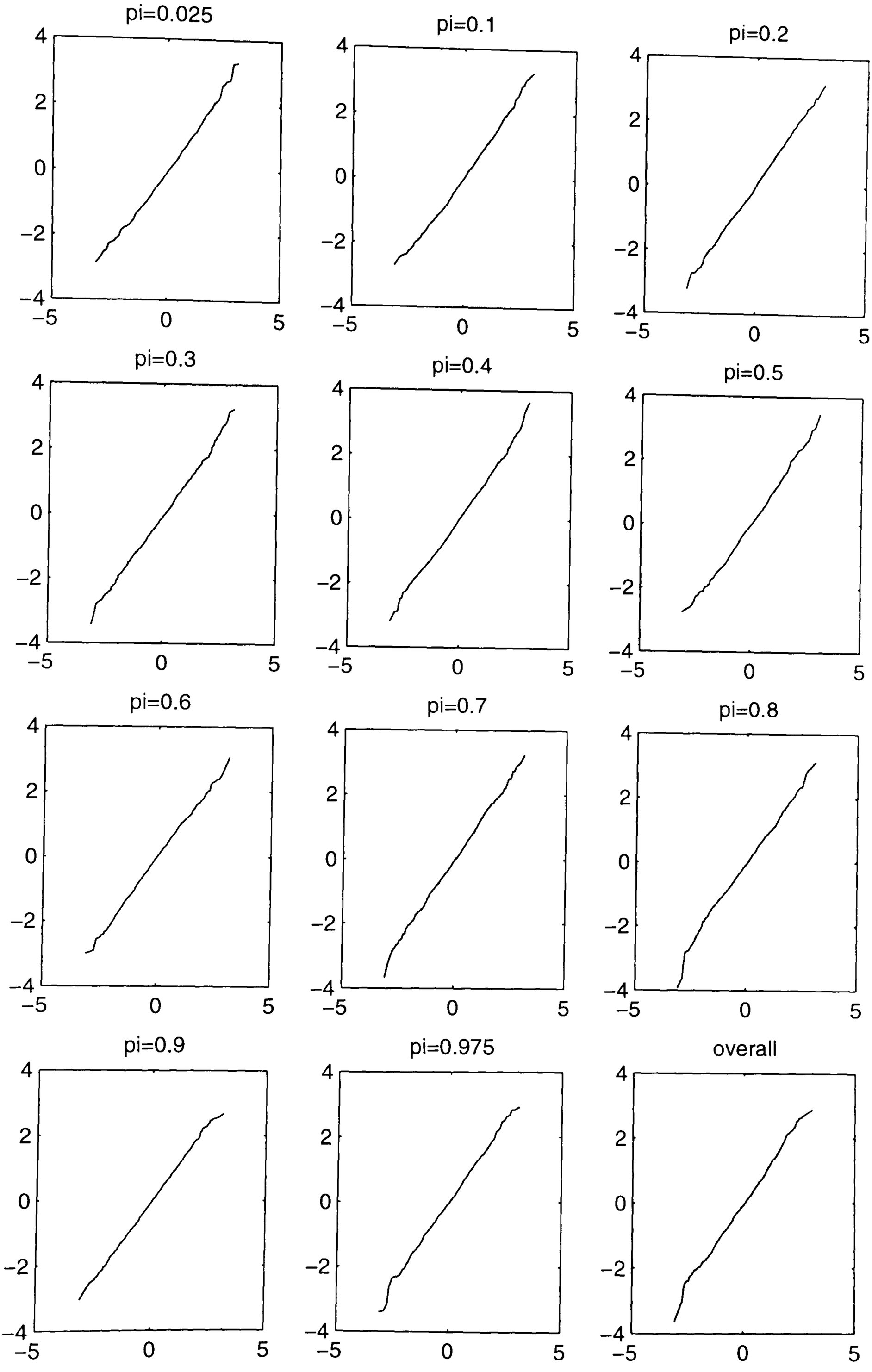

Figure 6.7: The normal probability plots, simulated Z versus n-score, for the production simulated test statistics.

6.5 Discussion

By using the test statistic introduced in section 6.3, it is possible to assess the calibration for an infinite sequence of (p_i, q_i) pairs. The test statistic tests the hypothesis that the p_i's and q_i's are completely calibrated. Using a variation of the central limit theorem for martingales, the distribution of test statistic is shown to be asymptotically standard Normal. The p-value's for the eleven test statistics evaluated for the DNA sequence show that, in general, the forecasts and updates are well calibrated.

Although Theorem 6.1 assures that the test statistic distribution is asymptotically Normal, the test statistics distribution is examined to see if this result remains true for the small sample sizes used in this example. This is done by simulating an empirical distribution of the test statistic and comparing the empirical p-value with that of the Normal. The comparison of the two p-values and normal probability plots of the simulated distribution both show that the Normal distribution is the distribution of the test statistics evaluated for the given data.

Chapter 7

Estimation

7.1 Introduction

Excluding Ex. 2b in Chapter 5, assessment of the probability forecasts' calibration has been based on a HMM with fixed parameter values. For comparative purposes, different estimation techniques are applied to the analysis of the CpG island example to examine how the estimation of the transition probabilities affects the calibration of the forecasts.

The HMM for the CpG island example consists of 36 Markov transition probabilities: four possible state transitions and sixteen possible transitions between nucleotides within each state. Together, these transition probabilities make up the model parameters to be estimated.

In this chapter, two estimation techniques are compared: the Baum-Welch estimation procedure and the prequential estimation method. Both methods are analysed to determine the impact estimation has on the calibration of the forecasts.

7.2 Baum-Welch estimation

The Baum-Welch algorithm (Buam, 1972, Rabiner, 1989) is used to estimate the transition probabilities for a test segment of the sequence. It is a special case of the EM algorithm, commonly used to estimate parameter values for HMMs. Given an observation sequence as training data, the algorithm uses an iterative re-estimation procedure to find parameter estimates for the model to maximise the probability of the observed sequence. However, it can produce estimates that may only locally maximise the likelihood.

7.2.1 The EM algorithm

The Expectation-Maximisation (EM) algorithm (Dempster et al, 1977) is a general algorithm that provides a procedure for executing maximum likelihood estimation in the presence of missing data. Let θ denote the set of all model parameters determining a statistical model. Furthermore, let $\mathbf{y}$ denote the vector of observed quantities and $\mathbf{x}$ denote the missing data. The data vector $\mathbf{y}$ is regarded as *incomplete* and is considered to be an observable function of the complete data. The notion of incomplete data is used to refer to situations where there is missing data and also refers to situations where the data contains variables that are never observed.

The purpose of the EM algorithm is to find a value for θ that maximises the log likelihood,

$$\log P(\mathbf{y}|\theta) = \log \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{y}|\theta),$$

using iterative re-estimation. Using $P(\mathbf{x}, \mathbf{y}|\theta) = P(\mathbf{x}|\mathbf{y}, \theta)P(\mathbf{y}|\theta)$, the log likelihood can be expressed as

$$\log P(\mathbf{y}|\theta) = \log P(\mathbf{y}, \mathbf{x}|\theta) - \log P(\mathbf{x}|\mathbf{y}, \theta). \tag{7.1}$$

Assume that the current model is determined by θ^t. the aim is to find a new and better model determined by θ^{t+1}.

Multiplying (7.1) by $P(\mathbf{x}|\mathbf{y}, \theta^t)$ and sum over $\mathbf{x}$ yeilds

$$\log P(\mathbf{y}|\theta) = Q(\theta|\theta^t) - \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}, \theta^t) \log P(\mathbf{x}|\mathbf{y}. \theta),$$

where

$$Q(\theta|\theta^t) = \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}, \theta^t) \log P(\mathbf{x}. \mathbf{y}|\theta). \tag{7.2}$$

If $\log P(\mathbf{y}|\theta)$ is expected to be larger than $\log P(\mathbf{y}|\theta^t)$, then the difference

$$\log P(\mathbf{y}|\theta) - \log P(\mathbf{y}|\theta^t) =$$
$$Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}, \theta^t) \log \frac{P(\mathbf{x}|\mathbf{y}, \theta^t)}{P(\mathbf{x}|\mathbf{y}, \theta)}$$

should always be positive. This difference can be expressed as

$$\log P(\mathbf{y}|\theta) - \log P(\mathbf{y}|\theta^t) \geq Q(\theta|\theta^t) - Q(\theta^t|\theta^t) \tag{7.3}$$

since

$$\sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}, \theta^t) \log \frac{P(\mathbf{x}|\mathbf{y}, \theta^t)}{P(\mathbf{x}|\mathbf{y}, \theta)}$$

is the relative entropy of $P(\mathbf{x}|\mathbf{y}, \theta^t)$ relative to $P(\mathbf{x}|\mathbf{y}, \theta)$ and is. therefore. always non-negative. The expression in (7.3) becomes an equality only if $\theta = \theta^t$ or $P(\mathbf{x}|\mathbf{y}, \theta) = P(\mathbf{x}|\mathbf{y}, \theta^t)$ for $\theta \neq \theta^t$. A positive difference (and thus a larger likelihood for the new model) can be derived by taking

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t).$$

If a maximum has already been reached, then $\theta^{t+1} = \theta^t$ and the likelihood will not change, otherwise the likelihood increases with each iteration of the algorithm. The EM algorithm is given below.

Algorithm 7.2.1 EM

E-step: *Calculate function* $Q(\theta|\theta^t)$.

M-step: *Maximise* $Q(\theta|\theta^t)$ *with respect to* θ.

An EM interpretation of the Baum-Welch algorithm can be found in Durban et al (1997). For the hidden Markov model, the hidden states compromise the missing data. The likelihood is expressed as

$$\log P(Y_1, \ldots, Y_N) = \sum_{\mathbf{X}} \log P(X_1, \ldots, X_N, Y_1, \ldots, Y_N | \theta).$$

where θ is the set of all parameters to be estimated, and $\sum_{\mathbf{X}}$ denotes the sum over all the sequence of hidden states $(X_1, \ldots, X_N)$ required to obtain the marginal probability of the observation sequence. The function Q is given by

$$Q(\theta|\theta^t) = \sum_{\mathbf{X}} P(X_1, \ldots, X_N | Y_1, \ldots, Y_N, \theta) \log P(X_1, \ldots, X_N, Y_1, \ldots, Y_n).$$

For a given sequence of states, each parameter will appear a given number of times in $P(Y_1, \ldots, Y_N, X_1, \ldots, X_N|\theta)$. Let x_{kj} denote the number of times a transition from state k to state j occurs, and let y_{sk} denote the number of times s is observed while in state k. Using x_{kj} and y_{sk}, $P(Y_1, \ldots, Y_N, X_1, \ldots, X_N|\theta)$ can be expressed as:

$$P(Y_1, \ldots, Y_N, X_1, \ldots, X_N|\theta) = \tag{7.4}$$

$$\prod_k \prod_s P(Y_i = s|X_i = k)^{y_{sk}} \prod_k \prod_j P(X_{i+1} = j|X_i = k)^{x_{kj}} \tag{7.5}$$

Using (7.5), the function $Q(\theta|\theta^t)$ for the hidden Markov model is:

$$Q(\theta|\theta^t) = \sum_{\mathbf{X}} P(X_1, \ldots, X_N|Y_1, \ldots Y_N, \theta^t) \times$$

$$\left[\sum_k \sum_s P(Y_i = s|X_i = k) \log\ y_{ks} + \sum_k \sum_j P(X_{i+1} = j|X_i = k) \log\ x_{ks} \right].$$

The expectations of y_{sk} and x_{kj} with respect to $P(X_1, \ldots, X_n|Y_1, \ldots, Y_N, \theta^t)$ is given by

$$E[y_{sk}] = \sum_{\mathbf{X}} P(X_1, \ldots, X_n|Y_1, \ldots, Y_N, \theta^t) y_{sk}.$$

and

$$E\left[x_{kj}\right] = \sum_{\mathbf{X}} P(X_1, \ldots, X_n | Y_1, \ldots, Y_N, \theta^t) x_{kj}.$$

respectively. The function $Q(\theta|\theta^t)$ can be restated in terms of the above expectations:

$$Q(\theta|\theta^t) = \sum_k \sum_s E\left[y_{sk}\right] \log\ y_{ks} + \sum_k \sum_j E\left[x_{kj}\right] \log\ x_{ks}.$$

The E-step of the HMM application of the EM algorithm consists of calculating $E[y_{sk}]$ and $E[x_{kj}]$, which completely determined the Q function. The M-step consists of plugging in the values of these expectations into the re-estimation formulas.

7.2.2 The estimation procedure

For the CpG island model the algorithm was altered slightly to incorporate the Markov structure in the observation sequence. Table 7.1 describes the standard Baum-Welch formulas and the alternative computations used for the CpG island example. The expectations listed in Table 7.1 can be evaluated by using the formulas given below:

$$\begin{aligned} P(X_i = k|D_n) &= \sum_j P(X_i = k, X_{i+1} = j|D_n), \\ P(Y_i = r, Y_{i-1} = s, X_i = k|D_n) &= \sum_{i=1 \ s.t. \ y_{i-1}=s, \ y_i=r}^{N} P(X_i = k|D_n), \\ P(Y_{i-1} = s, X_i = k|D_n) &= \sum_{i=1 \ s.t. \ y_{i-1}=s}^{N} P(X_i = k|D_n). \end{aligned}$$

The marginal distribution of $Y_i = r$, is

$$P(Y_i = r|X_i = k) = \frac{\sum_{i=1}^n \sum_s P(Y_i = r, Y_{i-1} = s, X_i = k|D_n)}{\sum_{i=1}^n \sum_s P(Y_{i-1} = s, X_i = k|D_n)}.$$

Thus, once

$$P(X_i = k, X_{i+1} = j|D_n) \tag{7.6}$$

Baum-Welch	CpG Island Model
$P(X_{i+1}=j\|X_i=k)$	$P(X_{i+1}=j\|X_i=k)$
$=\frac{E[\text{num of transitions from } k \text{ to } j]}{E[\text{num of transitions from } k]}$	$=\frac{E[\text{num of transitions from } k \text{ to } j]}{E[\text{num of times } k \text{ is visited}]}$
$=\frac{\sum P(X_i=k,X_{i+1}=j\|D_n)}{\sum P(X_i\|D_n)}$	$=\frac{\sum P(X_i=k,X_{i+1}=j\|D_n)}{\sum P(X_i\|D_n)}$
$P(Y_i=r\|X_i=k)$	$P(Y_i=r\|Y_{i-1}=s,X_i=k)$
$=\frac{E\left[\begin{array}{c}\text{num of times in state } k \\ \text{and observing } r\end{array}\right]}{E[\text{ num of visits to } k\ ]}$	$=\frac{E\left[\begin{array}{c}\text{num of times } s \text{ is observed} \\ \text{followed by } r \text{ in state } k\end{array}\right]}{E\left[\begin{array}{c}\text{num of times } s \text{ is observed} \\ \text{followed by visit to state } k\end{array}\right]}$
$=\frac{\sum P(Y_i=r,X_i=k\|D_n)}{\sum P(X_i=k\|D_n)}$	$=\frac{\sum P(Y_i=r,Y_{i-1}=s,X_i=k\|D_n)}{\sum P(Y_{i-1}=s,X_i=k\|D_n)}$
$=0$ if $Y_i\neq r$	$=0$ if $Y_i\neq r$ and $Y_{i-1}\neq s$

Table 7.1: The estimation updating formulas where $r,s = A,T,G,C$ and $j,k=0,1$.

is computed for all i, the remaining probabilities can be derived.

Before the estimation procedure is described, it is first necessary to explain the evaluation of $P(X_i=k,X_{i+1}=j|D_n)$. Expressed as

$$\frac{P(Y_1,\ldots,Y_i,\ X_i=k)P(X_{i+1}=j|X_i=k)P(Y_{i+1}|Y_i,X_{i+1}=j)P(Y_{i+1},\ldots,Y_n|X_{i+1}=j,Y_i)}{\sum_k\sum_j P(Y_1,\ldots,Y_i,X_i=k)P(X_{i+1}=j|X_i=k)P(Y_{i+1}|Y_i,X_{i+1}=j)P(Y_{i+1},\ldots,Y_n|X_{i+1}=j.Y_i)},$$

it can be seen that (7.6) is made up of two main components:

$$P(Y_1,\ \ldots,\ Y_i,\ X_i=k) \tag{7.7}$$

and,

$$P(Y_{i+1},\ \ldots,\ Y_n|X_{i+1}=j,Y_i). \tag{7.8}$$

The *forward-backward* algorithm is a compilation of two recursive algorithms which between them compute the two components required to evaluate (7.6).

Given the values for the transition probablities, the forward algorithm described in Algorithm 7.2.2, computes (7.7), the first part of (7.6), by using a *forward* pass through the data. In contrast, the backward algorithm, described in Algorithm 7.2.3 computes the second component, (7.8), by working backwards through the data.

Algorithm 7.2.2 Forward

1. *Initialisation:*

$$P(Y_1, X_1 = k) = P(X_1 = k)P(Y_1|X_1 = k), \qquad k = 0, 1.$$

2. *Recursion:*

$$P(Y_1, \ldots, Y_i, X_i = k) = \left(\sum_s P(Y_1, \ldots, Y_{i-1}, X_{i-1} = s)P(X_i = k|X_{i-1} = s)\right) P(Y_i|X_i = k, Y_{i-1}),$$

where $s = 0, 1$ *and* $i = 2, \ldots, n$.

3. *Stop when* $i = n$.

Algorithm 7.2.3 ***Backward***

1. *Initialisation:*

$$P(Y_n|X_n = j) = 1, \qquad j = 0, 1.$$

2. *Recursion:*

$$P(Y_{i+1}, \ldots, Y_n|X_i = j, Y_i) = \sum_s P(X_{i+1} = s|X_i = j)P(Y_{i+1}, \ldots, Y_n|X_{i+1} = s)P(Y_{i+1}|X_{i+1} = s, Y_i).$$

where $s = 0, 1$ *and* $i = 1, \ldots, (n-1)$.

3. *Stop when* $i = 1$.

Algorithm 7.2.4 describes the estimation procedure.

Algorithm 7.2.4 Baum-Welch

1. *Initialisation: select model parameters' starting values.*

2. *Using all the data and current parameter values, evaluate*

$$P(X_i = k, X_{i+1} = j | D_n)$$

   using the forward-backward algorithm in 7.2.2 and 7.2.3.

3. *Compute new parameter estimates, using Table 7.1 and the above results.*

4. *With new parameter estimates, return to step* 2

5. *Stop if there is either no change in the likelihood or the estimates have converged.*

Note that the new or updated estimates are derived using the values of the estimates evaluated in the previous iterations.

7.2.3 Results

The parameter values given for the model listed in Table 7.2 are used as initial values for the estimation procedure. The model in Table 7.2 has been taken from the literature and is the same model used in Chapter 6. The estimation procedure is performed on a data segment of 2000 nucleotides which constitute the 7000^{th} to the 9000^{th} nucleotide of the original DNA sequence. The

X_{i+1}

X_i	0	1
0	0.9997	0.000245
1	0.1171	0.998829

non-CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	0.300	0.210	0.205	0.285
T	0.177	0.292	0.239	0.292
C	0.322	0.302	0.298	0.078
G	0.248	0.208	0.246	0.298

CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	0.180	0.120	0.274	0.426
T	0.079	0.182	0.355	0.384
C	0.170	0.188	0.368	0.274
G	0.161	0.125	0.339	0.375

Table 7.2: The starting values for the parameters of the CpG island model.

algorithm converges after about 65 iterations giving the parameter estimates listed in Table 7.3.

The resulting estimates are very different from their starting values. The initial transition matrix for the hidden state gives very low probability for switching between states and forces a pattern of long segments of non-island regions interrupted by short segments of CpG islands. After estimation, the transition probabilities are more flexible giving a higher probability for a switch between states than before. Non-island segments are now shorter and are more likely to be interrupted by CpG islands. The probability of being in a non CpG island state is

$$\begin{aligned} P(X_i = 0) &= \frac{P(X_i = 0|X_{i-1} = 1)}{P(X_i = 0|X_{i-1} = 1)P(X_i = 1|X_{i-1} = 1)} \\ &= \frac{0.0138}{(0.0138)(0.0072)} \\ &= 0.6571. \end{aligned}$$

The transition probabilities for the observation sequence also change sharply after estimation, but with ambiguous results. Of primary interest is the transition probability $P(Y_{i+1} = G|Y_i = C)$. In both states this probability is low having a value of 0.0966 in the non-CpG island state and a value of 0.0949 in the CpG island state. These estimate values show that there is little distinction between the two states with regard to a $C \rightarrow G$ transition that characterise a CpG island.

Another important feature of a CpG island is the more frequent occurrence of C and G nucleotides then elsewhere in the DNA sequence. The estimated marginal probabilities for the CpG island state give a high probability of observing a C nucleotide, $P(Y_i = C) = 0.4786$, and relatively low probabilities of observing the remaining nucleotides with $P(Y_i = G) = 0.1602$. $P(Y_i = A) = 0.0774$, and $P(Y_i = T) = 0.2828$. For the non-CpG island state the probability of observing a C and the probability of observing a G are rather close in value with $P(Y_i = C) = 0.3014$ and $P(Y_i = G) = 0.3226$ and are also higher than the probability of observing an A, $P(Y_i = A) = 0.1879$, or a T, $P(Y_i = T) = 0.1880$, nucleotide.

The joint probability, $P(C \text{ followed by } G) = P(Y_i = C|Y_{i-1} = G)P(Y_{i-1} = G)$, for both the non-island state and island state, 0.030 and 0.045 respectively, are also not drastically different in value. This further emphasises that the estimates do not illustrate a clear distinction between the states based on the properties described above. The results can, therefore, be interpreted in one of two ways: either the estimated transition probabilities are good estimates and are exhibiting behavioural aspects characteristic to the sequence and yet invisible to the researcher, or the estimates have reached a very bad local maximum.

	X_{i+1}		
X_i	0	1
0	.9928	.0072
1	.0138	.9862

non-CpG island state

	Y_{i+1}			
Y_i	A	T	C	G
A	.1935	.0899	.2367	.4799
T	.1091	.1879	.3006	.4027
C	.2376	.2696	.3963	.0966
G	.1842	.1684	.2500	.3975

CpG island state

	Y_{i+1}			
Y_i	A	T	C	G
A	.1007	.2729	.2707	.3557
T	.0279	.2306	.5357	.2059
C	.1074	.3481	.4496	.0949
G	.0638	.1878	.5689	.1795

Table 7.3: The Baum-Welch parameter estimates.

7.3 Prequential Estimation

The formulation of Baum-Welch estimates examined in section 7.2 violates the fundamentals of prequential theory and therefore, has no place in the prequential framework. The nature of the Baum-Welch estimation process does not make use of new information as it becomes available. Instead, it requires that the forecaster use all the information simultaneously to estimate the model parameters. The forecaster can then either go back and make "*forecasts*" for what is now essentially the past, or use these estimates to make forecasts for future events without updating the estimates with new information as it becomes available.

7.3.1 Prequential estimation method

In order to generate *prequential forecasts*, it is first necessary to produce *prequential estimates*. In much the same way as a prequential forecast at

time i is based on all the information available at time $i-1$, a prequential estimate used to generate that forecast should also be based only on the information available at time $i-1$. By restricting the information available to it, the prequential estimate will not allow the event in question (or knowledge of future outcomes of events) to contribute in any way to the event's own forecast. It is only as new information becomes available that the prequential parameter estimate is updated to incorporate it.

A major impediment in the formulation of sequentially updated parameter estimates in the Baum-Welch procedure is the forward-backward algorithm (Rabiner, 1989). In the Baum-Welch procedure, the forward-backward algorithm (described in Algorithm 7.2.2 and Algortihm 7.2.3) is a recursive computational algorithm used to derive the conditional expectations summarised in Table 7.1. Various authors in the HMM literature have developed methods of parameter estimation that circumvent the forward-backward algorithm. Elliot et al (1995) describes an online recursive estimation procedure for sequentially updated parameter estimates which avoids the use of the forward-backward algorithm. Baldi and Chauvin (1994) also derived an online estimation approach based on gradient descent techniques.

It is also possible to derive the required expectations while avoiding the use of the forward-backward algorithm using Algorithm 7.3.1 described below. The evaluation and computation of probabilities used in this algorithm will be discussed in greater detail later in the section.

Algorithm 7.3.1 Prequential Estimation

1. *Initialise with a given sequence of length n and evaluate*

$$P(X_i = k, X_{i+1} = j|D_n) \ \forall \ i = 1, \ldots, n-1, \tag{7.9}$$

2. *Evaluate:* $P(X_n = k, X_{n+1} = j|D_{n+1})$.

$$
\begin{array}{ccc}
n = 3 & & n = 4 \\
\left[\begin{array}{l} P(X_1 = k, X_2 = j|D_3) \\ P(X_2 = k, X_3 = j|D_3) \end{array} \right] & \longrightarrow & \left[\begin{array}{l} P(X_1 = k, X_2 = j|D_4) \\ P(X_2 = k, X_3 = j|D_4) \\ P(X_3 = k. X_4 = j|D_4) \end{array} \right] \longrightarrow \\
n = 5 & & \cdots \\
\left[\begin{array}{l} P(X_1 = k, X_2 = j|D_5) \\ P(X_2 = k, X_3 = j|D_5) \\ P(X_3 = k, X_4 = j|D_5) \\ P(X_4 = k, X_5 = j|D_5) \end{array} \right] & \longrightarrow & \cdots
\end{array}
$$

Figure 7.1: The evolution of Equation (7.9) as more data is observed.

3. *Update:* $P(X_i = k, X_{i+1} = j|D_{n+1})\ \forall\ i = 1, \ldots, n-1.$

4. *Compute parameter estimates.*

5. *Return to step 2 until no new data is available.*

At each $n = 2, \ldots, N-1$, the conditional expectations in Table 7.1 must be evaluated. This requires that the probability in equation (7.9) be evaluated for all $i = 1, \ldots, n-1$. From these expectations the values of the prequential parameter estimates can be obtained. Essentially, what the above algorithm generates is illustrated in the Figure 7.1.

For any first order hidden Markov model, the formulas for obtaining and updating the probabilities in Figure 7.1 are derived using Bayes formula. Let $\beta_n = (X_1, \ldots, X_n)$. Then

$$P(X_1 = k, X_2 = j|D_{n+1}) =$$

$$\sum_{X_{n+1}} \frac{P(X_1 = k, X_2 = j|D_n)P(X_{n+1}|\beta_n, D_n)P(Y_{n+1}|\beta_{n+1}, D_n)}{P(Y_{n+1}|D_n)} \quad (7.10)$$

$$P(X_2 = k, X_3 = j|D_{n+1}) =$$
$$\sum_{X_{n+1}} \frac{P(X_2 = k, X_3 = j|D_n)P(X_{n+1}|\beta_n, D_n)P(Y_{n+1}|\beta_{n+1}, D_n)}{P(Y_{n+1}|D_n)} \quad (7.11)$$

$$\vdots$$

$$P(X_i = k, X_{i+1} = j|D_{n+1}) =$$
$$\sum X_{n+1} \frac{P(X_i = k, X_{n+1} = j|D_n)P(X_{n+1}|\beta_n, D_n)P(Y_{n+1}|\beta_{n+1}, D_n)}{P(Y_{n+1}|D_n)} \quad (7.12)$$

$$\vdots$$

$$P(X_{n-1} = k, X_n = j|D_{n+1}) =$$
$$\sum_{X_{n+1}} \frac{P(X_{n-1} = k, X_n = j|D_n)P(X_{n+1}|\beta_n, D_n)P(Y_{n+1}|\beta_{n+1}, D_n)}{P(Y_{n+1}|D_n)} \quad (7.13)$$

$$P(X_n = k, X_{n+1} = j|D_{n+1}) =$$
$$\sum_{X_{n+1}} \frac{P(X_n = k|D_n)P(X_{n+1}|\beta_n, D_n)P(Y_{n+1}|\beta_{n+1}, D_n)}{P(Y_{n+1}|D_n)}. \quad (7.14)$$

The conditional independence properties of the HMM:

$$Y_{n+1} \perp\!\!\!\perp (Y_1, \ldots, Y_n, X_1, \ldots, X_n)|X_{n+1},\ Y_n$$

and

$$X_{n+1} \perp\!\!\!\perp (Y_1, \ldots, Y_n, X_1, \ldots, X_{n-1}\}|X_n,$$

reduce equations (7.10) - (7.13) to the form given below:

$$P(X_1 = k, X_2 = j|D_{n+1}) =$$
$$\sum_{X_{n+1}} \frac{P(X_1 = k, X_2 = j|D_n)P(X_{n+1}|X_2, D^3)P(Y_{n+1}|X_{n+1}, Y_n)}{P(Y_{n+1}|D_n)} \quad (7.15)$$

$$P(X_2 = k, X_3 = j|D_{n+1}) =$$
$$\sum_{X_{n+1}} \frac{P(X_2 = k, X_3 = j|D_n)P(X_{n+1}|X_3, D^4)P(Y_{n+1}|X_{n+1}, Y_n)}{P(Y_{n+1}|D_n)} \quad (7.16)$$

$$\vdots$$

$$P(X_i = k, X_{i+1} = j|D_{n+1}) =$$
$$\sum_{X_{n+1}} \frac{P(X_i = k, X_{i+1} = j|D_n)P(X_{n+1}|X_{i+1}, D^{i+2})P(Y_{n+1}|X_{n+1}, Y_n)}{P(Y_{n+1}|D_n)} \quad (7.17)$$

$$\vdots$$

$$P(X_{n-1} = k, X_n = j|D_{n+1}) =$$
$$\sum_{X_{n+1}} \frac{P(X_{n-1} = k, X_n = j|D_n)P(X_{n+1}|X_n)P(Y_{n+1}|X_{n+1}, Y_n)}{P(Y_{n+1}|D_n)} \quad (7.18)$$

$$P(X_n = k, X_{n+1} = j|D_{n+1}) =$$
$$\frac{P(X_n = k|D_n)P(X_{n+1} = j|X_n = k)P(Y_{n+1}|X_{n+1} = j, Y_n)}{P(Y_{n+1}|D_n)}, \quad (7.19)$$

where $D^i = (Y_i, \ldots, Y_n)$.

In implementation, only equation (7.19) can be computed directly. The remaining probabilities, (7.15)-(7.18), require the use of parallel recursive computations to incorporate the new information sequentially as it becomes available. To perform these computations, Algorithm 7.3.2 is used in parallel with Algorithm 7.3.1 to keep track of $P(X_n|X_i, D^{i+1})$. The algorithm is given below.

Algorithm 7.3.2

1. $n = 3$,
2. $$a_n(n-2) = P(X_n|X_{n-1} = j).$$
3. *If* $n = N - 1$ *exit algorithm. otherwise* $n = n + 1$.

$$\mathbf{a}_{n=3} = \left[\; P(X_3|X_2 = j) \; \right] \quad \longrightarrow \quad \mathbf{a}_{n=4} = \left[\begin{array}{l} P(X_4|X_2 = j, Y_3) \\ P(X_4|X_3 = j) \end{array} \right] \quad \longrightarrow$$

$$\mathbf{a}_{n=5} = \left[\begin{array}{l} P(X_5|X_2 = j, Y_3, Y_4) \\ P(X_5|X_3 = j, Y_4) \\ P(X_5|X_4 = j) \end{array} \right] \quad \longrightarrow \quad \ldots$$

Figure 7.2: The development of $\mathbf{a}$ as n increases.

4. Recursion:

$$a_n(i) = \sum_{X_{n-1}} \frac{a_{n-1}(i) P(X_n|X_{n-1}) P(Y_n|X_n, Y_{n-1})}{P(Y_n|D_n)} \quad \forall \; i = 1, \ldots, n-3.$$

5. go to 2.

Algorithm 7.3.2 produces a vector, $\mathbf{a}$, containing the correct formulation of $P(X_n|X_i, D^{i+1})$ required to update

$$P(X_i = k, X_{i+1} = j|D_n) \; \forall \;\; i = 1, \ldots, n-1. \tag{7.20}$$

Figure 7.2 illustrates the development of $\mathbf{a}$ as n becomes larger.

7.3.2 Implementation and results

Unlike the Baum-Welch method, the prequential method, as described in Algorithm 7.3.1 does not allow the estimates to converge on a fixed set of data. With each new observation, the algorithm performs only one *Expectation* step and one *Maximisation* step. Ideally, the algorithm should allow the estimates to converge every time a new observation is incorporated. In

implementation, however, this is not possible due to computational limitations.

Prequential estimation is performed on the same data sequence used in section 7.2.3 (nucleotides $7000-9000$). The first 1000 nucleotides are used to initialise the values of (7.20) so that $n = 7999$ and $D_{7999} = \{Y_{7000}, \ldots, Y_{7999}\}$. Two different starting values for the parameters are used:

1. Using maximum likelihood estimates as starting values.

2. Using the parameter values in Table 7.2.

In the first scenario, the Baum-Welch algorithm is applied to the initial information base, D_{7999} using model 7.2 for starting values. This results in parameter estimates that have converged to a local maximum of the likelihood function of the first 1000 nucleotides of the sequence. The mle's are then used to compute the initial values of the probabilities in (7.20). The remaining 1000 observations are then added sequentially to the process as the estimates are updated using prequential estimation. Table 7.4 gives the maximum likelihood starting values and Table 7.5 gives the final values of the estimated transition probabilities after performing prequential estimation on the remaining 1000 observations.

The prequential model in Table 7.5 shows that the state transition probabilities have changed dramatically. Contrary to expectation, the model describes a DNA sequence with long stretches of CpG island segments interrupted by short and very infrequent non-CpG island regions.

Similar to the Baum-Welch results in Table 7.3, the observation transition probabilities show no clear distinction between the two states. The probability of a $C \rightarrow G$ transition in the non-CpG island state is slightly higher than in the island state, with $P(Y_i = C|Y_{i-1} = G, X_i = 1) = 0.0956$ and

X_{i+1}

X_i	0	1
0	.9805	.0195
1	.0084	.9916

non-CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	.2228	.2696	.2583	.2493
T	.0633	.2657	.4949	.1761
C	.1467	.2300	.3125	.0753
G	.2940	.4656	.3215	.1508

CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	.1867	.0776	.2256	.5100
T	.1261	.1796	.2789	.4105
C	.2565	.2360	.4396	.0679
G	.1845	.1442	.2530	.4183

Table 7.4: Baum-Welch estimates using the first 1000 observations only. The starting values for the prequential estimation procedure.

X_{i+1}

X_i	0	1
0	.1910	.8090
1	.0001	.9999

non-CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	.2050	.1569	.2519	.3862
T	.0616	.2705	.4306	.2373
C	.1218	.3780	.3685	.1317
G	.0934	.3675	.2093	.3298

CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	.1745	.1215	.2442	.4598
T	.0726	.2025	.4077	.3172
C	.1771	.3046	.4227	.0956
G	.1587	.1721	.3186	.3506

Table 7.5: The final prequential model estimates using mle starting values.

X_i \ X_{i+1}

	0	1
0	1	0
1	.0188	.9812

non-CpG island state

Y_i \ Y_{i+1}

	A	T	C	G
A	.1742	.1255	.2443	.4560
T	.0730	.2076	.4061	.3134
C	.1771	.3072	.4211	.0945
G	.1583	.1739	.3191	.3487

CpG island state

Y_i \ Y_{i+1}

	A	T	C	G
A	.1163	.0649	.2685	.5503
T	.0163	.1472	.6283	.2081
C	.0647	.2439	.5915	.0999
G	.0573	.1021	.6053	.2352

Table 7.6: The final prequential model estimates derived using the parameter values in Table 7.2 as starting values.

$P(Y_i = C|Y_{i-1} = G, X_i = 0) = 0.1317$. This also contradicts expectations.

In the second scenario, the information base, D_{7999}, is used to compute equation (7.20) without performing any form of estimation. The transition probabilities in model 7.2 are the starting parameter values used to perform these computations. The estimation process begins when prequential parameter estimation is performed on the remaining 1000 nucleotides of the segment. The results listed in Table 7.6 show the prequential estimates after they have been updated with the last observation.

This scenario also produces estimates that are *unacceptable*. As observations are incorporated in the estimation process, the value of the state transition probability, $P(X_{i+1} = 0|X_i = 0)$, fluctuates repeatedly from 1 to 0.9999. Eventually as the last few observations are added, the state transition matrix gives a probability of 0 for the transition from a non-CpG island

region to an island region. This means that unless the sequence begins with a CpG island segment then there is zero probability of a region occurring.

The observation transition matrices, like the observation transition probabilities estimated before, show no evident distinction between the two states.

7.3.3 Validation

Even though the estimates evaluated prequentially do not look promising, it still remains to be determined what exactly the prequential procedure is producing. It is clear from visual examinations of the evolution of the estimates that the prequential estimation method is highly sensitive to the data. The re-evaluated probabilities change continuously with each new addition to an extent that a state transition probability of zero (such as the case in Table 7.6) would actually increase in value given more data. The resulting probabilities are, therefore, subject to the last observation included in the estimation process. This is not unusual since this is also the case with other *prequential* estimation methods (i.e. recursive least squares, Kalman filter, etc...).

These other methods, however, are known to produce estimates that maximise the likelihood of the data used to evaluate them. This is not known to be the case for the prequential estimation procedure for the HMM. Due to the nature of the estimation process, one *Expectation* step and one *Maximisation* step with each new observation, it is difficult for the estimates to converge.

To determine if they are maximum likelihood estimates, the estimates obtained using prequential estimation, Table 7.5 and Table 7.6, are allowed to iterate using Buam-Welch until *convergence* is reached. The results of the iterations on the prequential estimates in Table 7.5 and Table 7.6 are

given in Table 7.7 and Table 7.8, respectively. The estimates in Table 7.5 are derived using mle as starting values. Regardless of this fact, the parameter values only converged after more than 1500 iterations. On the other hand, the transition probabilities in Table 7.6 converged to the values in Table 7.8 in less than 25 iterations.

Comparison of the converged prequential estimates with the Baum-Welch estimates shows some interesting results. All three models have state transition probabilities that are remarkably close in value. The same, however, cannot be said about the nucleotide's transition probabilities in the two states. The parameter values in Table 7.7 that used mle starting values are very different from their Baum-Welch counterparts and yet the converged estimates in Table 7.8 are more or less in the same vicinity. These results indicate that the likelihood function is not unimodal and, depending on the starting value, a different local maximum is delivered by the estimation algorithm.

To test this assumption, the Baum-Welch algorithm is executed using different starting values. The resulting parameter estimates are then compared to determine if the likelihood function of the designated data sequence is in fact unimodal. The results of this test run is given: Table 7.9 lists the starting values of the parameters and Table 7.10 displays the converged estimates. In this case, even the transition probabilities for the state are notably different. The results strongly support the assumption that the likelihood function has more than one mode.

Hence, not only are the estimated values sensitive to the data sequence used, but they are also largely determined by the starting values used to initialise the estimation process.

	X_{i+1}	
X_i	0	1
0	.9597	.0403
1	.0151	.9849

non-CpG island state

	Y_{i+1}			
Y_i	A	T	C	G
A	.1608	.0321	.3248	.4822
T	.2221	.1171	.1049	.5581
C	.2842	.1705	.4068	.1385
G	.2085	.1352	.1683	.4876

CpG island state

	Y_{i+1}			
Y_i	A	T	C	G
A	.1852	.1848	.1907	.4392
T	.0450	.2235	.4623	.2692
C	.1505	.3392	.4249	.0855
G	.1227	.1980	.4245	.2548

Table 7.7: The converged prequential model estimates of Table 7.5.

	X_{i+1}	
X_i	0	1
0	.9936	.0064
1	.0180	.9820

non-CpG island state

	Y_{i+1}			
Y_i	A	T	C	G
A	.1963	.1150	.2264	.4624
T	.1002	.2110	.3128	.3760
C	.2383	.2795	.3897	.0924
G	.1777	.1745	.2632	.3845

CpG island state

	Y_{i+1}			
Y_i	A	T	C	G
A	0	.2105	.3874	.4021
T	.0161	.1996	.5962	.1880
C	.0803	.3483	.4705	.1010
G	.0525	.1630	.6193	.1653

Table 7.8: The converged prequential model estimates of Table 7.6.

X_{i+1}

X_i	0	1
0	.8090	.1910
1	.25	.75

non-CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	.25	.25	.25	.25
T	.25	.25	.25	.25
C	.25	.25	.25	.25
G	.25	.25	.25	.25

CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	.25	.25	.25	.25
T	.25	.25	.25	.25
C	.25	.25	.25	.25
G	.25	.25	.25	.25

Table 7.9: The starting values for the test run.

X_{i+1}

X_i	0	1
0	.0036	.9964
1	.5545	.4455

non-CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	.2831	.1961	.2905	.2304
T	.0268	.4537	.0558	.4637
C	.1571	.5967	.2462	.0000
G	.0014	.3169	.3890	.2927

CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	.1052	.0784	.2123	.6041
T	.0817	.1547	.4826	.2810
C	.1898	.1065	.5420	.1617
G	.2701	.0686	.2714	.3899

Table 7.10: The converged estimates for the test run after 1200 iterations of the Baum-Welch algorithm.

7.4 Performance

This section examines how calibration is affected by estimation. Table 7.11 gives a summary of the calibration results for the forecasts based on both the Baum-Welch model in Table 7.3 and the prequential model described in Table 7.6.

As mentioned earlier, the Baum-Welch algorithm is executed on a segment of the DNA sequence, nucleotides 7000-9000. The estimates are then used to generate forecasts for the state of the system of nucleotide 9001 to nucleotide 10500. This is to ensure that the estimates used to generate a forecast do not embody information about future observations.

For fixed intervals of p_i, the average of the q_i computed using the Baum-Welch estimates is close in value to the midpoint of p_i. As indicated by the straight line through the calibration plot in Figure 7.3, the calibration of these estimation based forecasts are good.

For the prequential estimation procedure, nucleotide 7000 to 7999 are used only to form the information base of (7.20). No estimation is performed on the first 1000 nucleotides. Prequential estimation begins with the 8000^{th} nucleotide and the algorithm is allowed to run continuously to the nucleotide in position 10500. This corresponds to the second implementation scenario of section 7.3.2. In order to compare the calibration results of the prequential forecasts with those of the Baum-Welch model, only the last 1500 forecasts are used.

In the prequential framework with each new observation come new updated parameter estimates. Hence, a forecast for the probability of CpG island at position i is made with different parameter values than those used to generate the forecast at position $i+1$. As the results in Table 7.11 show, the forecasts are very consistent showing a very low probability of a CpG is-

p_i Range	Literature		Baum-Welch		Prequential	
	n	Average q	n	Average q	n	Average q
$0 \leq p_i \leq 0.05$	274	0.0161	65	0.0419	1493	0.001
$0.05 \leq p_i \leq 0.15$	74	0.0939	260	0.1023	6	0.0743
$0.15 \leq p_i \leq 0.25$	51	0.2329	163	0.2012	1	0.0597
$0.25 \leq p_i \leq 0.35$	79	0.3198	116	0.3017	0	n/a
$0.35 \leq p_i \leq 0.45$	80	0.4195	105	0.3986	0	n/a
$0.45 \leq p_i \leq 0.55$	78	0.4917	86	0.4962	0	n/a
$0.55 \leq p_i \leq 0.65$	81	0.5846	90	0.6189	0	n/a
$0.65 \leq p_i \leq 0.75$	94	0.7191	137	0.7120	0	n/a
$0.75 \leq p_i \leq 0.85$	147	0.7915	260	0.8073	0	n/a
$0.85 \leq p_i \leq 0.95$	147	0.8899	205	0.8817	0	n/a
$0.95 \leq p_i \leq 1.00$	395	0.9779	13	0.9577	0	n/a
total	1500		1500		1500	

Table 7.11: The calibration results obtained using the literature based model, the Baum-Welch model and the prequential model.

land occurrence. This is due primarily to the fact that throughout the length of the last 1500 observations the probability of a transition from non-CpG island state to an island state remains constant at zero. As the results in Table 7.11 and the calibration plot in Figure 7.4 show, calibration in the forecasts generated from prequential estimates is not good.

In the Baum-Welch case calibration is good, but does not seem to be improved by estimation. The prequential case presents a situation where calibration becomes worse with an estimated model. Table 7.11 shows that the estimation of the model drastically changes the transition probabilities and thus the forecast values.

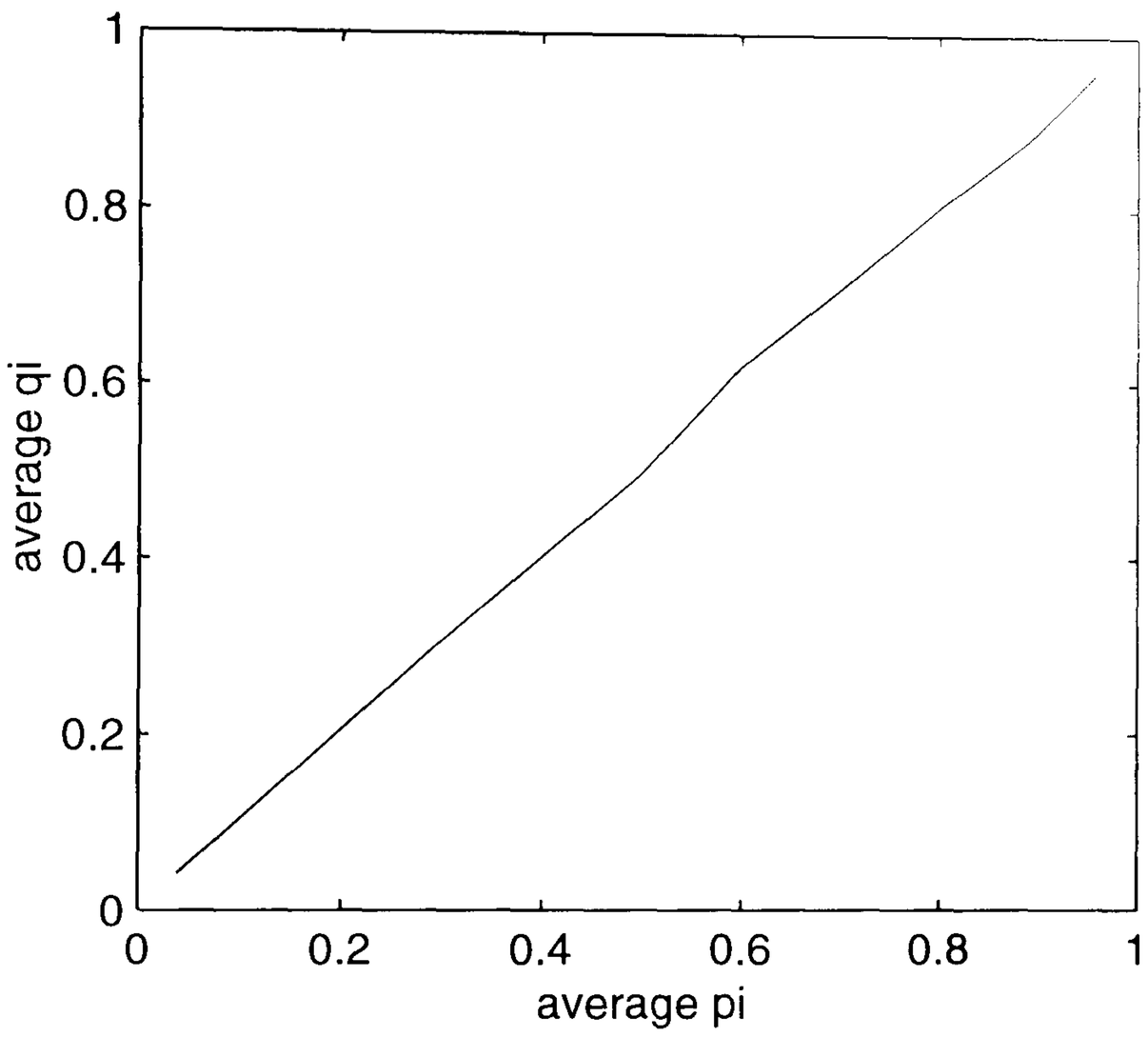

Figure 7.3: Calibration plot of the average q_i at fixed value of p_i for the Baum-Welch model.

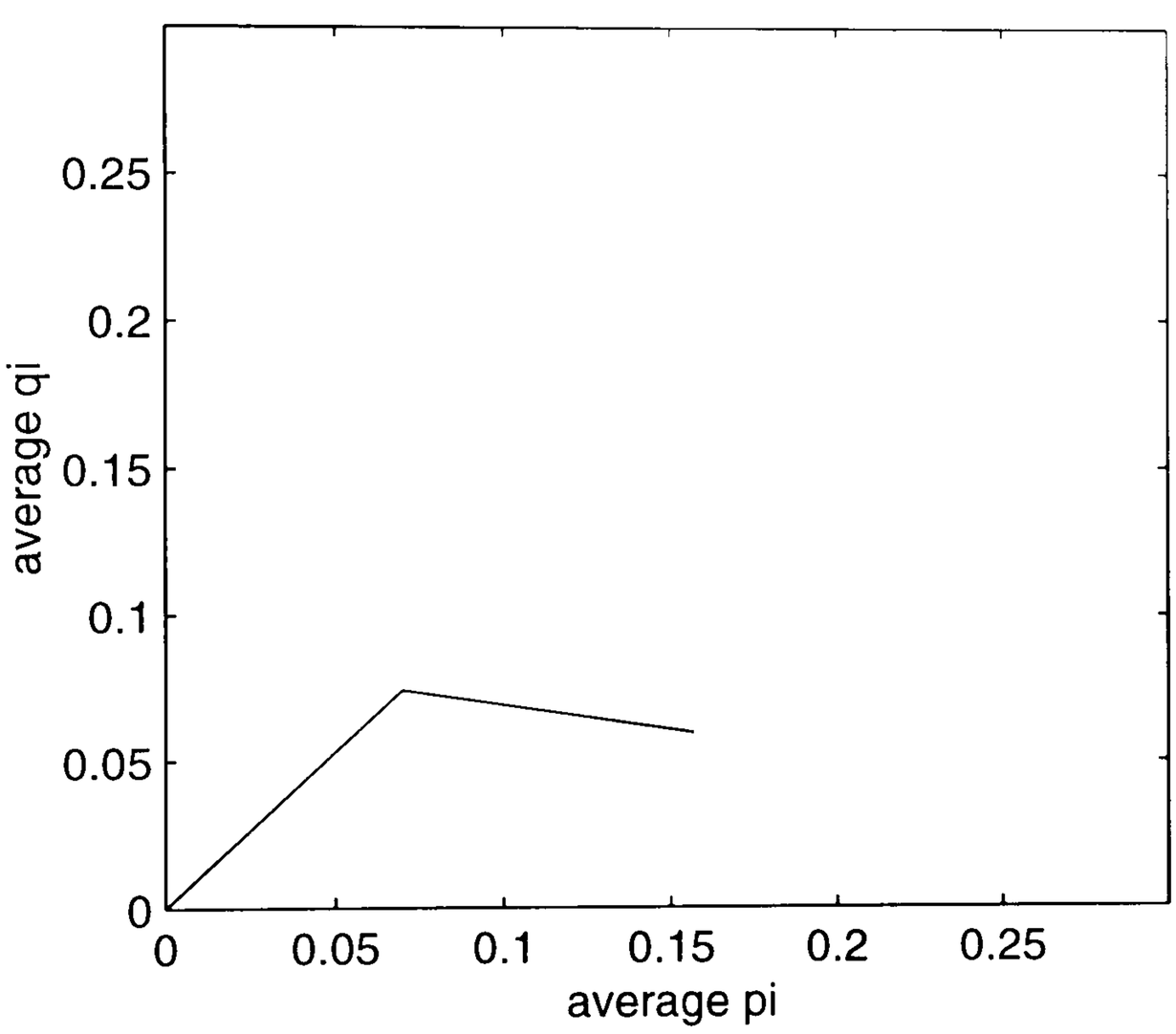

Figure 7.4: Calibration plot of the average q_i at fixed value of p_i for the prequential model.

7.5 Discussion

The investigation in this chapter attempts to explore the affects of estimation on forecasting validation. Estimates are derived using the Baum-Welch estimation procedure and to remain consistent with the theoretical concepts of Chapter 5 and Chapter 6, a prequential estimation procedure is also introduced. Both methods are iterative re-estimation procedures. In the Baum-Welch case, the resulting estimates, having converged at a local maximum, are local maximum likelihood estimates. The prequential procedure operates by incorporating each data point sequentially as it is observed. Ideally the prequential algorithm should iterate until it converges (maximisation) for each new observation. Due to computational limitations this is not possible. Hence, the prequential algorithm as it is applied in this chapter does not produce estimates that have converged to a local maximum. In application, it is found that the likelihood is not unimodal and, therefore, the estimates, whether derived by the Baum-Welch procedure or the prequential method, are subject to the initial values of the estimates.

There is no evidence here to suggest that estimation improves calibration. The calibration of the forecasts produced using the Baum-Welch estimates, although good, do not seem to improve on the calibration of the forecasts produced without the estimation of transition probabilities. The prequential estimates, by giving a probability of zero for a transition from a CpG island state to a non-CpG island state, give little room for the exploration of calibration behaviour under this type of estimation procedure. For what results that are obtained, the forecasts produced using prequential estimates show very poor calibration.

Chapter 8

Smoothed Predictions

8.1 Introduction

The prequential principle forms the foundation for much of the work presented in this thesis and is the predominant characteristic in the definition and formulation of the forecasts used in this work and the methods used to assess them. The forecast of an event at time i is formulated in such a way as to include only information available up to the time of the forecast, specifically all the information available up to and including time $i-1$.

In the HMM case, since the observation is not available to make a proper assessment of the forecast made, it is replaced with an approximation. This approximation is the updated forecast $q_i = P[X_i = 1|D_{i-1}, Y_i]$. In the previous chapters assessment of the forecast, $p_i = P[X_i = 1|D_{i-1}]$, is performed by comparing the forecast with the update. This method of assessment, up to now, has in a sense also been prequential limiting the knowledge of additional information for the forecast assessment to only the next sequential observation.

Although q_i adheres to the prequential principle and its many advantages,

it is mostly likely not the best replacement for the outcome of X_i. A much better approximation would make use of *all* the information (both past and future observations) such as the smoothed prediction $s_i = E[X_i|D_N]$.

The use of a more informative approximation of X has implications on the assessment of the forecasts. For example in Chapter 5 and Chapter 6 the performance of the p_i forecasts was measured by its overall empirical calibration with its q_i update. This is in essence the gauging of one forecast by another, albeit slightly more informative, forecast. Although s_i is also a prediction it contains all the information available to the forecaster. It is, therefore, safe to assume that s_i will make a more *truthful* probability statement about the outcome of X_i than q_i. Hence, comparison of the forecast with the smoothed prediction s_i will give a more insightful and critical assessment of the forecast's performance.

The aim of this chapter is twofold. The first is to provide a more critical evaluation of forecast performance. This is done by assessing the empirical calibration of the p_i forecasts with the s_i predictions. Superficially such a construction is expected to yield more viable results. Unfortunately, no theory is available to substantiate this claim.

The prequential calibration methods introduced in Chapter 5 and explored in Chapter 6, although limited in their use of information, have the advantage of having a strong theoretical basis detailing the behaviour of a well calibrated forecast from which conclusions about the forecasts can be drawn. Central to this theoretical foundation (the complete calibration criterion in Chapter 5 and the calibration test statistic Chapter 6) is the martingale difference property:

$$E[(p_i - q_i)|D_{i-1}] = 0. \tag{8.1}$$

Since $p_i - q_i$ is D_i-measurable, the sequence of such quantities $(p_i - q_i)$, for

$i = 1, \ldots, n$, forms a martingale difference series. For the smoothed case the conditional expectation,

$$E\left[(s_i - p_i)\,|D_{i-1}\right], \tag{8.2}$$

is also equal to 0. However, since s_i is a not a prequential estimate (i.e. not D_i-measurable) the series of $(s_i - p_i)$ quantities does not form a martingale difference series. The theory established in the previous chapters can not be used in this application, therefore, the $(p_i,\ s_i)$ calibration in section 8.4 will be assessed on a purely empirical basis.

The second aim of this chapter is the analysis of calibration outside of the prequential framework in which it was presented in the previous chapters. Remaining within the calibration structure designed for hidden Markov models, the analysis still calibrates the forecast with a more knowledgeable prediction; however, in this case both the forecasts and prediction make use of both past and future observations. Cross-validation, a method of comparing s_i with c_i, the forecast evaluated by using all but the i^{th} observations in the sequence, presents itself as an obvious choice for the implementation of such a calibration scheme. Section 8.5 investigates the calibration of $(s_i,\ c_i)$ in a manner similar to Chapter 6 and entails, firstly, examination of the empirical calibration and then the analysis of a purposed test statistic.

8.2 The Data

The CpG island example is again used in this chapter to illustrate the assessment techniques described. The data set used throughout this chapter is sequence of 5000 nucleotides simulated from the transition probabilities listed in Table 8.1. The variable of interest is the state of the system at time i, X_i, which denotes the presence, $X_i = 1$, or absence, $X_i = 0$, of a CpG

X_{i+1}

X_i	0	1
0	0.99755	0.00245
1	0.1171	0.998829

non-CpG island state

Y_{i+1}

Y_i	A	T	C	G
A	0.300	0.210	0.205	0.285
T	0.177	0.292	0.239	0.292
C	0.322	0.302	0.298	0.078
G	0.248	0.208	0.246	0.298

CpG island state

Y_{i+1}

Y_i	A	T	G	C
A	0.180	0.120	0.274	0.426
T	0.079	0.182	0.355	0.384
C	0.170	0.188	0.368	0.274
G	0.161	0.125	0.339	0.375

Table 8.1: The transition probabilities used to simulate the CpG island data.

island. The transition probabilities in Table 8.1 are identical to those used in Chapter 6 save for $P(X_{i+1} = 1|X_i = 0)$ and $P(X_{i+1} = 0|X_i = 0)$ which have been altered slightly to allow for more activity in the sequence, hence making a more interesting case study.

One of the advantages of using a simulated sequence is that the *hidden* sequence of states is no longer hidden. The DNA sequence is constructed by first simulating a state sequence X using the state transition matrices in Table 8.1. The Y_i's are then simulated from one of the two observation transition matrices conditional on the value of the corresponding X_i. The realised sequence of X's is shown in Figure 8.1.

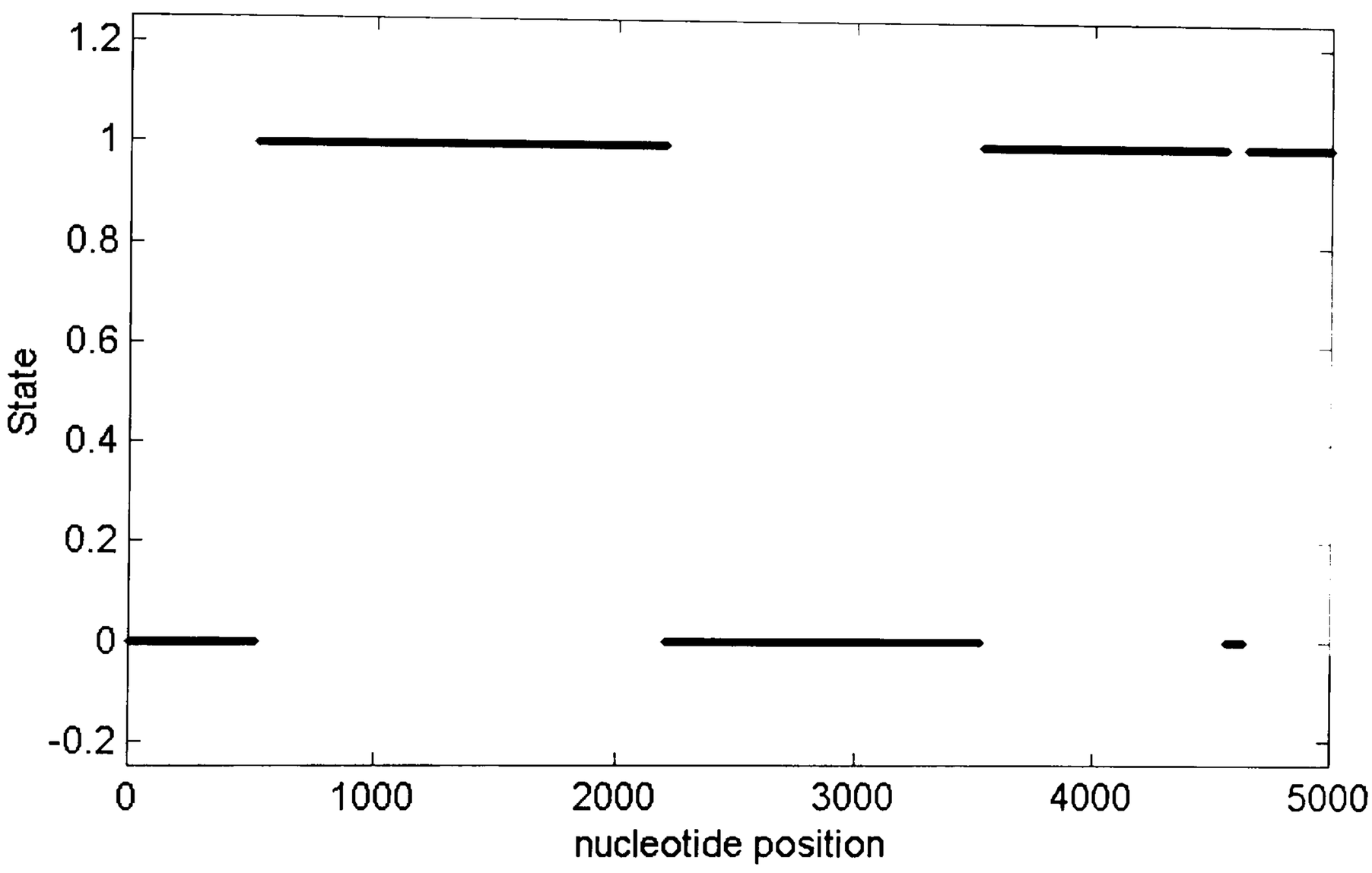

Figure 8.1: The simulated state sequence.

8.3 Computing s_i

The smoothed prediction,

$$s_i = P\left(X_i = 1 | Y_1, \ \ldots, \ Y_n\right), \tag{8.3}$$

can be evaluated using the forward-backward algorithm discussed in Chapter 7. For each time point i, the forward algorithm calculates

$$f_i(0) = P\left(Y_1, \ \ldots, \ Y_i, X_i = 0\right) \quad \text{and} \quad f_i(1) = P\left(Y_1, \ \ldots, \ Y_i, X_i = 1\right).$$

and the backward algorithm evaluates

$$b_i(0) = P\left(Y_{i+1}, \ \ldots, \ Y_N | X_i = 0, Y_i\right) \quad \text{and} \quad b_i(1) = P\left(Y_{i+1}, \ \ldots, \ Y_N | X_i = 1, Y_i\right).$$

Using the output from these two algorithms s_i can be easily evaluated once expressed as

$$P\left(X_i = 1 | Y_1, \ \ldots, \ Y_N\right) = \frac{f_i(1) b_i(1)}{\left(f_i(0) b_i(0) + f_i(1) b_i(1)\right)}.$$

Figure 8.2 shows a plot of the smoothed predictions together with plots of the p_i forecasts and q_i updates. The probability forecasts of both p's and the q's roughly follow the state transitions throughout the sequence giving a fairly general indication of where the CpG islands are located. The smoothed predictions are much more refined. With the majority of predictions very close in value to either 0 or 1, the s's give a very clear picture of what is happening in the sequence. It is only when there is a change in state that the smoothed predictions stray from the $\{0, \ 1\}$ extreme showing any reasonable measure of uncertainty.

8.4 Calibration

This section analyses the performance of the p_i forecasts by assessing the calibration of these forecasts with their smoothed counterparts, the s_i's. By

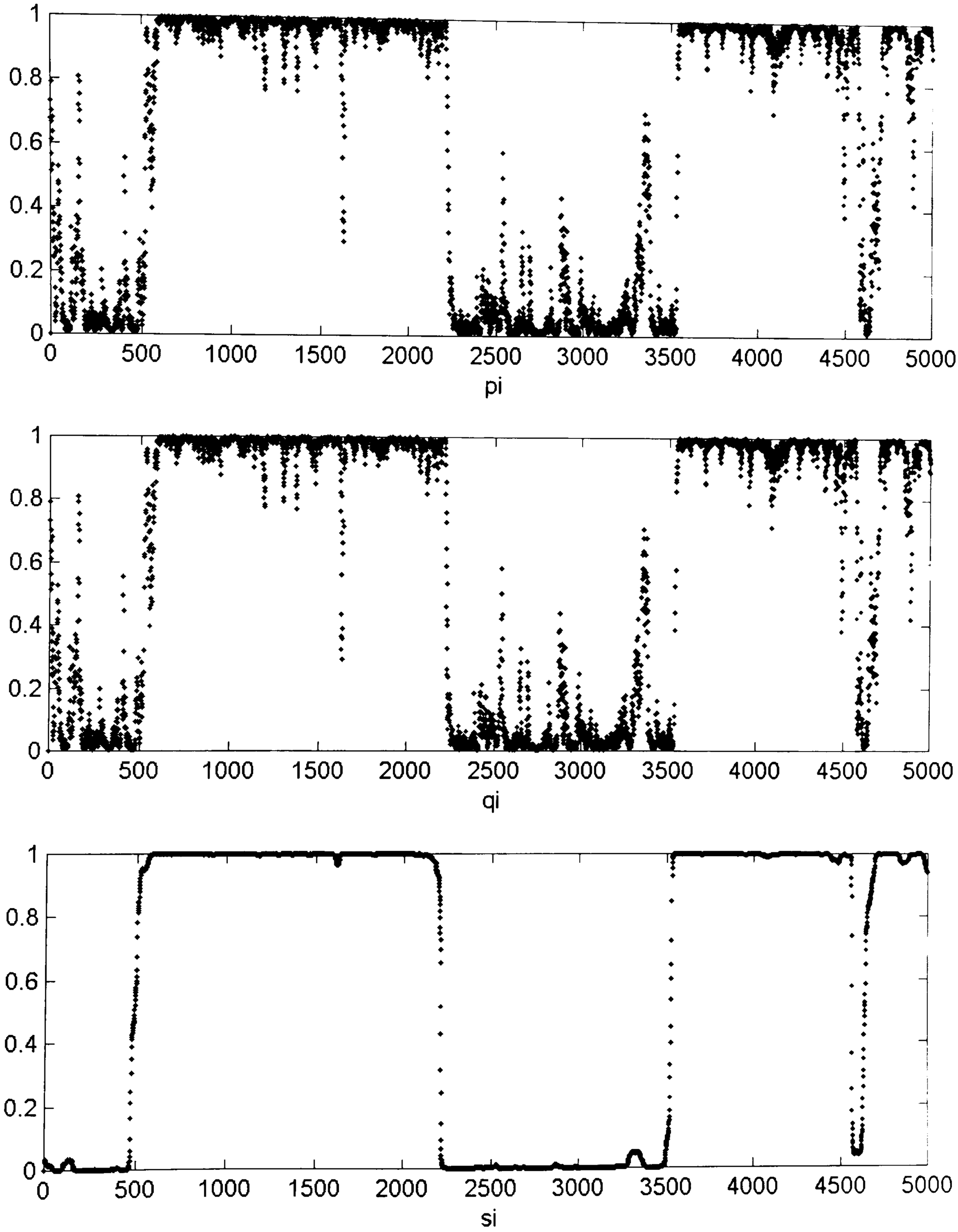

Figure 8.2: Plots of the p_i's, q_i's, and s_i's.

using the smoothed predictions for calibration all the observations in the sequence are used in the formulation of this prediction. This prediction is believed to be the best possible estimate of the hidden state. Such a method is expected to give a more precise assessment of the forecasts' performance.

As mentioned in the introduction of this chapter the martingale quality that characterised the previous calibration apparatus does not hold when replaced by the smoothed prediction. Therefore, the empirical results obtained in this section can not be supported by any theory that can detail the expected or asymptotic behaviour of such a calibration.

As the plots in Figure 8.2 in the previous section show, the s_i's at times make very different statements about the presence of a CpG island than their corresponding q_i's. By comparing the plots in Figure 8.2 with the plot of Y_i in Figure 8.1 it is also clear that the s_i out perform the q_i and give a much clearer indication of the actual value of Y_i.

Table 8.2 gives the results of the calibration, and the calibration plot of the average s_i against the average p_i for fixed intervals of p_i is shown in Figure 8.3. The plot shows that the smoothed prediction and forecasts are well calibrated toward the 1 end, but stray, slightly, everywhere else. It is also worth noting that the number of p_i forecasts with a value between 0.2 and 0.6 is relatively low at 421. This is a possible explanation for the poorly calibrated forecasts within this range. In contrast, the calibration line of the p_i and q_i is almost perfect.

8.5 Cross-Validation

Cross validation is a method commonly used to assess the predictive capability of a forecasting system. In concept, cross validation, introduced by Stone

Interval	Average p	Average q	Average s	n
$0 \leq p_i \leq 0.05$	0.0250	0.0243	0.0181	1084
$0.05 \leq p_i \leq 0.15$	0.0843	0.0856	0.0431	482
$0.15 \leq p_i \leq 0.25$	0.1878	0.1949	0.1324	170
$0.25 \leq p_i \leq 0.35$	0.2941	0.2814	0.1915	89
$0.35 \leq p_i \leq 0.45$	0.4021	0.3849	0.4231	58
$0.45 \leq p_i \leq 0.55$	0.5058	0.5037	0.4797	52
$0.55 \leq p_i \leq 0.65$	0.6018	0.5920	0.5169	52
$0.65 \leq p_i \leq 0.75$	0.6983	0.6728	0.6736	44
$0.75 \leq p_i \leq 0.85$	0.8041	0.8160	0.8984	79
$0.85 \leq p_i \leq 0.95$	0.9150	0.9224	0.9735	394
$0.95 \leq p_i \leq 1.00$	0.9842	0.9844	0.9927	2493
overall	0.5002	0.4965	0.5047	4997

Table 8.2: The calibration results of p_i forecasts with the smoothed prediction s_i.

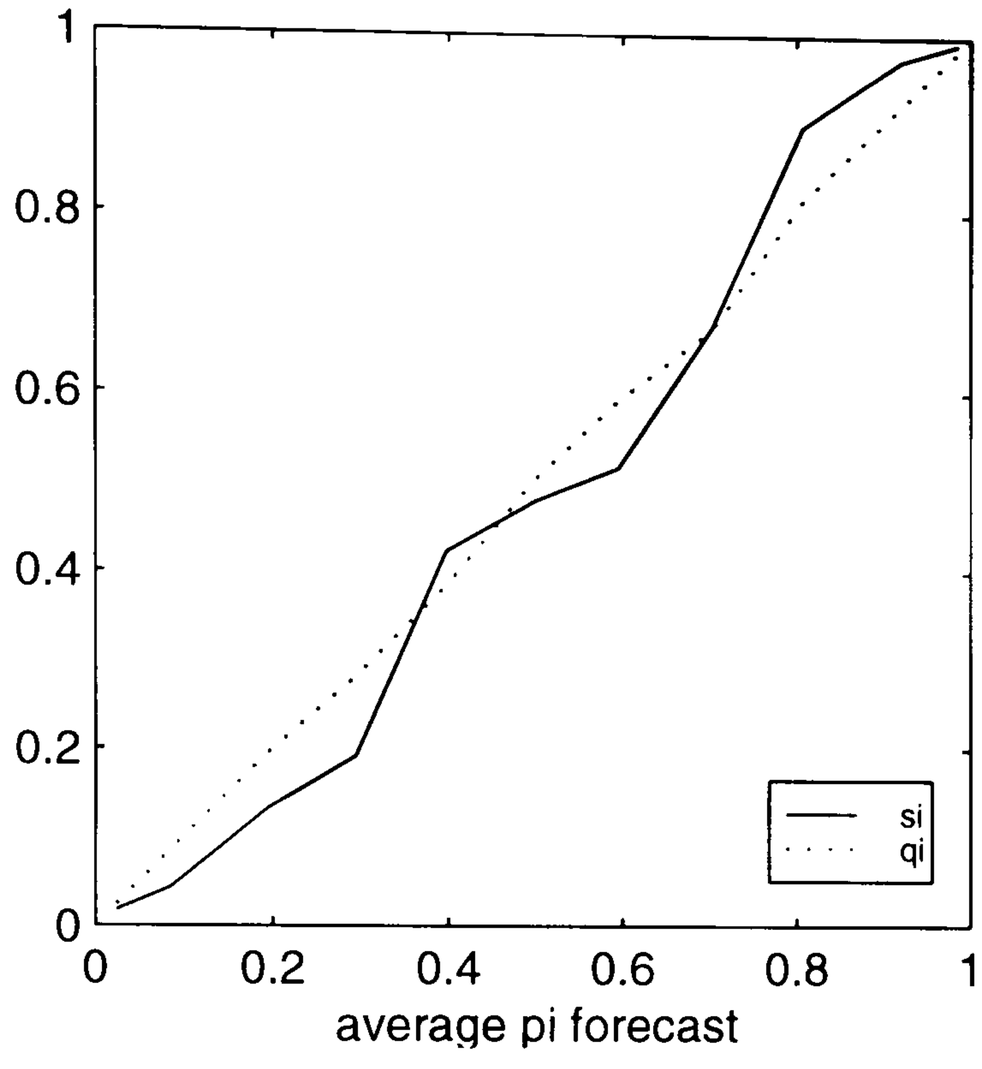

Figure 8.3: Calibration plot.

(1974), is the comparison of the outcome of an event with the prediction of that event such that the outcome of the event in question does not play any role in its own prediction. Letting y_i denote the outcome of an event Y_i at time i, the prediction required by cross validation is

$$c_i = P\left(Y_i = 1 | y_1, \ \ldots, \ y_n, \ {}^{-}y_i\right) \tag{8.4}$$

where ${}^{-}y_i$ denotes the absence of y_i. Once the c_i's are computed for the entire sequence, the average mean squared error for cross validation (MSCV),

$$MSCV = \frac{\sum_{i=1}^{n} \left(y_i - c_i\right)^2}{n}, \tag{8.5}$$

can be evaluated. Using the MSCV the predictive performance of competing models can be compared and assessed. Although it is not a prequential prediction, c_i is considered a *fair* prediction because y_i does not contribute in any way to its own prediction. In this light, cross validation is a fair assessment of a forecasting system making the distinction between a forecasting

system that provides a good fit to the data and a forecasting system that predicts well.

For a HMM, c_i is defined as

$$c_i = P(X_i = 1|Y_1, \ldots, Y_N, \neg Y_i).$$

To compute c_i the forward-backward algorithm must first be used to compute the following probabilities:

$$f_i(k,\ o) = P(Y_1,\ \ldots,\ Y_i = o,\ X_i = k), \tag{8.6}$$

$$b_i(k,\ o) = P(Y_{i+1},\ \ldots,\ Y_N|X_i = k,\ Y_i = o), \tag{8.7}$$

where $o \in \{A,\ T,\ G,\ C\}$ and $k \in \{0,\ 1\}$. Using (8.6) and (8.7) c_i can be expressed as

$$c_i = \frac{\sum_o f_i(1, o)\, b_i(1,\ o)}{\sum_k \sum_o f_i(k, o)\, b_i(k,\ o)}.$$

8.5.1 Calibration

If the prequential framework for calibration is to be abandoned, then the closest alternative to such a framework is cross-validation. This is because the cross-validation concept remains true to the nature of the prequential framework in that the event in question is not allowed to contribute to its own prediction, thereby providing a fair assessment.

In the cross-validation calibration scenario, the smoothed predictions are used in place of the observed outcome of an event X and the c_i are used as the forecast of that event. The behaviour of well calibrated forecasts is defined by the conditional expectation,

$$E\left[s_i - c_i|Y_1,\ \ldots,\ Y_n,\ \neg Y_i\right] = 0. \tag{8.8}$$

Equation (8.8) holds since $E[s_i|Y_1, \ldots, Y_n, {}^{-}Y_i] = c_i$ and it is expected that the well calibrated forecasts will show this. As discussed in section 8.4, the sequence of s_i's do not form a martingale process.

In the absence of the martingale property it is not possible to determine the generalised asymptotic behaviour of any suitably selected infinite subset of (s_i, c_i) pairs. Therefore, the empirical calibration witnessed in this example does not have any broader implications on the general validity of infinite subsequences of c_i forecasts.

Similar to Chapter 6, the average s_i is taken for fixed intervals of c_i. Table 8.3 gives the results of the calibration and Figure 8.4 shows plots of the calibration results. Like the calibration results in section 8.4 (Figure 8.3), the results show that the forecasts are well calibrated towards the $\{0, 1\}$ end of the plot, but go astray slightly near 0.5, but only a small number of forecasts probabilities lie within this interval range. It would seem that the greater the uncertainty conveyed in the forecast, the worse the calibration.

8.5.2 Test Statistic

In a manner identical to Chapter 6, a test statistic to test the calibration of cross-validation forecasts is analysed. The test statistic tests the null hypothesis that the overall discrepancy between s_i and its forecast c_i is equal to 0; a measure of the overall calibration of the forecasts. In the previous arrangement such a null hypothesis would be synonymous with testing the complete calibration criterion. This validity criterion, however, has not been defined for the nonprequential forecasts.

The proposed test statistic is

$$V_k = \frac{\sum_{i=1}^{N} U_i (s_i - c_i)}{\left(\sum_{i=1}^{N} U_i var(s_i|Y_1, \ldots, Y_N, {}^{-}Y_i)\right)^{1/2}}. \tag{8.9}$$

Interval	Average c	Average s	n
$0 \leq c_i \leq 0.05$	0.0062	0.0068	1258
$0.05 \leq c_i \leq 0.15$	0.0873	0.0902	110
$0.15 \leq c_i \leq 0.25$	0.1898	0.1669	53
$0.25 \leq c_i \leq 0.35$	0.2947	0.3055	37
$0.35 \leq c_i \leq 0.45$	0.3973	0.5130	24
$0.45 \leq c_i \leq 0.55$	0.5063	0.5862	19
$0.55 \leq c_i \leq 0.65$	0.6106	0.5961	31
$0.65 \leq c_i \leq 0.75$	0.7007	0.7774	28
$0.75 \leq c_i \leq 0.85$	0.7980	0.8361	40
$0.85 \leq c_i \leq 0.95$	0.9128	0.9402	173
$0.95 \leq c_i \leq 1.00$	0.9947	0.9958	3224
overall	0.4998	0.5285	4997

Table 8.3: The calibration results of c_i forecasts with the smoothed prediction s_i.

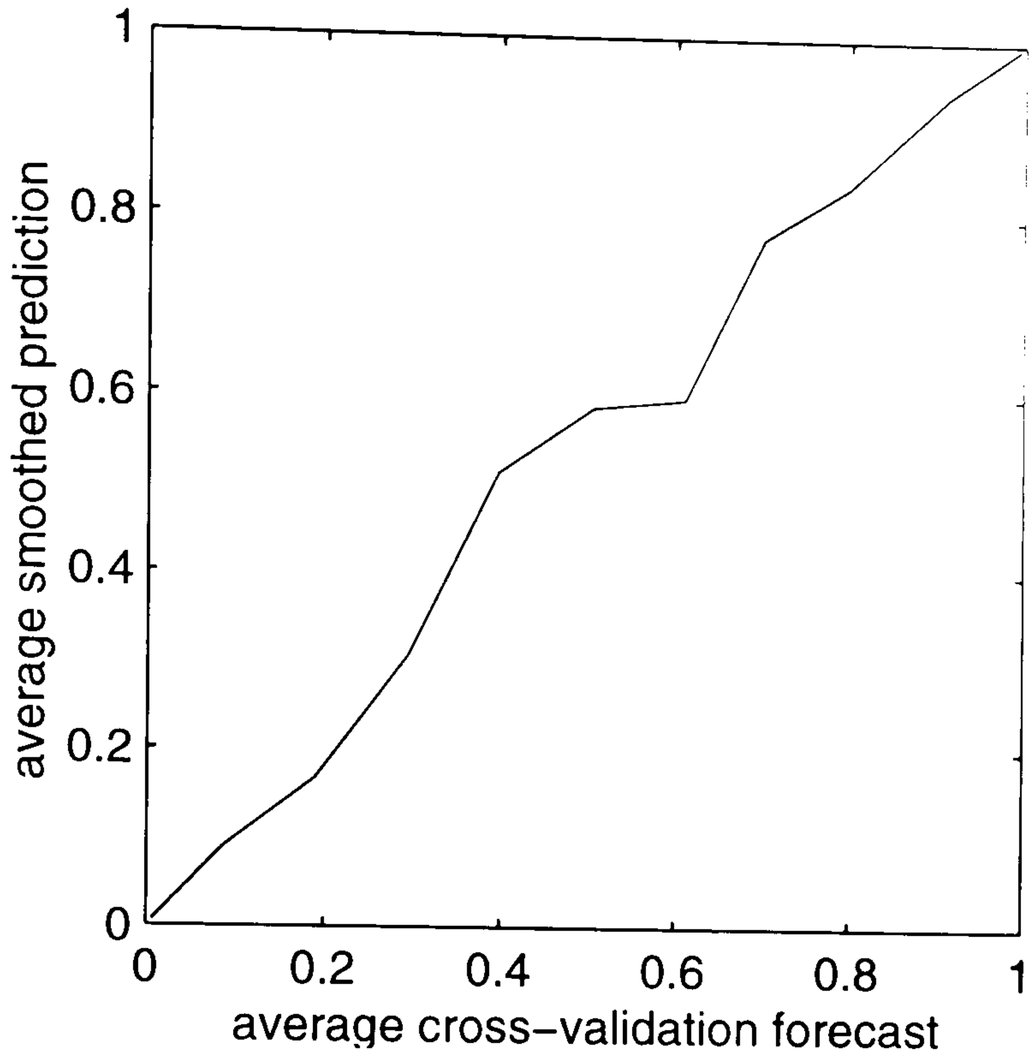

Figure 8.4: Calibration plot.

The U_i's are $\{0,\ 1\}$ indicator variables indicating the inclusion or seclusion of a $(c_i,\ s_i)$ pair. The U_i's are used to create subsequences of forecasts and predictions for the application of hypothesis testing. In this example the value of U_i is used to create subsequence of $(c_i,\ s_i)$ pairs based on prespecified intervals of c_i. When $U_i = 1$ for all i, V_k becomes the overall test of empirical calibration, V_0. If both c_i and $var\,(s_i|Y_1,\ \ldots,\ Y_N,\ {}^{-}Y_i)$ are fixed, then the only random element in the formation of V_k is s_i. This would make V_k a linear combination of independent random variables with mean zero and variance one.

The test statistic is used to test the empirical calibration of s_i with its forecast c_i. The asymptotic distribution of the test statistic can not be determined theoretically since the Central Limit Theorem for Martingales can not be applied to this example. However, if V_k is a linear combination of independent random variables with mean zero and variance one, then

by the central limit theorem for the sum of independent random variables the distribution of V_k can be approximated by a standard Normal. It is also possible to construct and analyse an empirical distribution for the test statistic and see how close it is to the standard Normal distribution.

Eleven disjoint subsequences along with the sequence as whole are considered in this section, giving a total of 12 test statistics to examine. The construction of the empirical distribution functions for each of the 12 test statistics is performed by simulating 1000 values of the test statistics using both the production method and a method analogous to the prequential simulation method. The diagram in Figure 8.5 explains the simulation of the latter method which will be referred to as the cross-simulation method. The frame of reference used is a *cross-validation* frame of reference. In the cross-simulation method the test statistics are evaluated using the data sequence described in section 8.2. At each point i in the data sequence, Y_i is simulated from the conditional distribution $P(Y_i|Y_1, \ldots, Y_N, {}^{-}Y_i)$. It is important to note that the values of $\{Y_1, \ldots, Y_N, {}^{-}Y_i\}$ are not simulated values, but come from the original sequence and therefore, the values of c_i and $var\,(s_i|Y_1, \ldots, Y_N, {}^{-}Y_i)$ remain the same throughout the 1000 iterations.

The diagram in Figure 8.6 describes the simulation method using the production model. This method is identical to that used in Chapter 6. The 1000 data sequences are simulated using the transition probabilities given in Table 8.1 which represent the production model. For each simulated sequence the test statistics are computed constructing an empirical distribution of 1000 values for each of the test statistics.

The empirical distributions constructed under both schemes are analysed using Normal probability plots, a plot of the ascending simulated test statistics versus their n-scores. For the cross-simulation method the Normal

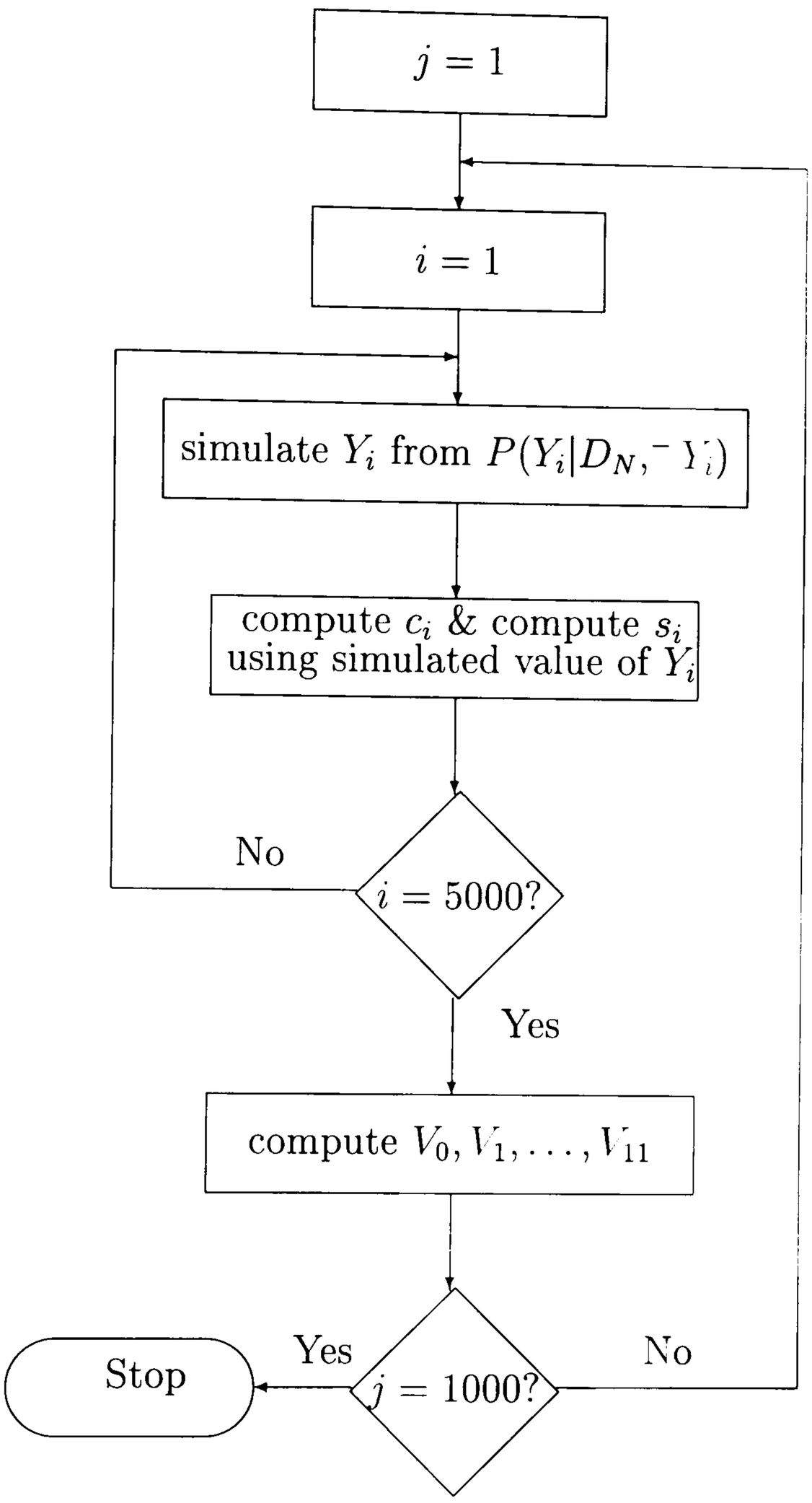

Figure 8.5: Flowchart of the test statistic simulation using the cross-simulation method. Note that the value of c_i's do not change (since the value of D_N is fixed throughout) and for efficiency can calculated prior to the simulation procedure.

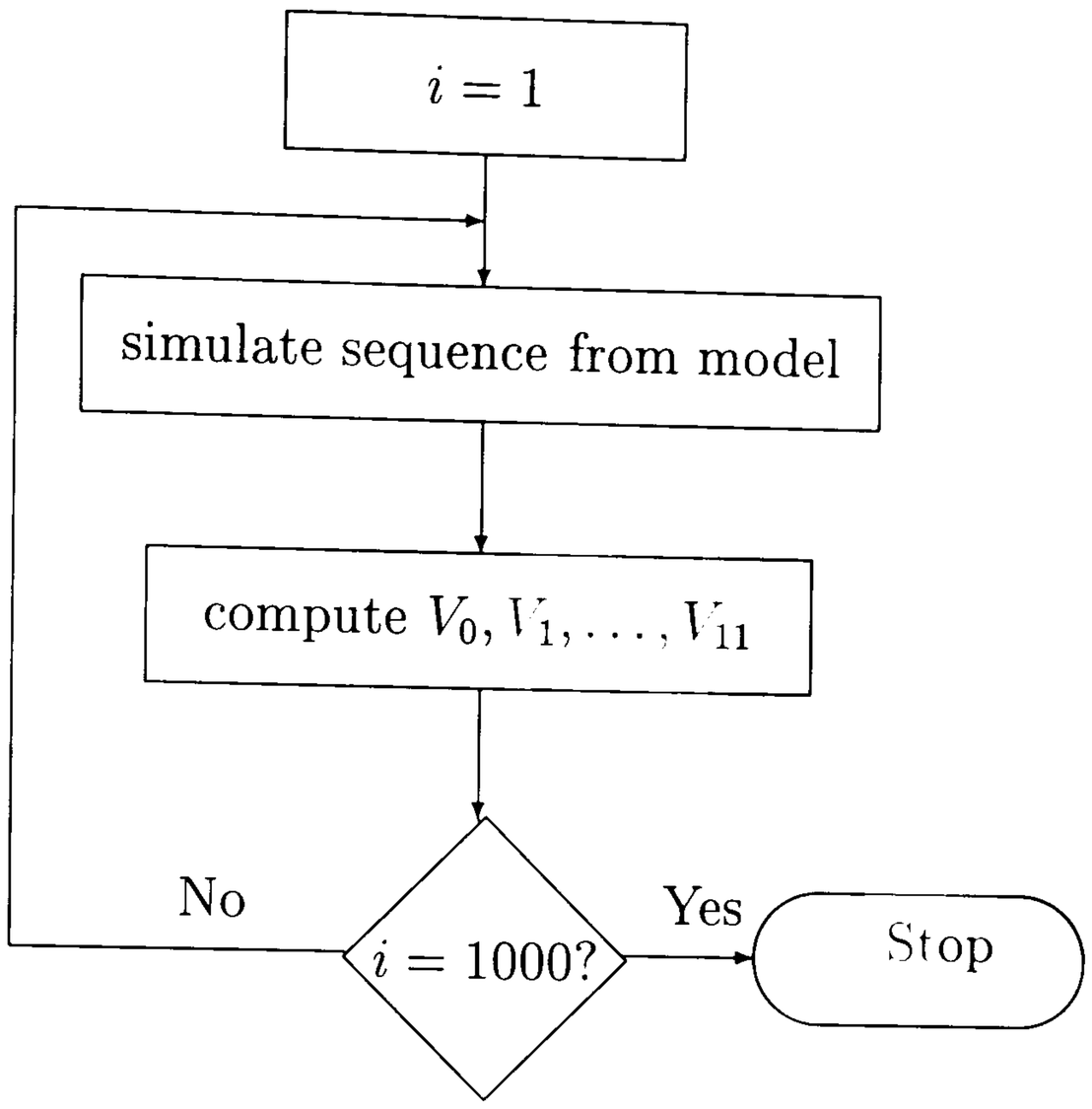

Figure 8.6: Flowchart of the test statistic simulation using the production model.

probability plots, shown in Figure 8.7, of the simulated distributions show that they do have a standard Normal distribution. The results can be scrutinised further by looking at the mean and standard deviation of the simulated distributions given in Table 8.4. As shown in Table 8.4, all the simulated distributions have a mean and standard deviation close in value to zero and one respectively.

Figure 8.8 shows the Normal probability plots for the test statistic distributions simulated using the production method. The results for this case are drastically different. The plots show that the distributions for all of the 12 test statistics are approximately Normal, but not standard Normal. It is clearly evident from the examination of the results in Table 8.4 that the means are not equal to zero and that the standard deviations are not equal to one. The empirical distribution of V_{11} and V_0 in particular both have large

V_k	Interval	Cross-Simulation		Production	
		mean	std	mean	std
V_1	$0 \leq c_i \leq 0.05$	-0.0090	0.9546	-1.9063	2.6522
V_2	$0.05 \leq c_i \leq 0.15$	0.0260	1.0612	0.9644	2.3444
V_3	$0.15 \leq c_i \leq 0.25$	-0.0266	1.0014	1.1585	2.0855
V_4	$0.25 \leq c_i \leq 0.35$	0.0299	0.9971	1.1498	1.8918
V_5	$0.35 \leq c_i \leq 0.45$	-0.0317	1.0176	1.0549	1.7500
V_6	$0.45 \leq c_i \leq 0.55$	0.0283	1.0017	1.0036	1.7627
V_7	$0.55 \leq c_i \leq 0.65$	-0.0118	1.0235	1.0490	1.7823
V_8	$0.65 \leq c_i \leq 0.75$	-0.0290	0.9986	1.0689	1.8604
V_9	$0.75 \leq c_i \leq 0.85$	0.0031	1.0147	1.3386	2.0472
V_{10}	$0.85 \leq c_i \leq 0.95$	0.0258	0.9881	2.7768	2.4297
V_{11}	$0.95 \leq c_i \leq 1.00$	-0.0067	1.0135	11.5286	3.8509
V_0	overall	0.0072	1.0403	26.2857	50.5663

Table 8.4: The means and standard deviations for the simulated test statistic distributions.

values for the mean and standard deviation.

In this example, the probability statements of the inferential model provided by the cross-validation frame of reference do not adhere to the sampling probabilities of the production model.

Despite the results, the test statistics are computed for c_i and s_i using the data described in section 8.2. The p-values are computed both from the standard Normal distribution and empirically using the two simulated distributions. The results are summarised in Table 8.5. The cross-simulation p-values are remarkably close to their Normal counterparts. This is not suprising since under the cross-validation frame of reference the values of both c_i and $var\,(s_i|Y_1,\ \ldots,\ Y_N,\ {}^{-}Y_i)$ are fixed and hence the V_k test statistics are linear combinations of independent random variables with mean zero and variance one. With this in mind, the central limit theorem for the sum of idependent random variables is applicable in this situation. It is, therefore, expected that the test statistic distributions under the cross-validation frame of reference be approximately standard Normal.

As Table 8.5 shows, the same is not true of the production simulation where the p-values are far from equal to their Normal counterparts. The calibration of the c_i's, based on both the Normal and cross-simulation p-values, is questionable since the p-values of the test statistics for the subsequences V_2, V_{10}, and V_{11} all show a very high level of significance. The p-value of the overall test statistic at 0.0702 is also slightly low.

8.6 Discussion

In this chapter, the updated forecast, q_i, used to assess the calibration of the p_i forecasts, is replaced with a more informative smoothed prediction.

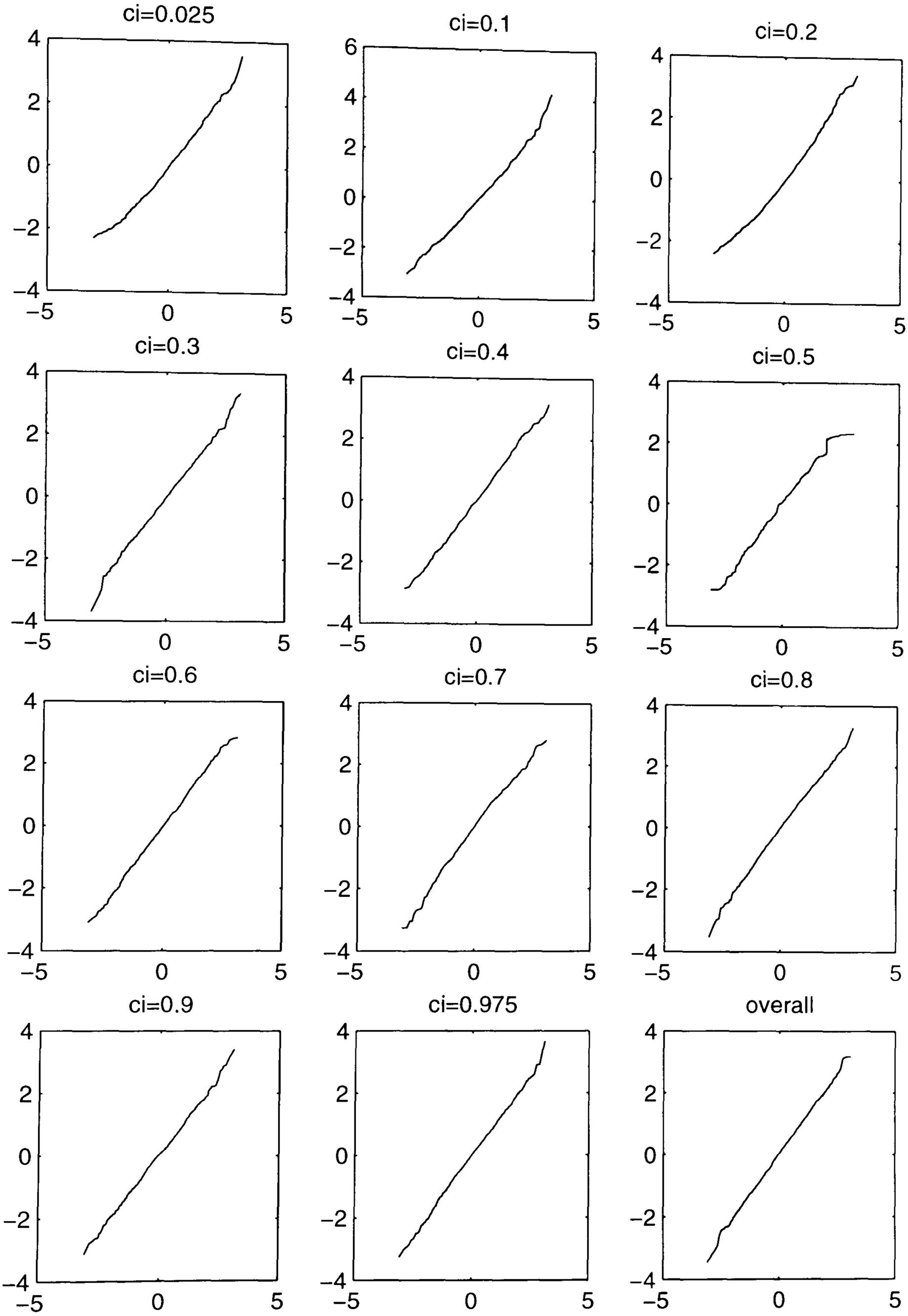

Figure 8.7: Normal probability plots for the cross-simulated test statistic distribution.

V_k	Interval	Test Statistic Z	Two-tailed Normal *p-value*	Simulated *p-value Cross-Simulation*	Simulated *p-value Production*
V_1	$0 \leq c_i \leq 0.05$	-0.0693	0.9448	0.9560	0.9750
V_2	$0.05 \leq c_i \leq 0.15$	-4.8099	0.0000	0.0000	0.0580
V_3	$0.15 \leq c_i \leq 0.25$	-0.2660	0.7903	0.7920	0.8870
V_4	$0.25 \leq c_i \leq 0.35$	0.5405	0.5889	0.5900	0.7760
V_5	$0.35 \leq c_i \leq 0.45$	0.5151	0.6064	0.6090	0.7990
V_6	$0.45 \leq c_i \leq 0.55$	0.7981	0.4248	0.4350	0.6830
V_7	$0.55 \leq c_i \leq 0.65$	0.1373	0.8908	0.8870	0.9470
V_8	$0.65 \leq c_i \leq 0.75$	-1.1534	0.2488	0.2450	0.5710
V_9	$0.75 \leq c_i \leq 0.85$	0.8193	0.4126	0.4300	0.7310
V_{10}	$0.85 \leq c_i \leq 0.95$	4.9312	0.0000	0.0000	0.1860
V_{11}	$0.95 \leq c_i \leq 1.00$	11.4021	0.0000	0.0000	0.5180
V_{12}	overall	1.8583	0.0631	0.0720	0.9560

Table 8.5: Summary of the test statistic results.

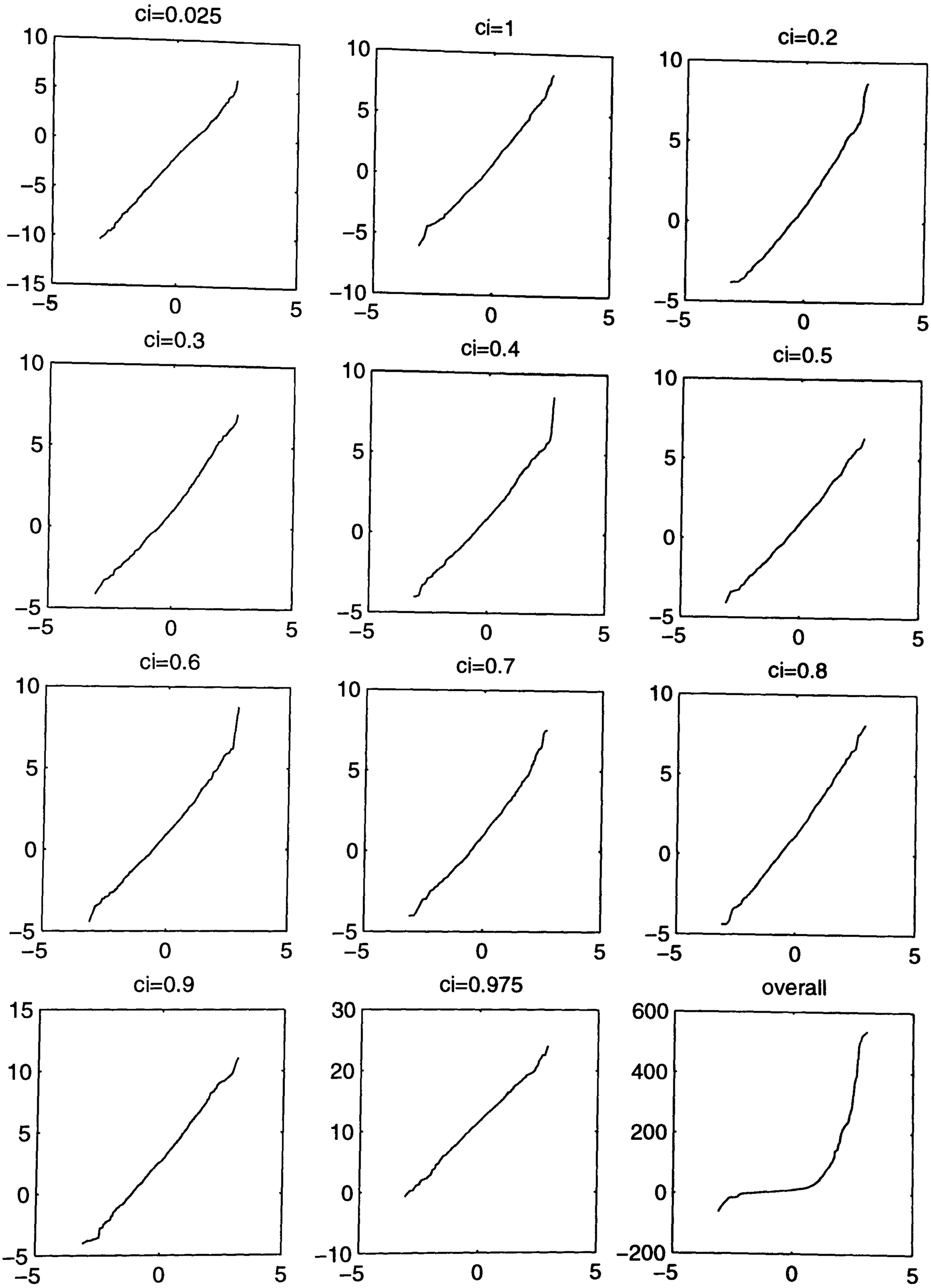

Figure 8.8: Normal probability plots for the production simulated test statistics.

s_i in the hope of obtaining a more enhanced forecast assessment. Using a simulated sequence of 5000 nucleotides the calibration of (p_i, q_i) pairs is compared with that of (p_i, s_i). As shown, although the smoothed predictions are more accurate indicators of the hidden state than their q_i counterparts, the calibration of p_i and q_i is slightly better than that of p_i and s_i.

The use of all the available information in the formulation of forecasting assessment comes at the cost of the prequential framework. Since prequential theory no longer applies when the smoothed predictions are used, p_i is replaced with a more information rich and yet *fair* cross-validation forecast, c_i. A test statistic similar to that in Chapter 6 is devised to assess the c_i's validity in explaining the s_i's in a manner synonymous with the complete calibration criterion. The absence of the martingale property inhibits the determination of the distribution of the purposed test statistic, V_k.

In a manner similar to Chapter 6, the distribution of the V_k's is examined in greater detail by simulating empirical distributions for them. The empirical distributions are simulated using both a cross-simulation method, which only simulates a value for Y_i while keeping all the remaining observations fixed at their original values, and the production method, which repeatedly samples from the production model. The formulation of the cross-simulation method makes it possible to invoke the central limit theorem for the sum of independent random variables, since the V_k's are linear combinations of independent random variables with mean zero and variance one and therefore, can be approximated by $N(0,\ 1)$. The cross-simulated empirical distributions of the test statistics reiterate this claim.

Unlike the prequential case, the cross-validation frame of reference presents a situation where the inferential method does not correspond to the production principle. Examination of the cross-simulated empirical distribution

shows that they test statistics have a $N(0,\ 1)$ distribution. The empirical distributions simulated from the production model, however, do not have a standard Normal distribution and because of this the cross-simulated distributions do not meet the minimum validity requirement of the production principle.

Chapter 9

Conclusion

The medley of various data-driven methods which has been presented here provides two different approaches to forecasting assessment and improvement. The first explored point forecasts constructed from Normal linear models and the second undertook the extension of the calibration techniques of probability forecasting assessments.

In the examination of the first case, emphasis was placed on the analysis of the recursive residual. Not only is this residual neatly suited to a prequential framework, but, as has been shown, the recursive residual is unique to the properties it possess. For a standard linear regression model it has been illustrated how the recursive residual, commonly used to detect model misspecification, can be used to correct it. The linear structure of a recursive residual vector of a deficient model can be exploited in the construction of a new model formation which when regressed on the missing components corrects the original misspecification.

Recursive residuals have also been introduced to new areas. The predictive distributions of a Normal variable with unknown mean and known precision has a recursive residual formation. In a manner similar to the mis-

specification correction strategy, it is shown how this residual can be used in a hierarchical modelling scheme and produce results identical to that of standard Bayesian analysis. The residual analysis concepts developed here were also extended to hidden Markov models. Recursive residuals are defined for the state of a system of various HMMs, as the difference between the one-step ahead prediction and the prediction's update. These definitions are used to show that, for the special case of a HMM with univariate state and multivariate observation sequences, the dimensionality of the data can be reduced to a univariate sufficient statistic without loss of information. This result also illustrates the correspondence between predictive sufficiency for a hidden state and conditional independence in a HMM configuration. In HMM applications where there are many variables, this compression technique can be used to simplify the analysis without compromising the model's predictive performance.

The modelling of residuals discussed earlier introduces a novel approach to modelling strategy in both the standard linear regression and Bayesian applications, and also to statistical analysis as seen in the HMM case. Three examples of very different residual applications have been given here, but, in all three cases, conclusions about statistical methods are drawn based on the residual content they produce. If, for a given set of data, the residuals of two varying statistical models are the same, then, regardless of the methods used to derive them, their corresponding forecasts are also the same. This gives an "ends justifies the means" approach to development and evaluation of statistical models based on the forecasts they produce.

This sort of residual analysis draws only on the linear correspondence between residual and observation. As such, the concepts presented are applicable to a limitless number of linear models from autoregressive models to

infeasible. Computational limitations also arise in the calculation of standard errors for the estimated parameters. The conditional independence structure of the observation sequence complicates the calculations to the extent that the standard errors can not be computed. The standard errors, had they been available, would have provided an indication of the quality of the estimates produced. The analysis of the estimation procedure is further hindered by the example DNA sequence used. The likelihood has been found to have more then one mode, which makes it difficult to determine the quality of the estimates produced prequentially.

The prequential framework has the advantage of possessing mathematical properties which simplify the statistical theory associated with prediction. As seen in the calibration case, the adherence to prequential theory made possible the use of the central limit theorem for martingales, enabling the possibility of hypothesis testing in empirical forecasting validation. However, when using HMMs, the prequential approach restricts the use of information in an already information-deprived situation. Leaving the prequential framework makes it possible to use more *information-rich* forecasts, such as cross-validation forecasts and smoothed predictions, and the assessing of one against the other using calibration. Although it is possible to judge the performance of the forecasts on a basic calibration level, it is not possible to extend the calibration criterion to forecasts of this type so that stronger statements can be made about the validity of such forecasts and the forecasting systems used to construct them. Empirical investigations show that a test statistic testing the complete calibration of smoothed and cross-validation forecasts, under the cross-validation frame of reference, is $N(0,1)$. However, in this case, the inferential model presented by the cross-validation frame of reference does not uphold the probability statements of the production

higher order HMMs and switching-state models.

Using HMMs, probability forecasting assessment is made possible by examining the calibration of the one-step ahead forecast with its filtered prediction. Both the complete calibration theorem and the calibration test statistic are extended to the HMM case. The application of these extended concepts in forecasting the occurrence of a CpG island in a DNA sequence shows that they perform well even for small samples.

Estimation is also carried out which allows for the examination of model-based calibration. The point of estimation is not to improve the model but the forecasts the model generates. For the forecasts calculated using Baum-Welch estimates, the calibration of the forecasts as indicated by their calibration plot is good; however, there is little evidence to suggest that model based forecasts are empirically better then their counterparts computed from unestimated transition probabilities. Since both forecasts are completely well calibrated then by the calibration criterion both the estimation based forecasts and the non-estimation based forecasts are indistinguishable as the number of forecasts made approaches infinity. A prequential estimation procedure which allows the sequential integration of observations as they become available is also introduced, but the calibration results in this case are very poor.

Computational limitations and the example data used both proved to be major obstacles in the development of a proper prequential estimation procedure. The prequential algorithm is a variation of the EM algorithm which require that the estimated parameters be allowed to reach a local maximum for a fixed set of data. The continuous incorporation of new data in the prequential procedure would require the estimates to converge at a local maximum for each new observation which is, unfortunately, computationally

infeasible. Computational limitations also arise in the calculation of standard errors for the estimated parameters. The conditional independence structure of the observation sequence complicates the calculations to the extent that the standard errors can not be computed. The standard errors, had they been available, would have provided an indication of the quality of the estimates produced. The analysis of the estimation procedure is further hindered by the example DNA sequence used. The likelihood has been found to have more then one mode, which makes it difficult to determine the quality of the estimates produced prequentially.

The prequential framework has the advantage of possessing mathematical properties which simplify the statistical theory associated with prediction. As seen in the calibration case, the adherence to prequential theory made possible the use of the central limit theorem for martingales, enabling the possibility of hypothesis testing in empirical forecasting validation. However, when using HMMs, the prequential approach restricts the use of information in an already information-deprived situation. Leaving the prequential framework makes it possible to use more *information-rich* forecasts, such as cross-validation forecasts and smoothed predictions, and the assessing of one against the other using calibration. Although it is possible to judge the performance of the forecasts on a basic calibration level, it is not possible to extend the calibration criterion to forecasts of this type so that stronger statements can be made about the validity of such forecasts and the forecasting systems used to construct them. Empirical investigations show that a test statistic testing the complete calibration of smoothed and cross-validation forecasts, under the cross-validation frame of reference, is $N(0,1)$. However, in this case, the inferential model presented by the cross-validation frame of reference does not uphold the probability statements of the production

model, since the production simulated test statistic distributions are not $N(0,1)$. This result is an example of the limitations of cross-validation as an inferential method and displays the superiority of prequential forecasts over their cross-validation associates. More research, however, is need to see the extent of such a claim.

The HMM calibration applications presented here bear witness to the easy adaptability and wide applicability of prequential theory and probability forecasting. It seems to reasonable to assume that these concepts can be extended further and applied to more structured HMMs.

Bibliography

[1] Baldi, P., and Chauvin, Y. (1994). Smooth on-line learning algorithms for hidden Markov models. *Neural Computation*, **63**, 307-318.

[2] Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**, 1-8.

[3] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Statistics*, **41**(1), 164-171.

[4] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L. (2000). GenBank. *Nucleic Acids Research*, **28**(1), 8-15.

[5] Bessler, D. A., and Kling, J. L. (1990). Prequential analysis of cattle prices. *Applied Statistics*, **39**, 95-106.

[6] Bird, A. (1987). CpG islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, **3**, 342-347.

[7] Broemeling, L. D. (1985). *Bayesian Analysis of Linear Models*. M. Decker.

[8] Brown, R. L., Durbin, J., and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society* B, **37**, 149-192.

[9] Churchill, G. A. (1992). Hidden Markov chains and the analysis of genome structure. *Computers and Chemistry,* **163**, 107-115.

[10] Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* Springer.

[11] Crowley, E. M., Roeder, K., and Bina, M. (1997). *Journal of Molecular Biology,* **268**, 8-14.

[12] Dawid, A. P. (1982). The well-calibrated Bayesian (with Discussion). *Journal of the American Statistical Association* **77**, 604-613.

[13] Dawid, A. P. (1983). Inference, Statistical. *Encyclopaedia of Statistical Science* Vol. 4, edited by S. Kotz, N. L. Johnson, and C. B. Read. Wiley-Interscience, 89-105.

[14] Dawid, A. P. (1984). Present position and potential developments: some personal views. Statistical theory the prequential approach (with Discussion). *Journal of the Royal Statistical Society* A, **147**, 278-292.

[15] Dawid, A. P. (1985). The impossibility of inductive inference. (Invited discussion of 'Self-calibrating priors do not exist,' by D. Oakes.) *Journal of the American Statistical Association*, **80**, 340-341.

[16] Dawid, A. P. (1985). Calibration based empirical probability (with Discussion). *Annals of Statistics.* **13**, 1251-1285.

[17] Dawid, A. P. (1986). Probability Forecasting. In *Encyclopaedia of Statistical Science* Vol. 7, edited by by S. Kotz. N. L. Johnson. and C. B. Read. Wiley-Interscience, 210-218.

[18] Dawid, A. P. (1991). Fisherian inference likelihood and prequential and prequential frames of frames (with Discussion). *Journal of the Royal Statistical Society* B, **53**, 79-109.

[19] Dawid, A. P. (1992). Prequential data analysis. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, edited by M. Gosh. and P. K. Pathak. IMS Lecture Notes-Monograph Series **17**, 113-126.

[20] Dawid, A. P. (1997). Prequential analysis. In *Encyclopedia of Statistical Science,* Updated Vol. 1, edited by S. Kotz, C. B. Read, and D. L. Banks. Wiley-Interscience, 467-470.

[21] Dawid, A. P. (1998). Conditional Independence. In *Encyclopedia of Statistical Science,* Updated Vol. 2, edited by S. Kotz, C. B. Read, and D. L. Banks. Wiley-Interscience, 146-155.

[22] de Finnetti, B. (1975). *The Theory of Probability,* Vol. 1. Wiley.

[23] de Finnetti, B. (1975). *The Theory of Probability,* Vol. 2. Wiley.

[24] DeGroot, M. H. (1989). *Probability and Statistics.* Addison-Wesley.

[25] DeGroot, M. H., and Fienberg, S. E. (1982). Assessing probability assessors: calibration and refinement. *Statistical Decision Theory and related topics. III* Vol. 1 edited by S. S. Gupta, and J. O. Berger. 291-314.

[26] DeGroot. M. H., and Fienberg, S. E. (1983). The comparison and evaluation of forecasts. *The Statistician,* **32**, 12-22.

[27] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B, **39**, 1-38.

[28] Durbin, R., Eddy, S., Krogh, A., and Mitchinson. G. (1998). *Biological Sequence Analysis.* Cambridge University Press.

[29] Elliot, R. J., Aggoun, L., Moore, J. B. (1995). *Hidden Markov Models: estimation and control.* Springer-Verlag.

[30] Everitt, B. S. (1984). *An Introduction to Latent Variable Models.* Chapman and Hill.

[31] Feller, W. (1971). *An Introduction to Probability Theory and its Applications,* Vol. 2, John Wiley.

[32] Ghahramani, Z. (1998). Learning Dynamic Bayesian Networks . In *Adaptive Processing of Sequences and Data Structures* . Lecture Notes in Artificial Intelligence, C.L. Giles and M. Gori (eds.), 168-197. Springer-Verlag.

[33] Ghahramani, Z., and Roweis, S. (1999). Learning linear dynamic systems using the EM algorithm. *Advances in Neural Information Processing Systems,* **11**, 431-437.

[34] Hadi, A. S. and Son, M. S. (1990). Some properties of and relationships among several uncorrelated residuals. *Communications in Statistics: Theory and Method,* **19**, 2625-2642.

[35] Hall, P., and Heyde, C. C. (1980). *Martingale Central Limit Theory and its applications,* Academic Press.

[36] Hamilton, J. D. (1988). Rational-expectations econometric analysis of changes in regime: an investigation of the term structure of interest rates. *Journal of Econometrics, Dynamics, and Control,* **12**. 385-423.

[37] Hamilton, J. D. (1989). A new approach to the econometric analysis of nonstationary time series and the business cycle. *Econometrica*. **2**. 357-384.

[38] Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics,* **45**. 39-70.

[39] Hamilton, J. D. (1993). State-space models. In *Handbook of Econometrics* Vol. 4, edited by R. Engle, and D. McFadden.

[40] Harvey, A. C. (1990). *The Econometric Analysis of Time Series.* Philip Allen.

[41] Harvey, A. C. (1993). *Time Series Models.* Harvester Wheatsheaf.

[42] Helland, I. S. (1982). Central limit theorem for Martingales with discrete and continuous time. *Scandinavian Journal of Statistics,* **9**, 79-94.

[43] Hinton, G., Revow, M., and Dayan, P. (1995). Recognizing handwritten digits using mixtures of linear models. In *Advances in Neural Information,* Vol. 7, edited by G. Tesauro, D. Touretzky, and T. Leen. 1015-1022.

[44] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of ASME, Journal of Basic Engineering,* 82D. 35-45.

[45] Kalman, R. E., and Buey, R. S. (1961). New results in linear filtering and prediction. *Journal of Basic Engineering (ASME).* **83**D. 95-108.

[46] Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, **82**, 1032-1063.

[47] Kling, J. L., and Bessler, D. A. (1989). Calibration-based predictive distributions: an application of prequential analysis to interest rates, money, prices, and output. *Journal of Business*, **62**, 477-499.

[48] Krogh, A. (1994). Hidden Markov models for labeled sequences. In *Proceedings of the* 12^{th} *IAPR International conference on Pattern Recognition*, 140-144. IEEE Computer Society Press.

[49] Krogh, A. (1998). An introduction to hidden Markov modelsfor biological sequences. In *Computational Biology: Pattern Analysis and Machine learning Methods*. edited by S. Salzberg, D. Searls, and S. Kasif. Elsevier.

[50] Lichenstein, S., Fischoff, B., and Philips, L. D. (1982). Calibration of probability: The state of the art to 1980. In *Judgement under uncertainty: Heuristics and Biases*, edited by D. Kahneman, P. Slovic, and Tversky. Cambridge University Press.

[51] Lumsdaine, R., and Ng, S. (1999). Testing for ARCH in the presence of a possible misspecified conditional mean. *Journal of Econometrics*, **93**. 257-79.

[52] McCullouch, R. E., and Tsay, R. S. (1994). Statistical analysis of economic time series via Markov switching models. Journal of Time Series Analysis, **155**, 523-539.

[53] McLachlan, G. J. (1997). *The EM Algorithm and Extensions*. Wiley-Interscience.

[54] Miller, R. J. (1962). Statistical prediction by discrimination analysis. *Meteorological Monographs*, **4**, no. 25.

[55] Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595-600.

[56] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, **77**, 257-285.

[57] Rabiner, L. R., and Juang, B. H. (1986). An introduction to hidden Markov Models. *IEEE ASSP Magazine*, **3**(1), 4-16.

[58] Roweis, S., and Ghahramani, Z. (1999). A Unifying review of linear Gaussian Models. *Neural Computation*, **11**(2), 305-345.

[59] Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, **2**, 191-201.

[60] Schott, J. R. (1997). *Matrix Analysis for Statistics*. Wiley-Interscience.

[61] Seillier-Moiseiwitsch, F., and Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*, **88**, 355-359.

[62] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). *Journal of the Royal Statistical Society* B, **36**, 111-147.

[63] Symthe, P., Heckerman D., and Jordon, M. I. (1997). Probabilistic independence networks for hidden Markov models. *Neural Computation*, **9**(2), 227-269.

[64] Theil, H. (1965). The analysis of the disturbances in regression analysis. *Journal of the American Statistical Association,* **60**, 149-192.

[65] Theil, H. (1968). A simplification of the BLUS procedure for analysing regression disturbances. *Journal of the American Statistical Association,* **63**, 242-251.

[66] Theil, H. (1971). *Principles of Econometrics.* John Wiley & Sons.

[67] West, M., and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models.* Springer.

[68] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* Wiley.