Kent Academic Repository

Rodrigues, David Francis (2009) *Modelling election poll data using time series analysis.* Doctor of Philosophy (PhD) thesis, University of Kent.

Downloaded from

https://kar.kent.ac.uk/94615/ The University of Kent's Academic Repository KAR

The version of record is available from

https://doi.org/10.22024/UniKent/01.02.94615

This document version

UNSPECIFIED

DOI for this version

Licence for this version

CC BY-NC-ND (Attribution-NonCommercial-NoDerivatives)

Additional information

This thesis has been digitised by EThOS, the British Library digitisation service, for purposes of preservation and dissemination. It was uploaded to KAR on 25 April 2022 in order to hold its content and record within University of Kent systems. It is available Open Access using a Creative Commons Attribution, Non-commercial, No Derivatives (https://creativecommons.org/licenses/by-nc-nd/4.0/) licence so that the thesis and its author, can benefit from opportunities for increased readership and citation. This was done in line with University of Kent policies (https://www.kent.ac.uk/is/strategy/docs/Kent%20Open%20Access%20policy.pdf). If you ...

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in *Title of Journal* , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

MODELLING ELECTION POLL DATA USING TIME SERIES ANALYSIS

by

David Rodrigues

Ph.D Thesis

University of Kent, 2009

Acknowledgements

I am greatly indebted to my supervisor, Professor Stephen Walker, for providing focus and consistent guidance throughout the three years, and without whom this thesis would have been nearly impossible. Thank you for your patience.

Thank you also to my family and friends for their encouraging words of support through the harder times over the course, and not to mention, their belief in me finishing this amalgamation of work, which has been both challenging and rewarding.

Finally, I am grateful to the Engineering and Physical Sciences Research Council for its generous funding over the course of my study, without which it I doubt I would have managed the completion of this thesis within a full-time duration of three years.

This thesis is dedicated affectionately to my mother, father, sister and two brothers.

Abstract

There is much interest in election forecasting in the UK. On election night, forecasts are made and revised as the night progresses and seats declare results. We propose a new time series model which may be used in this context. Firstly, we have statistical models for the polls conducted in a run-up to the election; the model produces the distribution of voting amongst the parties. The key here is the use of modelling the probability of voting each poll as latent variables. Secondly, we use this information in the forecasting of the inevitable outcome, continually revising our forecasts as the actual declarations are made, until we can actually determine what we believe the final outcome to be, before it actually happens. We outline the nature and history of elections in the UK, and provide an account of time series analysis. These tools, as well as the theoretical basis of our method, the h-likelihood, are then applied to the creation of each of our models proposed. We study simulations of the models and then fit the models to actual data to assess forecasting accuracy, using existing models for comparison.

List of Common Notations

y_t: observed data for a particular party at poll t

x_t: probability of voting for a particular party at poll t

r_t: total number of votes at poll t

δ_t: time interval (days or hours) between poll t and poll $t-1$

θ: vector of parameters

$\hat{\theta}$: maximum likelihood estimate vector

$\tilde{\theta}$: maximum hierarchical likelihood estimate vector

L: likelihood

l: log likelihood

h: hierarchical likelihood

ARCH(k, j): ARCH model with k parties and j lags

SV(k, j): stochastic volatility model with k parties and j lags

GARCH(k, j): GARCH model with k parties and j lags

i.i.d.: identically and independently distributed

$\mathbb{E}(X)$: expectation of variable X

$\mathbb{V}ar(X)$: variance of variable X

$\mathbb{C}ov(X, Y)$: covariance of variables X and Y

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Elections and polls

Before a general election, various opinion polls are taken, asking samples of the electorate for whom they *will* vote. These are carried out over a certain length of time up to election night. Following voting, exit polls are carried out, asking samples of the electorate for whom they actually *did* vote. The UK is split into what are known as 'constituencies'. Throughout election night, each constituency declares the number of votes per candidate. The winning party in that constituency then represents the seat. At the end of election night, the total number of seats per party is added, and the party with the majority of seats forms the Government. The main parties in the UK overall are now Conservative, Labour and Liberal Democrats, who all have candidates standing in almost all constituencies.

Politicians, among other groups, are interested in forecasting the final outcome before it happens; not only for the final result but also broken down by constituency. Various organisations seek to produce the best forecasts and to do this requires a robust and accurate forecasting method. The forecasts (mainly of the final total outcome and for the key parties in the UK) are quoted in the media. Also, the country begins to plan for the economy, implications of the winning party are made based on these forecasts, and a whole host of 'what if?' scenarios are widely debated. This has led to various approaches, statistical and non-statistical, to satisfy demand.

For instance, UK Elect ([106]) is an online election viewer, forecaster, processor and simulator which can be used at any time to predict the outcome of an election. It can take current opinion poll percentages for the main parties plus the results of one previous election and predict the results of the next election by constituency. It may also be used to extrapolate these results from the results of by-elections. In Chapter 4 we will review key statistical approaches to election forecasting which have been developed, looking at their specific methodologies. These include the poll of polls, the cube law as well as some newer ones.

Lewis-Beck (2005, [74]) distinguishes between two types of forecasting: scientific and non-scientific. The former, with which we are concerned, offers estimates based on some scientific procedure, such as a simulation, market analysis, sample survey or statistical model. The latter are guesses based on hunches, intuition, casual conversation, non-systematic interviews or coincidence. The paper of Lewis Beck (2005) reviews the leading scientific approaches to election forecasting, looking at models which come from the US, the UK and France. Forecasting from statistical models has been popular in the US and France, but has also been existent in the UK, although to a lesser extent. The UK focus has been on using opinion polls to forecast. Whiteley (1979, [108]) made the first attempt, with a Box-Jenkins model of monthly data based on 'popularity'.

Lewis-Beck (1985, [75]) concluded that a forecasting method should be compared using four criteria: accuracy, lead time, parsimony and reproducibility:

- Accuracy - in classical regression analysis the two measures of goodness-of-fit are the R^2 value and standard errors.

- Lead time - this is the time duration between observations of election data. For poll data the polls are collected over a short time period. The same is clearly true for seat declarations on election night. By contrast, election final outcomes usually only occur every four years and so lead time is somewhat larger.

- Parsimony - other things being equal, a few well-specified variables will work better than many questionable ones. Fewer parameters to estimate means less error from parameter estimation. Parsimony is important since the sample sizes, of poll data for instance, are so small.

- Reproducibility - generally, parsimonious models are easier to understand and reproduce. This issue is important if the model is to be used by other analysts, for example. This is harder if the measures are costly, in time or money, say.

Lewis-Beck (1985) proposed putting these factors into a single formula, known as a 'quality index', Q, in order to get a rough idea of the model's usefulness and to compare different models:

$$Q = \frac{(3A + P + R)L}{20},$$

where accuracy A, parsimony P, reproducibility R and lead time L all take values between 0 and 2, and the 20 is the maximum possible score of the numerator to enable an upper limit of 1 for Q. Note that accuracy has been given three times the weight of parsimony and reproducibility. A model which is highly accurate, parsimonious, easily reproducible and with good lead time would score $Q = 1.0$, whereas one which is the complete opposite would score $Q = 0.0$. The limitation of this index is that one may only *subjectively* assign values to each factor, which may be difficult to do.

To forecast election night using tools within time series analysis makes sense. This is particularly due to the progressive way in which the constituencies declare results. It would be useful to adapt models according to declared results, rather than just stating a model which remains unchanged as results arrive. It is also a natural choice because the polls in the weeks before the night itself are implemented at different stages, whose outcomes should be correlated with the state of the competing parties, which is prone to variation with time. There is considerable evidence that the minds of voters are made up well before election day (Bean, 1948, [8]). An average of poll outcomes cannot tell us about swings amongst parties for instance, but provides merely a point estimate. By contrast, time series analysis can reveal such issues.

1.1.2 Time series

Kendall and Ord (1990, [63]) state that recording occurrences on a graph whose horizontal axis comprises equal intervals to represent equal spaces of time must have occurred over a thousand years ago. However, until about 1925, a time series was regarded as being generated deterministically; evident departures from trends, cycles or other systematic patterns of behaviour that were observed in Nature were simply regarded as 'errors'. In 1927, such irregularities in amplitude and distances between successive peaks and troughs were identified by Yule (1927, [111]) as a series of **shocks** which are incorporated into the motion of a system. This leads to the theory of stochastic processes of which the theory of stationary time series, which includes Box-Jenkins analysis (Box and Jenkins, 1970, [18]), is part.

Graphical examination of data series illustrates what is going on generally; we aim to identify and explain this formally in a time series analysis. Subsequently, analyses may be separated into two classes: **time-domain** methods and **frequency-domain** methods. Time-domain methods have a model-free subset consisting of the examination of auto-correlation and cross-correlation analysis, but it is here that partly and fully-specified time series models appear. Frequency-domain methods centre around spectral analysis and recently wavelet analysis, and may be regarded as model-free analyses well suited to exploratory investigations. Note that in this thesis we focus on the former type.

As shown by Box and Jenkins (1970), models for time series data may have many forms and represent different stochastic processes. When modelling variations in the level of a process, three broad classes of importance are the autoregressive (**AR**) models, the integrated (I) models, and the moving average (**MA**) models. These depend *linearly* on previous data. Combinations of these ideas produce autoregressive moving average (**ARMA**) and autoregressive integrated moving average (**ARIMA**) models. Extensions to deal with vector-valued data are available. Other extensions are available where the observed time series is driven by some 'forcing' time series (which may not have a causal effect on the observed series); the distinction from the multivariate case is that the forcing series may be deterministic or under control. Non-linear dependence of the level of a series on previous data points is also of interest, partly to try to account for more chaotic behaviour. Empirical investigations can indicate the advantage of using predictions derived from non-linear models, over those from linear models. Among other types of non-linear time series models, there are models to represent the changes of **variance** along time (**heteroskedasticity**). Such models are called autoregressive conditional heteroskedasticity (**ARCH**) (Engle (1982), [40]) and they comprise a wide variety of representation (including **GARCH** models). Here changes in variability are related to, or predicted by, recent past values of the observed series. There are, however, other possible representations of locally-varying variability, where the variability might be modelled as being driven by a separate time-varying process, as in a doubly-stochastic model.

In recent work on model-free analyses, wavelet transform based methods (for example locally stationary wavelets and wavelet decomposed neural networks) have been used. Multi-scale techniques decompose a given time series, attempting to illustrate time dependence at multiple scales.

An important feature of time series analysis is to be able to track trends in observed data in order to use the information gained to forecast how the time series is likely to continue. Attached to this is the need to assess the likely forecasting error; for example, providing confidence intervals of the predictions.

Time series modelling has subsequently seen a growth of popularity recently, and there is plenty of theory to deal with the various types of data. As with modelling in general, we often have a set of candidate models to which we might fit the data, and must examine the underlying features and implications of each before choosing the 'best'.

In this thesis we will use formal statistical methodology for this important type of time series, placing emphasis on the probabilities of voting for each party. Our key component is the modelling of voting probabilities, which are **latent**, in that they are clearly unobserved.

The idea of using latent variables in statistical modelling has seen increasing popularity. Latent variables are not directly observed, but rather inferred through a mathematical model from variables which are observed[1]. Their uses are diverse, appearing in social sciences, medicine, computer science and economics, but the exact definition varies in each field. Examples of latent variables from the field of economics, for instance, include the quality of life, business confidence and happiness - all variables which cannot be measured directly. Aigner and Goldberger (1977, [2]) contains a selection of papers, published and unpublished, which concern latent variables in economic analysis, such as the estimation of regression relationships containing unobservable independent variables. Goodman (1978, [49]) considers a wide range of latent-structure measurement models, fitting them to particular datasets for illustration. Hagenaars (1993, [54]) looks at contingency table analysis, covering latent variables in loglinear modelling, causal models and longitudinal models. Bollen (1989, [15]) provides a thorough discussion of structural equations with latent variables.

In the context of time series, Bondon (2005, [17]) points out two required assumptions, stating first that if a parametric model admits a finite dimensional state space representation then it may accommodate such 'missing' data. One assumption is that the responses on the indicators are the result of an individual's position on the latent variable(s). The second is that the response variables have nothing in common after controlling for the latent variable(s), which is also known as the 'axiom of local independence'.

Bondon (2005) investigates the influence of missing data on the linear prediction of stationary[2] time series. Lower and upper bounds for the prediction error variance are established, properties of the predictor presented and asymptotic behaviour for the prediction error variance obtained, for short and long-memory

[1] Definition from http://en.wikipedia.org/wiki/Latentvariable.

[2] 'Stationary' here refers to covariance stationarity, which means that the first two moments (and therefore mean, variance and covariance) remain constant in any time space.

processes, respectively.

Estimation of the latent data will be important. With usual time series data, assumed to have Gaussian errors, the exact Gaussian likelihood function may be computed via the associated Kalman prediction recursions[3] and prediction error variances found via the Kalman fixed point smoothing algorithm (a way of computing smoothed estimates of the state vector at some fixed point in time). An alternative to estimate the missing data involves using explicit formulae of these estimates, as proposed by Brubacher and Wilson (1976 [26]) and Ljung (1989, [80]), which is often more efficient numerically, especially if the number of missing observations is small (Ferreiro, 1987, [43]). [Also, explicit formulae are useful for analysing theoretically the influence of the number and positions of the missing data on the error variances (Pourahmadi, 2001, [97]).] An expression for the best linear interpolation of a finite number of missing data with arbitrary pattern of any stationary time series, whose past and future are observed indefinitely, was given by Rozanov (1967, [99]). (This involves the inverse autocorrelations of the series, which may be expressed themselves in terms of the AR(∞) parameters[4].) An important departure for us is that our model will not be Gaussian.

1.1.3 Statistical models for election forecasting

Brown and Payne (1975, [23]) go into the *problems* encountered in the forecasting of elections. Firstly, **special seats** are somewhat independent of the remainder of seats in that they cannot be used to predict other seats, nor can they be predicted from other seats. The **boundaries** of constituencies often change, whereas previous election results are a vital source of information. Commonly, this problem is tackled by imagining that the change occurred before the previous election and then recalculating the outcome. However, this is not always an easy process. Importantly, the order of declaration is not a representative sample of the electorate as a whole; for example, urban areas (which are predominantly Labour) tend to declare earlier than rural areas (which are predominantly Conservative). Further, many Scottish and Welsh seats where nationalist parties stand tend to declare late. Finally, the list of candidates standing is not published until ten

[3] The Kalman filter is a recursive estimator with two phases: predict and update. The predict phase uses the state estimate from the previous timestep to produce an estimate of the state at the current timestep. In the update phase, measurement information at the current timestep is used to refine this prediction.

[4] $AR(\infty)$ refers to a time series model in which the current observation is an additive linear function of all previous observations as well as a random component; the parameters are coefficients of the previous observations.

days before the election, and so the political scenario may change rapidly in the run-up to election night. Therefore, a good forecasting model will need to allow for such issues.

Bean (1948) comments on the importance of assessing how business and agricultural conditions, religious preferences, nationality and cultural groups, and third parties affect the fortunes of the major parties. It should be possible to include any such factors into the forecasting method on election night, in the form of explanatory variables.

Another avenue of election forecasting is the prediction of transition rates between parties. Miller (1972, [86]) presents cross-tabulation models for **swing** (the net change when there are only two major parties), which includes the estimation of transition probabilities via ridge regression[5]. The key here is to model change in which the behaviour of individuals is related to their political environment as well as individual voting history. Therefore, the model involves both individual and aggregate (national) data.

1.2 The Problem and Our Approach

We have information on voting in the form of polls conducted in the period up to a general election. Clearly, this scenario is a time series framework. Therefore, we are interested in producing a statistical model which can take this into account for two reasons:

1. To be able to summarise statistically what these polls say about the voting distribution.

2. To use our modelling in the forecasting of the final outcome. This part would be done sequentially; that is to say, each time a constituency declares its results we refine our forecast, incorporating into our model all the known information, until we eventually have enough backing to declare a final distribution of seats amongst all competing parties in the UK. Also, we need to determine precisely *when* it is that we can make this final forecast.

We dedicate more effort on the first of the reasons stated above, since our forecasting approach is based on that already developed for the BBC and outlined in Brown and Payne (1975, [23], 1984, [24]) and Brown, Firth and Payne (1999, [22]).

[5] Ridge regression is a linear regression technique which modifies the usual residual sum of squares computation to include a penalty for large parameter estimates.

We aim to find parameters which each represent the strength of a party standing in the general election. To achieve this we make use of the collective opinion polls conducted by the various organisations in the build up to the election night. The polls are conducted at different stages and so, it is hoped, will reflect and track the strength of a party relative to others over time. A simple time series plot of proportion of votes is fine as an initial picture.

In arriving at these parameters, we rely on latent data. These data are the probabilities of voting for the parties at each poll. To be more specific, the latent data will be a *vector of probabilities*, summing to one, which represent the support for each party in a frequency sense. Suppose that there are N parties of interest and we have the latent vector $(p_1, p_2, \ldots, p_N)$ then $100p_1\%$, for instance, is the percentage of support for party 1 within a constituency *or* a collection of constituencies. Therefore, they correspond with the numbers who vote at that poll. We may then assess the change in dynamics between the parties, and infer of their relative strength over time.

These probabilities evolve over time, which forms the basis of our time series approach. Generally speaking, for each poll we will model the probabilities as random variables which follow Dirichlet distributions. Subsequently, we insert these probabilities as parameters of a multinomial distribution, along with the sample size of the poll, which in turn models the number who vote in that poll. Then, for each poll we have a unique mixed **Dirichlet-multinomial** distribution in a discrete-time setting. In time series analysis, we are often interested in assessing autocorrelation, that is, the extent to which lags affect the current value. If our assumption holds that each is representative of the UK opinion then it is possible that voting at some point tends to relate to voting at or up to some earlier point.

We focus on these latent probabilities, considering *three* types of modelling. The simplest is the assumption of stationarity of voting, whereby we do not consider time interval magnitudes in our modelling and assume that the voting pattern of a party does not vary (widely) over the build-up period. The other two, by contrast, consider volatility, whereby different time intervals result in different marginal Dirichlet distributions. As well as time intervals, these three types of model each use different historical information; one uses just observed votes, another previous probabilities and the final uses both. Combined with modelling lag polls, this means that we have a variety of information to use to determine the probabilities of voting, and thus we will have a number of candidate specific models to consider for any set of data. Note that we will operate in discrete time. However, having latent data raises issues for estimation, in that we cannot apply the usual methods of direct maximisation of likelihood. One option is to apply the EM algorithm introduced by Dempster, Laird and Rubin (1977, [34]), treating the latent probabilities as missing data. However, this depends on us evaluating integrals which may prove either difficult or impossible, particularly when we make the models more complicated. Therefore, we make use of a similar approach, called the **h-likelihood** method, introduced by Lee and Nelder (1996, [69]), which is a double maximisation method. First, we fix the latent variables, temporarily treating them as if they were observed, and maximise with a usual Newton-Raphson approach. Then, we switch this round to fixing the parameters and subsequently maximise a likelihood over the latent variables. We iteratively repeat this until convergence, when we simply use the 'optimal' parameters and formally 'disregard' the maximised latent data. However, the latter are still interesting to view in a plot with time, because of the fact that they are the estimated probabilities of voting at each time point.

Now we must address the second of our two objectives given above. The basic idea of our modelling on election night is as follows. For each constituency, we assume that there is a prior distribution on the vector of probabilities, which is determined by the exit poll information. On election night once other constituencies have declared their results we use both these results and the prior information to predict the vote outcome of the undeclared constituencies. This method is based on Brown and Payne (1975). We then use this information to update the prior distribution of each undeclared constituency. This process happens through the night, until all constituencies have declared their results.

1.3 Structure of the Thesis

We begin in Chapter 2 by outlining the *general election* in the UK, including the preceding run up to election night, the night itself in terms of the process by which the seats declare, a history of outcomes in past elections, the main candidate parties and how the UK is divided into constituencies.

Chapter 3 then provides a review of the subject of *time series* as a whole. We look at the history of its development, which covers both its applications as well as the key theory, firstly focusing on Box-Jenkins analysis. Next, we switch to aspects which we will make use of later in our modelling, which rely on the foundations of time series theory described in the first half of the chapter.

Chapter 4 moves on to discuss the statistical *forecasting* of election outcomes, which has been considered by others in the past. These include the cube rule and the multivariate structural time series model.

Chapter 5 is a particularly important part, in the sense that it provides the theoretical basis behind our main method of estimation, namely the *h-likelihood*; we look at established theory and proofs and give a brief review of its usefulness.

This leads naturally into Chapter 6, in which we take this method and make it specific to our problem, that is, put it into a time series environment, which involves a subtle check that we can use the method in our case. Here, we introduce the generic and then specific *models*, and outline goodness of fit tests available to be able to select the best model for further analysis. The chapter includes simulation analysis in order to study model behaviour. We also provide comparisons of our simplest model with the EM algorithm.

Chapter 7 is an *illustration* chapter. We take actual data, poll outcomes in the run up to historial elections, and fit each of our models to them in turn to summarise the poll information. We next perform various diagnostic checks to decide on the best model(s), including simulations and goodness-of-fit tests. Finally, here we compare our model with performances of comparative models from elsewhere, which were outlined in Chapter 4.

Chapter 8 firstly discusses our *forecasting* approach in some detail. The method adopted by the BBC in its all-night coverage of the election night is reviewed. The chapter then goes on to illustrate our method, albeit in a somewhat simplified scenario.

Finally, Chapter 9, as well as summarising our thesis, suggests possible *future* work which may develop all that we have discussed.

Chapter 2

The General Election

2.1 Introduction and Overview

In this chapter we give a broad account of the entire general election process in the UK. The aim of this is to provide a background to the data to which we will end up fitting models and data from which we will forecast. We start by providing the necessary definitions, and then outline the structure of the UK with respect to the general elections, the main parties, and also a history of recent outcomes. Finally, the role of pre-polls will become fundamental to our modelling later, and so we will discuss these.

In the UK, the general election is an election in which all or most members of Government stand for election to the House of Commons. This is to be distinguished from by-elections and local elections. Each **constituency**, or seat, in the UK elects one MP (Member of Parliament) to a seat in the House of Commons. By 'constituency' we mean a geographical area of the UK which is represented by one MP in the House of Commons. These MPs are elected from a choice of candidates by a simple majority system in which each person casts one vote. The candidate with the most votes then becomes the MP for that constituency. Candidates may come either from a political party registered with the Electoral Commission (an independent body accountable directly to the UK Parliament, which regulates elections in the UK, promotes voter awareness and works to build confidence in the electoral process[1]), or may stand as an 'Independent' rather than represent a registered party. The political party which wins a majority of **seats** (not individual votes) in the House of Commons typically forms the Government. On two occasions in the post-war period, the party winning the most number of votes did not win the most number of seats. Once was in 1951, when Labour

[1] Source: http://www.parliament.uk/about/how/elections/general.cfm.

had a larger vote share than the Conservatives but gained fewer seats, and the other was in 1974, when it was the other way round. Since both elections led to a change in government, suffice to say it is more important to analyse number of winning seats than votes.

These general elections traditionally occur on a Thursday; the last general election not on a Thursday was in 1931. They must be held within five years and one month of the previous one, but are often held before then, since it is up to the parties in government when to call a general election, and not all Parliaments run for the whole five-year period. For instance, the current Labour Party government has held general elections every four years since coming to power in May 1997 and thereafter in June 2001 and May 2005; therefore, another election is not *legally* obliged to occur until June 2010. This five-year limit may be varied by an Act of Parliament, as was so during both World Wars; the Parliament elected in 1910 was prolonged to 1918, and that elected in 1935 lasted until 1945. The House of Lords has an absolute veto on any Bill to extend the life of a Parliament.

We collectively define all the people in the UK eligible to vote as the **electorate**. Typically around three-quarters of this electorate vote in what is known as the **turnout**; however, this was only 59 per cent in 2001. Most voting occurs in polling stations, but anyone eligible to vote may apply for a postal vote. British citizens living abroad are also entitled to a postal vote as long as they have been living abroad for less than fifteen years.

2.2 Constituencies and the UK

Currently, there are 646 constituencies in the UK: 529 (82%) in England, 59 (9%) in Scotland, 40 (6%) in Wales and 18 (3%) in Northern Ireland. Recall that in each constituency there are a few candidates (typically between four and eight), from which one is voted to occupy the seat - there was a total of 3,320 in the UK in 2001. Constituencies declare their winner at some time in the period during election night, at varying rates of flow, which in practice may last a whole day after polls close (at 10 p.m.); for example, in 2001 the final seat declared at 10.17 p.m. the next day. Sometimes, a constituency may have to recount for whatever the reason.

Certain parties will not try to win certain constituencies due to location; an obvious example is the Scottish National Party, which will not try to win any constituencies within the London Boroughs due to their location outside Scotland.

Constituency boundaries are from time to time redrawn, usually to reflect changes in the population distribution; there were 630 in 1971. Subsequently, major and minor changes may be carried out within constituencies.

Constituencies in urban areas, where ballot boxes are fairly centralised, tend to declare earlier than those in rural areas. The urban seats, as well as being predominantly Labour, have in the past shown rather different behaviour, in terms of movements between parties, from the predominantly Conservative rural seats where Liberal also have their strongest support. Many Scottish and Welsh seats where nationalists stand tend to declare late. A particular difficulty in a close-run election is that many of the seats with contests of a unique character are among the latest declarers (Brown and Payne, 1975, [23]).

2.3 Parties

Between 1945 and 1970, the majority of seats were won by **Labour** or **Conservative**, and each has put up a candidate in almost every constituency. The **Liberal** party, until 1974, received a small percentage of votes cast and won few seats, despite having a large number of candidates. In 1974 however, they received over a quarter of the total vote in seats where they contested (517 constituencies in 1974, compared with 332 in 1970).

Also, since 1964 **nationalist** parties have had candidates in most of the Scottish and Welsh seats, and achieved some notable successes in 1974. In the Ulster constituencies political issues have always different from those in Great Britain; as a result, contests there are primarily between the Ulster Unionists (taken as Conservatives before 1974), a number of significant Independent candidates and a variety of local parties representing Republicans and Labour. Consequently, minor parties have become more noticeable in more recent elections - for instance, in 1992 all parties excluding the main three accounted for 35 per cent of all candidates up for election, compared to 18 per cent in the 1987 election (see Table 2.2). (Table 2.1 shows all parties which stood for election in 2001, as an example.) Also, there has been more volatility in voting behaviour.

Essentially now, we have a three-party (and in some places four-party) contest. As we have already mentioned, there is interest not only in predicting which party will win, but also by how many seats. Figure 2.1 summarises the percentage of seats won by the dominating three parties in the last eight elections, as well as a combined total for the remainder. We see that a change in government took place in 1979; Conservative then remained in power until 1997, when Labour regained power and they have held it ever since. From the graph it appears as if it is

Alliance Party of Northern Ireland	People's Justice Party
British National Party	Pro-Life Alliance
Conservative and Unionist Party	Reform 2000 Party
Communist League	Rock and Roll Loony Party
Countryside Party	Socialist Alliance
Christian Peoples' Alliance	Social Democratic and Labour Party
Communist Party of Britain	Sinn Fin [We Ourselves]
Communist Party of Great Britain*	Scottish Green Party
Democrat	Socialist Labour Party
Democratic Labour Party*	Scottish National Party
Green Party	Socialist Outlook*
Independent	Mr Speaker seeking re-election
Left Alliance	Socialist Party
Labour Party	Scottish Socialist Party
Labour Party and Co-operative Party joint candidate	Scottish Unionist Party
Legalise Cannabis Alliance	Socialist Workers' Party*
Liberal Democrat	Third Way
Mebyon Kernow - The Party for Cornwall	Ulster Democratic Unionist Party
Official Monster Raving Loony Party	UK Independence Party
New Britain Party	United Kingdom Unionist Party
National Front	Ulster Unionist Party
Northern Ireland Unionist Party	Women For Life on Earth
Northern Ireland Women's Coalition	Workers Power*
Plaid Cymru the Party of Wales	Wessex Regionalist
Progressive Democratic Party	Workers' Revolutionary Party
Pro Euro Conservative Party	-

Table 2.1: All parties with candidates standing for election in the 2001 election ('*': part of the Socialist Alliance)

still a two-party contest. However, Liberal Democrats have more recently won a continually-increasing proportion of seats, as may be seen in the graph. The combined remainder has stayed reasonably stable over the period shown.

Also, Table 2.2 shows the figures for percentage of votes won (which, as we expect on the whole, are positively correlated with the percentage of seats won) and percentage of candidates standing.

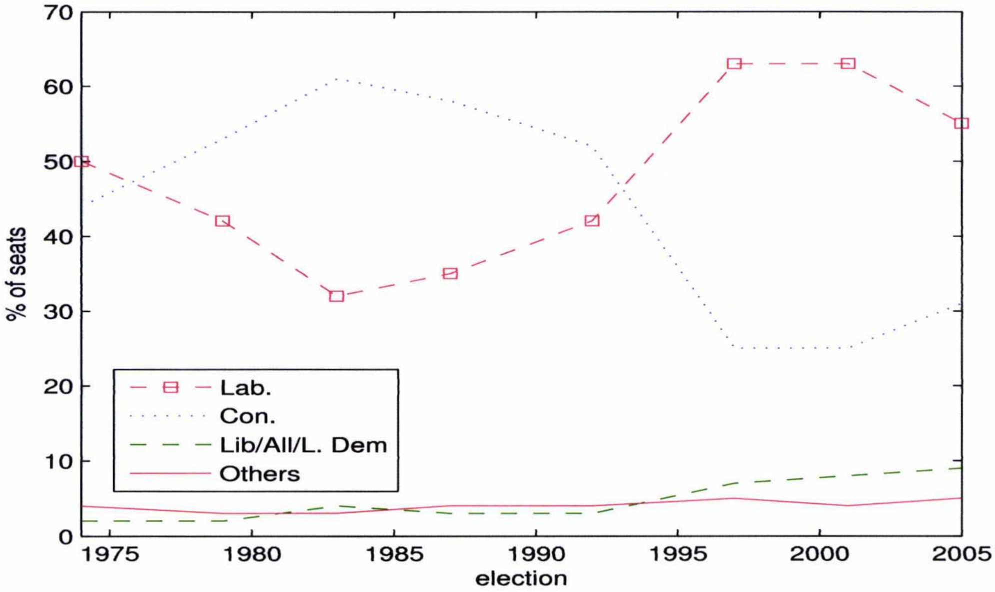


Figure 2.1: Percentage of seats won by Labour, Conservative, Liberal Democrats (Liberal before 1983, Alliance in 1983 and 1987 elections) and all Others in each general election between October 1974 and 2005.

2.4 Key Events

2.4.1 A recent history

The October 1974 election was the second one that year, due to a **hung parliament** (where no one party has the clear majority) in the February that year.

In 1983, the Alliance party formed, which was a pact between the Liberal Party and Social Democratic Party. Soon after the 1987 election this evolved into the Liberal Democrats. Also, new boundaries were drawn up; for comparison, some pollsters worked out how the previous election would have turned out if these new boundaries were set then. This is typical practice.

The 1992 opinion polls and, to a lesser extent, exit polls (see next section) suggested a hung parliament would occur, whereas Conservative won by a noticeable majority of 21 seats. (On the election night, the results-based forecasts moved quite lethargically towards the final outcome.) Due to this, the Market Research Society in 1994 produced a report analysing the whole election including its run-up. The Society concluded that there was too much reliance on opinion and exit polls in the forecasting. This led to considerable scepticism in the media about the accuracy of both opinion and exit polls in the run-up to the 1997 general

Election	Number (%)	Labour	Conservatives	Democrats	Others	Total
Oct 1974	Candidates	623 (28)	622 (28)	619 (27)	388 (17)	2,252
	Seats	319 (50)	277 (44)	13 (2)	26 (4)	635
	Votes	11,457,079 (39)	10,462,565 (36)	5,346,704 (18)	1,922,756 (7)	29,189,104
1979	Candidates	623 (24)	622 (24)	577 (22)	754 (29)	2,576
	Seats	269 (42)	339 (53)	11 (2)	16 (3)	635
	Votes	11,532,218 (37)	13,697,923 (44)	4,313,804 (14)	1,677,417 (5)	31,221,362
1983	Candidates	633 (25)	633 (25)	633 (25)	679 (26)	2,578
	Seats	209 (32)	397 (61)	23 (4)	21 (3)	650
	Votes	8,456,934 (28)	13,012,316 (42)	7,780,949 (25)	1,420,938 (5)	30,671,137
1987	Candidates	633 (27)	633 (27)	633 (27)	426 (18)	2,325
	Seats	229 (35)	376 (58)	22 (3)	23 (4)	650
	Votes	10,029,270 (31)	13,760,935 (42)	7,341,651 (23)	1,398,348 (4)	32,530,204
1992	Candidates	634 (21)	645 (22)	632 (21)	1,038 (35)	2,949
	Seats	271 (42)	336 (52)	20 (3)	24 (4)	651
	Votes	11,560,484 (34)	14,093,007 (42)	5,999,384 (18)	1,961,199 (6)	33,614,074
1997	Candidates	639 (17)	648 (17)	639 (17)	1,798 (48)	3,724
	Seats	418 (63)	165 (25)	46 (7)	30 (5)	659
	Votes	13,517,911 (43)	9,600,940 (31)	5,243,440 (17)	2,925,817 (9)	31,288,108
2001	Candidates	640 (19)	643 (19)	639 (19)	1,397 (42)	3,319
	Seats	412 (63)	166 (25)	52 (8)	29 (4)	659
	Votes	10,724,953 (41)	8,357,615 (32)	4,814,321 (18)	2,470,494 (9)	26,367,383

Table 2.2: Candidates standing, seats won and votes won for Labour, Conservatives, Democrats (Liberal before 1983, Alliance in 1983 and 1987 elections) and all Others in each UK election between October 1974 (there were *two* elections in 1974) and 2001.

election.

In 1997, due to the unpopularity of Conservative, the governing party, it was thought that there would be a lot of tactical voting occurring. New constituencies resulted from major boundary changes. The election took place on the same day as local elections in some parts of the UK, which meant that Scottish, Welsh and urban English (predominantly safe Labour) seats declared earlier than usual.

Boundaries were redrawn again in 2005. Another key feature of this election was that the BBC and ITV merged their exit poll data to form one set of results; the prediction of the final win result was consequently extremely accurate. Curtice and Firth (2008, [32]) cover possible reasons for this, which are summarised below:

- There was a shifted focus to estimate *change* in support rather than the level of support, using previous exit-poll data.
- An attempt was made to estimate systematic variation in the change in party support.

- There was a probabilistic approach to forecast the outcome in seats.
- Data regarding postal voters were obtained before polling day.
- Newer graphical techniques were exploited, as well as inference and reestimation using new developments in I.T.
- Refusals were estimated by guessing for whom respondents and non-respondents would have voted.

Full detail on these issues is covered in Curtice and Firth (2008).

The next general election is legally obliged to take place no later than June 2010, when new boundaries will be drawn up.

2.4.2 1983 revolution

The 1983 election was a landmark in British politics due to the extent to which across the board politicians, the media and voters used polls to inform themselves and others about what was happening in the election campaign. Telephones were used for interviewing, and there was widespread use of one-day 'quickie' polls. The political parties also commissioned their own polls during the campaign. Additionally, there were more potential sponsors, due to the increased realisation by newspapers that sponsoring a poll was good to secure free television advertising of the paper the night before. Developments in computer technology made it possible for opinion polls to conduct elaborate analyses of their data to check sample representativeness, and to provide more detailed breakdowns of the political outlooks of different groups within the population, such as the young and elderly, men and women, and different social classes. This was all reflected in the somewhat larger number of opinion poll recordings available that year (and so will be a natural choice of dataset when we perform our modelling later).

2.5 Sources of Information

We would like our forecasts to be as accurate as possible (and ideally as early as possible), to be of use in practice. To help steer our forecast, there are fundamentally four guides to consider: opinion poll data[2], exit poll data, results from previous election(s) and the results declared per constituency on election night.

[2] In this thesis, when we refer to opinion polls we mean polls of the voting intentions of the electorate.

2.5.1 Opinion polls

Bean (1948, [8]) defines 'polling' as essentially the art of obtaining from a remarkably small sample an indication as to how the entire population of the country, state, or community reacts to a particular issue or candidate. He refers to polls as a scientific device for obtaining a cross section of public opinion at a given time. Such a cross section makes it possible to bring historical trends up to date and to project them into the future. Thus, polls of public opinion are now widely used for general elections, and are a generally accepted feature of political analysis and judgement. When concerned with the prediction of elections, as well as the historical relationships between parties and past trends, polls provide invaluable current information, which supplement and bring up to date these former influences. 'As the few go, so go the many,' is the basic principle in public opinion polling.

It is important of course that the polls are reliable for use. For example, nationality and economic groups often vary greatly in party preferences, and have been seen to shift their political allegiances abruptly. Local polls, to be most successful, must represent all the elements in a community, geographic, economic, and cultural; national polls, to be most accurate, must embrace communities in each of the significant regions.

Bean (1948) provides a full discussion on polls. He states that it is after the conventions have chosen the candidates that the combination of political history and opinion polls find greatest usefulness in predicting the election outcome.

On the whole, polls tend to share common technical features. They typically involve a **quota sample** of about 1000 interviews, selected according to age, sex and social class, as well as region and constituency. A concern with speed in identifying respondents, as well as a marked deterioration in the electoral register, meant that no polling organisation wanted to undertake pure random samples, with all the additional expense in time and money.

Since 1970, organisations have taken major steps to accelerate the conduct of polls, especially the final poll forecasting the result. Obviously, organisations differ in what they do, and there is no agreement about how best to conduct a poll quickly as well as accurately. However, the belief with polls is that, in principle, different organisations asking the same question ought to get results within a few per cent of each other. Despite all this, it is interesting to note that in every general election since 1964, polls have consistently overestimated the vote for the governing party of the day.

In an election campaign in which there is significant last-minute change in electoral opinion, the ability to poll up to election day may be crucial in accurate forecasting. Otherwise, superior accuracy of late polls cannot be assumed; rather, it would be more reliable to consider as long a period of time as possible, over which polls are carried out. A **poll of polls** is conducted, which is simply the arithmetic average of each separate poll conducted over a build up of time. This shows far greater stability than the reports of any one poll from week to week. However, the volume of polling reflects the willingness of the media to spend money. Indeed, the influence of the polls has some (albeit minor) impact on steering the final choice.

It has been known that some politicians now read the polls in search of guidance for influencing voters, so as to try to change subsequent poll figures to their own advantage.

The creation of the Alliance (which evolved into the Liberal Democrats) expanded the demand for opinion polls by political parties. An opinion poll is a survey of opinion from a sample of the public, which is intended to represent opinions of the entire population (by extrapolating generalities in ratio or within confidence intervals).

Public opinion polls provide current information, and collectively cover a reasonable length of time in the run-up to the election night to be able to assess changes. Whiteley (1979, [108]) states that they are the only reliable data available for purposes of election forecasting prior to polling day. They offer politicians and voters a chance to improve the steering capacity of the electoral process, by providing more accurate information about voters' views than politicians could otherwise obtain at the ballot box. As the former Prime Minister James Callaghan once remarked, 'If the people cannot be trusted with opinion polls, then they cannot be trusted with the vote (Rose, 1985, [98]).'

Pollsters: Key pollsters (polling organisations) in the UK are MORI (Market & Opinion Research International), which only selects those who say that they are likely to vote, YouGov (online), Populus (by The Times), Communicate Research, ICM (Independent Communications and Marketing) and GfK NOP (National Opinion Polls). All the major television networks do their own polling, alone or in conjunction with the largest newspapers or magazines, either in collaboration or independently. Several organisations try to monitor the behaviour of polling arms and the use of polling and statistical data.

Sample and polling methods: In the past opinion polls were conducted via telephone or person-to-person contact. Methods and techniques vary, though they are widely accepted. Some pollsters such as YouGov use Internet surveys, where a sample is drawn from a large panel of volunteers. This is in contrast to using a scientific sample of the population, and are they therefore not generally considered as accurate.

Failures: Overall, prediction using polls failed to predict the Conservative victories of 1970 and 1992, and Labour's victory in 1974. However, poll figures at other elections have led to generally accurate predictions.

Sources of error:

- Polls are subject to **sampling error**, which measures the effects of chance and uncertainty in the sampling process. We express the uncertainty as a *margin of error* (usually the confidence interval for a particular statistic). This may be reduced with a larger sample in theory; however in practice, pollsters must balance the cost against the reduction in sampling error; a sample size of around 500 to 1,000 is a typical compromise for political polls.

- Nonresponse bias - some people do not answer calls from strangers, or refuse to answer the poll, and the characteristics of those who are interviewed may differ widely from those who decline. In terms of election polls, studies suggest that these selection bias effects are small, but each pollster performs its own way to minimise the bias.

- Response bias - this is where the answers given do not reflect true beliefs. This could be deliberately set up by a pollster to generate a certain result or please its clients, but more commonly is a result of the wording or ordering of questions. Respondents may seem more extreme than they actually are in order to boost their argument, or give rushed and thoughtless answers in order to end the survey. Respondents may also feel under social pressure not to give an unpopular answer, and thus polls would not reflect all attitudes within the population. This effect may be magnified if the results of surveys are widely publicised.

- Wording of questions - for instance, the public may be more likely to support a person described by the surveyor as one of the 'leading candidates'. Comparisons between polls often boil down to this issue. On some issues, question wording may lead to quite pronounced differences between surveys (Cantril and Hadley, 1951, [29]). This may also, however, be a result of conflicted feelings or evolving attitudes, rather than a poorly-designed survey. Common techniques are to rotate the order in which questions are asked or to split-sample. The latter involves having two different versions of a question, with each version presented to half the sample.

- Coverage bias - this is the use of samples unrepresentative of the population due to the methodology used. Blumenthal (2008, [12] and [13]) explains how this issue came to prominence during the 2008 US presidential election. In previous elections, the proportion of the general population using mobile phones was small, but as this proportion has increased, the worry is that polling only landlines is no longer representative of the general population. Pollsters have developed many techniques to help overcome this, to varying degrees of success.

- Failure to vote - the people surveyed in the opinion polls may not actually vote; the larger this problem is the less useful the opinion poll data will be.

Influence: Opinion polls may sometimes influence the behaviour of electors. There are three schools of thought:

1. A **bandwagon effect** is when the poll encourages voters to support the candidate shown to be winning in the poll. The opposite is the 'underdog effect', when people vote out of sympathy for the 'losing' candidate; there is less empirical evidence for this than for the former.

2. **Tactical voting** is when voters choose not the candidate whom they prefer but another, less-preferred, candidate from strategic considerations. An example was in the 1997 election. The constituency of Enfield Southgate was believed to be a safe seat for Conservative but opinion polls showed the Labour candidate steadily gaining support, which may have prompted undecided voters or supporters of other parties to support Labour here in order to remove the Conservatives.

3. A **boomerang effect** is when the likely supporters of the candidate shown to be winning feel that chances are slim and that their vote is not required, thus allowing another candidate to win.

A popular example of opinion polls leading to errors was the 1992 general election. Despite pollsters using different methodologies, almost all polls in the run-up to the vote, and to a lesser extent exit polls (see next part) taken on voting day, showed a lead for Labour but the actual vote gave victory to Conservative. Explanations were:

- Late swing - voters who changed their minds shortly before voting tended to favour the Conservatives;
- Nonresponse bias - Conservative voters were less likely to participate in surveys than in the past and were thus underrepresented; and
- The Shy Tory Factor - the Conservatives had a sustained period of unpopularity as a result of economic difficulties and minor scandals, leading to Conservative supporters being reluctant to disclose their sincere intentions to pollsters.

The relative importance of these factors was, and remains, a matter of controversy, but since then pollsters have adjusted methodologies and achieved more accurate results in subsequent elections.

Much work has been done to explain erroneous polling results. Some of this has blamed the errors on pollsters, some due to natural statistical error, and others have blamed the respondents for not giving reliable answers.

2.5.2 Exit polls

An exit poll is taken immediately after the electors have left the polling stations. Unlike an opinion poll, which asks whom the voter intends to vote for, an exit poll asks whom the voter actually voted for. The final outcome takes hours to count and so pollsters conduct exit polls to gain an early indication. Over the longer term, exit polls are also used to collect demographic data about voters and to find out why they voted as they did. Since actual votes are cast anonymously, polling is the only way of collecting this information. As with opinion polls, here we get current information on the public opinion, in fact, even more current than the former. If we also assumed honesty of answers and a good response rate then we would expect a good correlation of the results with that of the final outcome.

Problems: As with opinion polls, exit polls naturally come with a margin of error. A famous instance of error was in the 1992 election, when two exit polls predicted a hung parliament. The actual vote revealed that Conservative held their position, though with a significantly-reduced majority. Investigations into this failure identified a number of causes including differential response rates (the Shy Tory Factor mentioned earlier), the use of inadequate demographic data and poor choice of sampling points. Issues with exit polls have encouraged pollsters to pool data to enhance accuracy. For example, during the 2005 election the BBC and ITV merged data on exit polls, which led to an exact prediction of the number of seats won per main party.

Criticism: Criticism has occurred in cases where exit poll results have appeared to provide and/or have provided a basis for projecting winners before all real polls have closed, thereby possibly influencing election results. As a result, in the UK it is now illegal to release exit poll figures before the polling stations have closed.

2.5.3 Previous election results

The results of previous general elections, particularly the election directly before the current, would also help shape the forecast. The main drawback with them is that, compared to polling which obtains current information, they are old results, whereas public opinions and voting are volatile. However, in a forecasting model we could firstly derive forecasts solely using the previous outcomes, and then introduce variables, say, to account for any important changes in the period between the previous election(s) and the current one, which will or may have an impact. This is the method introduced by Brown and Payne (1975, [23]) and detailed in Chapter 8.

2.5.4 Election night

The polls close at 10 p.m. Various television channels, including the British Broadcasting Corporation (BBC), Independent Television (ITV) and Sky begin their all-night coverage. Primarily, this involves forecasting of the results of the as yet uncounted votes, to predict the winning candidate in each constituency, and from this the number of seats won per party. Particular interest is focused on the prediction of the party winning the most seats and its *majority* over other parties. Throughout the night there is commentary on particular constituencies as they declare their results, as well as on the subsequent state of the parties. Importantly, the forecasts are *constantly updated.*

These forecasts form the main discussion, involving both psephologists (election analysts) and politicians. In Chapter 4, we will look at the types of forecasting methods in detail, and in Chapter 8 we will concentrate on the BBC method, which has on the whole remain unchanged since it began in 1974.

2.6 Review

This chapter has summarised the nature of general elections in the UK to set the scene for subsequent chapters. The process in the run up to the election night involves various polls, mainly in order to gauge voting intention. We described election night itself, and how in recent years the parties most dominant in the voting have been Labour, Conservative and Liberal Democrats. This has led to a shift from a two-party contest to a three-party contest, although the former two have so far still proven to account for greater proportions of seats (and votes) than the latter party. Next, we assessed in some detail the part which polling plays, focusing on opinion polls (which ask whom the respondents *intend* to vote for) and exit polls (which ask voters whom they actually just *did* vote for).

As we have emphasised, there is a great interest in forecasting the number of seats won by each party before all seats have been declared on the election night. In response, several different approaches are taken to try to do this. Shortly, we will review some of these, before introducing our own new method. We outlined the various sources of information available to assist in this, and we will make use of these data in our approach. To appreciate some of these methods as well as the models which we will end up using, we provide next an overall account of the subject of time series.

Chapter 3

Time Series Modelling

3.1 Introduction

Time series is concerned with a number of observations ordered in time. The subject has several applications to real life. Examples are numerous, including the GNP of a country, sales or profit margins of a business, exchange rates, stock prices and insurance claims.

There are quite a few motivations for using time series models, an important one of which is forecasting into the future. This is clearly important for the examples stated above, as it helps governments and businesses plan the future effectively, implementing changes where necessary. For instance, concerning the stock price of tomorrow or next year the investor can forecast, using time series modelling, to decide whether or not to invest.

Another important use of time series is in testing economic theories. A simple yet classic model is the **random walk** hypothesis,

$$y_t = y_{t-1} + \epsilon_t,$$

with ϵ_t some random disturbance, where usually $\mathbb{E}(\epsilon_t) = 0$. A more general model is

$$y_t = a_0 + a_1 y_{t-1} + \epsilon_t, \tag{3.1}$$

in which a_1 is known as the **root**.

Time series may be used for either the sequence of random variables or their realisations[1]. The method involves decomposing the observed data into building blocks or components which we can interpret. We aim to forecast each component and then reassemble all forecasts to forecast the series as a whole. The

[1] For this reason, note that we will in this thesis alternate between the two.

decomposition generally takes the form

$$y_t = T_t + S_t + I_t, \tag{3.2}$$

in which

- T_t is the value of the trend (e.g. $T_t = 1 + 0.1t$)
- S_t is the seasonal component (e.g. $S_t = 16sin(t\pi/6)$); and
- I_t is the irregular part (e.g. $I_t = 0.7I_{t-1} + \epsilon_t$).

Above, T_t and S_t are both **deterministic** or predictable, whereas I_t contains both a predictable part and a random **stochastic** part, in which the ϵ_t is the random disturbance. T_t, S_t and I_t are all difference equations, in which variables are expressed as a function of their own **lagged** values, time and other variables.

Assume that we have observations as a function of a variable t, which take discrete and equally-spaced values, i.e. we go from t to $t+1$ and so on. Therefore, we have a sequence of random variables or their values

$$\{\ldots, y_{t-2}, y_{t-1}, y_t, y_{t+1}, y_{t+2}, \ldots\},$$

denoted by $\{y_t\}$, and where we let t stand for 'time' (measured in any unit such as days, months, years). The first difference is the **change**: $\triangle y_t = y_t - y_{t-1}$ and the second is the **acceleration**: $\triangle^2 y_t = \triangle(\triangle y_t) = y_t - 2y_{t-1} + y_{t-2}$.

This chapter breaks down into two main parts. The first part focuses on general time series theory for what is known as the '**autoregressive model**', which relies on the assumption of covariance stationarity, which we will define. We will look in some detail at various properties, including the conditions for covariance stationarity, autocovariance (covariance within a particular time series dataset), specific types of model and how to choose the best of these types given a dataset, methods to estimate parameters, particular inference tests, forecasting of future values and volatility in the data. The second part focuses first on a time series model with the assumption of strict stationarity, which we will also define, as well as more complex modifications of this stationary model to cater for non-stationarity. In these latter models we introduce the idea of probabilities as latent variables, which will become important in subsequent chapters. Again, we study in detail the properties of each type of model, providing illustrations for clarity.

3.2 The Autoregressive Model

3.2.1 Covariance stationarity

Observed time series is a realisation of a stochastic process, that is, any collection of random variables $y(t)$, $t \in \Omega$ defined on a common probability space, where t denotes time. If Ω is a discrete set then observations are only taken at (usually equally-spaced) specific times and we talk of a **discrete-time** time series. Instead, if Ω is an interval then observations are made continuously in time and we have a **continuous-time** time series. We only deal with the former in this thesis.

The stochastic process is described by a probability distribution for $\{y_t\}$, where elements are typically *not* independent. Underlying is a joint density function, say, $f(y_t,\, y_{t-1},\, y_{t-2},\, \ldots)$, which is usually expressed as a transition density function, say, $f(y_t|\, y_{t-1},\, y_{t-2},\, \ldots)$. We somehow need to characterise the probability distribution and may do so using moments ($\mathbb{E}(y_t) = \mu_t$, $\mathbb{V}ar(y_t) = \mathbb{E}[(y_t - \mu_t)^2]$, $\mathbb{C}ov(y_t,\, y_{t-s}) = \mathbb{E}[(y_t - \mu_t)(y_{t-s} - \mu_{t-s})]$, $\mathbb{E}_t(y_{t+i}) = \mathbb{E}(y_{t+i}|\, y_t,\, y_{t-1},\, y_{t-2},\, \ldots)$, $\forall t,\, s$, and so on).

To make all this usable in practice for the analysis of time series, we cannot have all these moments change with every time period. Therefore, we often assume that they remain constant over time.

Definition: A stochastic process with finite mean and variance is *covariance stationary* if, for all t and s,

$$\begin{aligned}
\mathbb{E}(y_t) &= \mu \\
\mathbb{V}ar(y_t) &= {\sigma_y}^2 \\
\mathbb{C}ov(y_t,\, y_{t-s}) &= \gamma_s,
\end{aligned}$$

in which μ, ${\sigma_y}^2$ and γ_s are constant parameters. Clearly, $\gamma_0 = {\sigma_y}^2$ and $\gamma_k = \gamma_{-k}$.

A stronger condition is strict stationarity (see later), in which the entire probability distribution is unaffected by a change of the time origin. Under normality, these concepts coincide.

3.2.2 The basic model

An **autoregressive (AR) equation** takes the form

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \epsilon_t,$$

in which p is the order (thus the model is denoted by AR(p)), that is, the total number of lags we wish to consider and ϵ_t is the outside influence. One way to solve such equations is by iteration.

Condition for stationarity

Equation (3.1), AR(1), converges if and only if

$$|\, a_1 \,| < 1. \tag{3.3}$$

Now, if we let $a_0 = 0 = \epsilon_t$ then we are left with the *homogeneous* part, which is solved by the homogeneous solution

$$y_t = A{a_1}^t,$$

for some constant A. If we have a difference equation of order n then there exist n roots, which may be real or imaginary.

Generally, stability requires that all characteristic roots lie *within the unit circle.* The homogeneous solution governs the stability of the variable, even after we add y_0 and ϵ_t to the process.

Under unit roots (roots equal to unity), the solution will have a polynomial time trend, of order equal to the number of unit roots.

3.2.3 White-noise process

Definition: A sequence $\{\epsilon_t\}$ is a white-noise process if, for all t,

$$\begin{aligned}
&\mathbb{E}(\epsilon_t) = 0\\
&\mathbb{V}ar(\epsilon_t) = \sigma^2\\
&\mathbb{C}ov(\epsilon_t,\, \epsilon_{t-s}) = 0,\ \forall s \neq 0.
\end{aligned}$$

This is the simplest possible time series model. Every identically and independently distributed (i.i.d.) process with mean 0 and variance σ^2 is white noise (but not conversely). This model is the basic building block of time series models because of the following theorem.

Wold decomposition: Every covariance stationary stochastic process may be written as the sum of a deterministic part and an infinite **moving average (MA)** of uncorrelated variables

$$y_t = \mu + \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}, \tag{3.4}$$

with $\psi_0 = 1$ and $\{\epsilon_t\}$ a white-noise process. This implies that

$$\mathbb{E}(y_t) = \mu$$

$$\mathbb{V}ar(y_t) = \sigma^2 \sum_{i=0}^{\infty} {\psi_i}^2.$$

Importantly, the $\{\epsilon_t\}$ are uncorrelated but the $\{y_t\}$ are not. Typically, it is assumed that $\sum_{i=0}^{\infty} | \psi_i | < \infty$, so that the variance exists.

The finite version of the so-called *linear-filter representation* in (3.4) is a moving-average (MA) process of order q or MA(q):

$$y_t = \sum_{i=0}^{q} \beta_i \epsilon_{t-i}.$$

3.2.4 ARMA models

We now combine the ideas of the AR(p) and MA(q) to obtain

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{t=0}^{q} \beta_i \epsilon_{t-i}, \tag{3.5}$$

which is known as the autoregressive moving average model ARMA(p, q). Obviously, the AR(p) and MA(q) are specific cases of this model. Similar conditions for stationarity apply as in (3.3), but now we have more characteristic roots due to having more lags:

$$y_t = \alpha^t \sum_{i=1}^{m} A_i t^{i-1} + \sum_{i=m+1}^{p} A_i {\alpha_i}^t,$$

where α is repeated m times and the other roots are $\alpha_{m+1}, \ldots, \alpha_p$. If one or more of these roots is on the unit circle then we say that $\{y_t\}$ is an autoregressive *integrated* moving average (or ARIMA) process. Brockwell and Davis (2002, [21]) refer to this property as *causality*, as it implies that y_t may be expressed entirely in terms of ϵ_s for $s \leq t$.

Invertibility

We may therefore assert that the MA part does not contribute to the stationarity, yet if we want to estimate a model with a MA part then we often impose that it is **invertible**, which essentially means that its parameters are uniquely determined by its autocorrelation function (see next part). This would hold if and only if the characteristic roots of the MA polynomial are all within the unit circle. Note that invertibility implies that the error term ϵ_t may be expressed as a function of y_s, for $s \leq t$.

Generally, a stationary finite order AR(p) may be written as an infinite order MA with restricted parameters, and an invertible finite order MA(q) may be written as a restricted infinite order AR.

3.2.5 Autocorrelation function

We will denote γ_s to represent the 'autocovariance' for the lag s. Using this,

$$\rho_s = \frac{\gamma_s}{\gamma_0} \tag{3.6}$$

is called the **autocorrelation** between y_t and y_{t-s}, which depends only on the lag s and not on the actual time t for the stationary series. Clearly, $\rho_0 = 1$. Then, a plot of ρ_s versus s is the autocorrelation function (ACF).

In practice, we have observations (y_1, y_2, ..., y_n) and can estimate the above by:

$$\begin{aligned}
\hat{\mu} = \bar{y} &= \left(\sum_{t=1}^{n} y_t\right) / n \\
\hat{\sigma}_y^2 &= \sum_{t=1}^{n} (y_t - \bar{y})^2 \\
\hat{\gamma}_s &= \left(\sum_{t=s+1}^{n} (y_t - \bar{y})(y_{t-s} - \bar{y})\right) / n.
\end{aligned}$$

These are consistent estimators if observations which are very far apart from one another are (almost) uncorrelated.

We may define the sample ACF, r_s say, as $r_s = \hat{\gamma}_s / \hat{\gamma}_0$, and then test for significance of each lag.

3.2.6 Partial autocorrelation function

Consider again AR(1). It may be shown that its ACF takes the form

$$\rho_s = {a_1}^s, \tag{3.7}$$

and so y_t is correlated with all y_{t-s}. This feature will become important in Subsection 3.3.1, when discussing one of the key models which we will end up using in the thesis. However, for $s > 1$ the correlation mentioned occurs indirectly, through *intermediate lags*. The **partial autocorrelations**, represented by ϕ_{ss}, eliminates this effect, whereby we subtract the mean of y and then consider the regression coefficient of y_{t-s} in an AR(s).

Generally, partial autocorrelations may be derived from the autocorrelations via the **Yule-Walker** equations: this is where ϕ_{ss} is the solution of a_s, which is expressed in terms of ρ_1, ρ_2, ..., ρ_s of the first s Yule-Walker equations for an AR(s).

As with r_s, in practice we require a sample partial ACF (PACF), and thus make use of r_s in deriving it. It may be shown that under the null hypothesis of an AR(p) the estimated $\hat{\phi}_{ss}$ for lags $s > p$ has $\mathbb{V}ar(\hat{\phi}_{ss}) \approx 1/n$.

Table 3.1 summarises the behaviour of both the ACF and PACF depending upon the types of time series models which we have currently considered.

Model	ACF	PACF
AR(p)	$\neq 0 \; \forall s$; decays to 0	$\neq 0$ for $s \leq p$; 0 for $s > p$
MA(q)	$\neq 0$ for $s \leq q$; 0 for $s > q$	$\neq 0 \; \forall s$; decays to 0
ARMA(p, q)	$\neq 0 \; \forall s$; decays for $s \geq q$	$\neq 0 \; \forall s$; decays for $s \geq p$

Table 3.1: Properties of the ACF and PACF.

3.2.7 Non-stationarity

Many time series which correspond to observed variables are not stationary, however. There are many instances in which series look as though either the variance or the mean are not constant over time.

Nonstationarity in the variance

When a variance changes over time, we call this 'heteroskedasticity'. Often, a transformation such as a logarithmic transformation stabilises the variance. We are often interested in modelling these changes in variance explicitly, as will be seen in Subsection 3.2.10.

Nonstationarity in the mean

One way to deal with a non-constant mean is to add a simple linear (or higher order) trend:

$$y_t = \alpha_0 + \alpha_1 t + \epsilon_t, \tag{3.8}$$

possibly including more dynamics (ARMA). However, this would imply that the trend will continue in the same way over time, which may be unrealistic. This is known as a **deterministic** trend. However, there may exist nonstationarity through the difference equation itself, known as a **stochastic** trend.

For example, consider the random walk model again, but now with a drift a_0, i.e.

$$y_t = a_0 + y_{t-1} + \epsilon_t. \tag{3.9}$$

Its solution (noting that $a_1 = 1$) is

$$y_t = a_0 t + y_0 + \sum_{i=1}^{t} \epsilon_i,$$

so there is both a deterministic ($a_0 t$) and stochastic trend (the sum of all previous disturbances). Also,

$$\begin{aligned} \mathbb{E}(y_t) &= y_0 + a_0 t && \text{dependent on } t \\ \mathbb{V}ar(y_t) &= t\sigma^2 && \text{dependent on } t \\ \mathbb{C}ov(y_t, y_{t-s}) &= (t-s)\sigma^2 && \text{dependent on } t, \end{aligned}$$

so (3.9) is clearly not stationary.

Furthermore, $\rho_s = (1 - s/t)^{1/2}$, which decreases slowly (linearly not exponentially), which implies the existence of a unit root. However, (3.9) is an integrated process, meaning that a *first difference* produces a stationary model:

$$\triangle y_t = a_0 + \epsilon_t.$$

Generally then, if a model needs a d^{th}-order differencing to make it a stationary-invertible ARMA(p, q) model then we call it an ARIMA(p, d, q) model, which has d unit roots. We use this method of **differencing** to remove a stochastic trend. Such a model is called a difference-stationary model. Reconsider (3.8). We would first regress $\{y_t\}$ on a polynomial trend and then estimate the difference between the actual and estimated values (detrended process). We use a method known as **detrending** to remove a deterministic trend. Such a model is called a trend-stationary model. Often, the realisations of these two types of model look very much alike.

The problem is often that stationary and unit-root processes look alike on the basis of small samples. It would be helpful to have a formal way of testing unit roots. However, the usual asymptotic theory no longer applies under the null of a unit-root process. Instead, we use **Monte Carlo** experiments, where a large number of samples is generated from known processes with a unit root, in order to investigate the sampling behaviour of test statistics.

3.2.8 Model fitting and testing

Box and Jenkins (1970, [18]) advocate three steps to choose autoregressive integrated moving average (ARIMA) models and estimate from them: identification (model selection), estimation and diagonistic checking.

Model selection

First, it may be necessary to stabilise the variance or induce some distributional form, via a transformation. Secondly, we need to choose the degree of differencing discussed above, so as to induce stationarity. We achieve this by testing for unit roots. Thirdly, we then determine the orders of the AR and MA polynomials. Tools here include examining the time plot of the series, as well as the sample ACFs and PACFs, which we may compare with the ACFs and PACFs of known models. A time plot provides useful information on outliers, missing values and structural breaks in the data.

Fundamental is the idea of *parsimony*, that is, a relatively small model producing better forecasts. Forecasting often becomes worse if we have extra regressors (lags) which are not really needed, despite the fact that the fit may improve. Also, it may happen that the AR and MA polynomials share a *common factor*. We spot this if t-ratios are low whereas there is high correlation between parameter estimates. Available as model selection criteria are, for instance, the Akaike Information Criterion (AIC)

$$AIC = 2p - 2\ln L,$$

where p is the number of parameters in the model and L is the maximised value of the likelihood function for the model and Bayesian Information Criterion (BIC)

$$BIC = p{\cdot}\ln n - 2\ln L,$$

where n is the number of observations. For both criteria we select models with the smallest such values. It is important to realise that different researchers may come up with different models of choice.

Estimation

There are two common methods for estimating time series models: the **Yule-Walker algorithm** discussed earlier, which is particularly for AR models, and **maximum likelihood** (see Appendix B). The estimators of both have approximately the same sampling distribution. We saw the Yule-Walker equations in a simple case, but note that they may be generalised to consider AR(p) models. From now on, we will focus on the maximum likelihood method.

Maximum likelihood estimation: Consider a stationary and invertible process y_t (transformed and differenced appropriately), which we want to express as a ARMA(p, q) model. Suppose the p are in some p-variate vector $\mathbf{a}$ and q in some q-variate vector β. We generally use (approximate) maximum likelihood and assume that $\epsilon_t \sim N(0, \sigma^2)$ (i.i.d.). Often, it is easier to write the likelihood if we use a recursive formulation.

The general form is

$$\begin{aligned} L(\mathbf{a}, \beta, \sigma^2) &= pr(y_1, y_2, \ldots, y_n \,|\, y_0, y_{-1}, \ldots; \mathbf{a}, \beta, \sigma^2) \\ &= \prod_{t=i}^{n} pr(y_t \,|\, y_{t-1}, \ldots, ; y_0, y_{-1}, \ldots; \mathbf{a}, \beta, \sigma^2). \end{aligned}$$

There are different ways of expressing the likelihood; typically, we will obtain some non-linear function of the parameters which needs to be optimised. We may write down the likelihood directly in terms of the joint distribution of all observables (the method which we will eventually adopt later). Another option is to approximate the maximum likelihood estimate by minimising the sum of squares; this is known as conditional least squares, where we fix the first p values of y_t for AR(p), the first q values of ϵ_t for MA(q), and both for ARMA(p, q), that is, enough to make the series 'self-sufficient'.

Diagnostic checking

Once we have estimated, we must for instance check whether the residuals correspond to what we have assumed them to be, whether the variance is constant, and if any residual correlation is gone.

We may check the mean and variance using a residual plot (versus time), and check for autocorrelation of residuals by computing their sample ACF and PACF and using their asymptotic standard errors.

'Portmanteau' tests: These are used to test a group of correlations at once, but they do not test against a specific alternative. Generally, we need a lot of evidence to reject the null in large samples, but if we do reject then we do not know how to change the model. Box and Pierce (1970, [19]) defined the test statistic

$$Q^* = n \sum_{k=1}^{s} {r_k}^2,$$

which follows an asymptotic χ^2_{s-p-q} distribution if the data come from the ARMA(p, q) model.

Ljung and Box (1978, [81]) modified this to

$$Q = n(n+2)\sum_{k=1}^{s} \frac{{r_k}^2}{n-k},$$

which converges to a χ^2_{s-p-q} distribution, and if the value is larger than, say, a 95th percentile then we question the adequacy of the fitted ARMA(p, q).

Lagrange-multiplier tests: These are used to test for the null of a white-noise process for the residuals against the alternative of a AR(s) or MA(s) process. Here, we need not estimate the alternative model. A convenient way of computing it is generally as n times the R^2 of a regression of ϵ_t on the partial derivatives of ϵ_t with respect to the model parameters evaluated by maximum likelihood. Therefore, we test in terms of the actual variable for an ARMA($p+s$, q) or an ARMA(p, $q+s$) versus the null of an ARMA(p, q). The Lagrange multiplier will be asymptotically χ^2_s under the null.

F-tests: We might wish to estimate the model over *subsets* of the entire sample, for instance, split the sample in half and see whether both lead to the same model. If we wish to test whether the same ARMA(p, q) specification leads to the same coefficients in both subsamples then we use an F-test.

If SSR denotes the sum of squared residuals for the whole sample, SSR_1 for the first subsample and SSR_2 for the other then we have

$$F = \frac{(SSR - SSR_1 - SSR_2)/(p+q)}{(SSR_1 + SSR_2)(n - 2(p+q))},$$

which follows an $F_{p+q,\, n-2(p+q)}$ distribution under the null of parameter equality.

Alternatively, we may forecast part of the sample not used in the estimation; a good model should predict well.

3.2.9 Forecasting

An important aim of time series models is to forecast the future. We assume that we know the actual data-generating process, that is, we have the correct model and model parameters. Therefore, we know the parameters and the $\{y_t\}$, and then the $\{\epsilon_t\}$ are just functions of the two, and so are known.

Stationary models

Let us now consider that we have n observations and want to forecast the next ones. Using the AR(∞) representation

$$y_t = \sum_{i=1}^{\infty} \pi_i y_{t-i} + \epsilon_t,$$

as well as the fact that $\mathbb{E}(\epsilon_{n+1}) = 0$, at time point n we obtain

$$\mathbb{E}_n(y_{n+j}) = \sum_{i=1}^{j-1} \pi_i \mathbb{E}_n(y_{n+j-i}) + \sum_{i=j}^{\infty} \pi_i y_{n+j-i}, \tag{3.10}$$

where the subscript n on the expectation means conditional on all information up to and including n. This predictor is called a **forecast function**, and we can see that the first sum comprises the forecasted values whereas the second are observed. Thus we must compute recursively. n is called the **origin** here and j the **lead time**. Note that it is possible to rewrite this to suit MA or ARMA models.

We also need some idea of uncertainty. For this we use what is known as the **forecast error** (fe)

$$fe_n(j) = y_{n+j} - \mathbb{E}_n(y_{n+j}) = \sum_{i=0}^{j-1} \psi_i \epsilon_{n+j-i}, \tag{3.11}$$

which has mean zero but prediction error variance of

$$\sigma^2(1 + \psi_1^2 + \psi_2^2 + \ldots + \psi_{j-1}^2);$$

this enables us to make prediction intervals, typically using the formula: predictor ± 2 × standard error.

Unit-root models

When we need to difference a nonstationary process y_t into a stationary y_t^*, we may use the above to forecast the latter but need another expression to forecast the y_t. We have here that

$$\mathbb{E}_n(y_{n+j}) = y_n + \sum_{h=1}^{j} \mathbb{E}_n(y_{n+h}^*)$$

and

$$\sum_{h=1}^{j} fe_n(h) = \sum_{h=1}^{j} \sum_{i=0}^{h-1} \psi_i \epsilon_{n+h-i},$$

where $fe_n(h)$ is the forecast error of the stationary process y_t^* as in (3.11).

Holt-Winters algorithm

However, real data are not often generated by such simple models as ARMA and ARIMA, and more heuristic algorithms have been considered. One example of such is the Holt-Winters algorithm.

Recall the classical decomposition which we saw in (3.2), but forgetting the seasonal part S_t for now. Note that $\mathbb{E}(I_t) = 0$. We focus on a 'local' trend, where we place more weight on more recent observations:

$$\mathbb{E}_n(y_{n+j}) = \alpha_n + b_n j,$$

where α_n is the local level and b_n the slope of the trend. We update these through **exponential smoothing**, that is to say, for the local level we take a linear combination of the observed and forecasted values:

$$\alpha_t = \lambda_0 y_t + (1 - \lambda_0)(\alpha_{t-1} + b_{t-1})$$

and for the local slope we take a linear combination of the change in level and the previous slope:

$$b_t = \lambda_1(\alpha_t - \alpha_{t-1}) + (1 - \lambda_1)b_{t-1},$$

with starting values $\alpha_2 = y_2$ and $b_2 = y_2 - y_1$.

Smoothing constants λ_0 and λ_1 lie in $(0, 1]$, and setting them equal to 1 gives forecasts which use only the last two observations (extreme local behaviour):

$$\mathbb{E}_n y_{n+j} = y_n + (y_n - y_{n-1})j;$$

by contrast, if equal to 0 then the behaviour will tend to the global-trend model (assumed to hold for all t).

It is not difficult to generalise this algorithm to include the seasonal component, if necessary.

3.2.10 Volatility

Consider Figure 3.1, showing the annual unemployment rate in the US between 1890 and 1970 (source: Hyndman, 2005, [60]). Time series often displays behaviour in which some periods seem to have much larger variance than other periods in the sample. This does not necessarily mean that the process is not stationary (which implies a constant variance over time), but we do need to model this heteroskedasticity explicitly. Even though the marginal or unconditional variance is constant for a stationary process, there may still be conditional

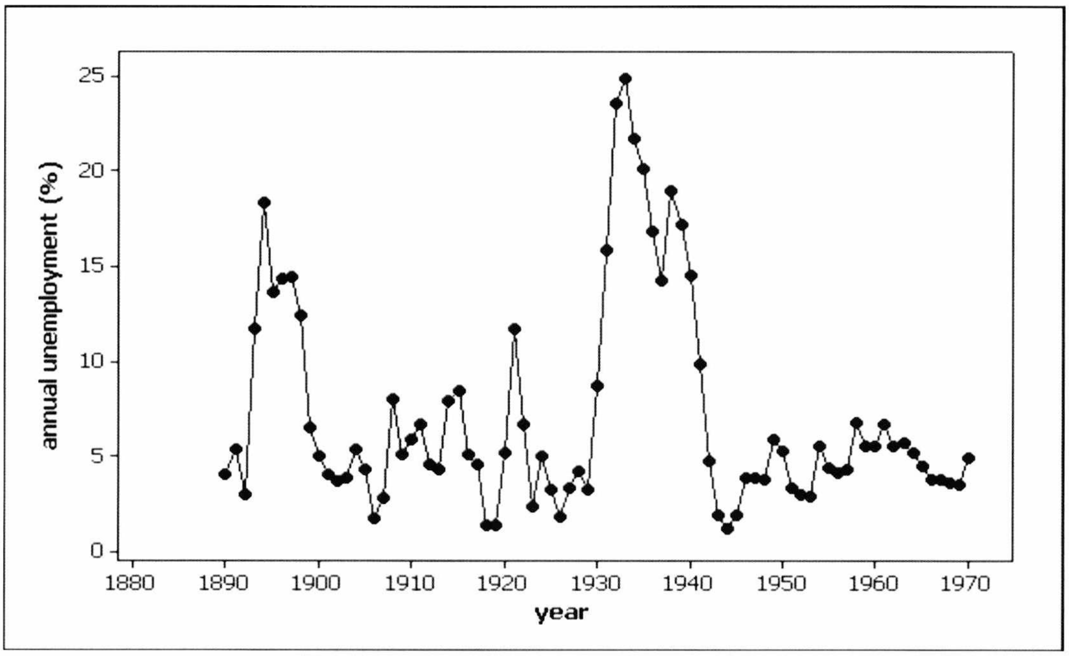


Figure 3.1: Annual unemployment in the US from 1890 to 1970.

heteroskedasticity. There is great interest in conditional, as opposed to unconditional, variance; for instance, an investor who wishes to sell shares in a week cares about the near future, not the long-run average variance.

The ARCH model

Here, we model the changing variance through an AR(q) process on the squares of past residuals, that is,

$$\mathbb{E}_t(\epsilon_{t+1}^2) = \alpha_0 + \alpha_1\epsilon_t^2 + \ldots + \alpha_q\epsilon_{t+1-q}^2,$$

where we must also introduce some disturbance (stochastics). We do so usually in a multiplicative way, for example,

$$\epsilon_t = \nu_t\sqrt{\alpha_0 + \alpha_1\epsilon_{t-1}^2},$$

where $\{\nu_t\}$ is a white-noise process with mean 0 and variance 1, independent of ϵ_{t-1}, and where $\alpha_0 > 0$ and $0 < \alpha_1 < 1$.

Some properties:

- $\mathbb{E}(\epsilon_t) = 0$;
- $\mathbb{V}ar(\epsilon_t) = \alpha_0/(1-\alpha_1)$ (constant unconditional variance);
- $\mathbb{E}(\epsilon_t | \, \epsilon_{t-1}, \, \epsilon_{t-2}, \, \ldots) = 0$; and

- $\mathbb{E}(\epsilon_t^2 \,|\, \epsilon_{t-1}, \epsilon_{t-2}, \dots) = \alpha_0 + \alpha_1 \epsilon_{t-1}^2$ (non-constant conditional variance).

From this we see that $\{\epsilon_t\}$ has conditional and unconditional mean zero, is serially uncorrelated but *not* independent, since they are linked through second moments. Clearly then, this means that $\{\epsilon_t\}$ can no longer be normal.

Generally, an (autoregressive conditional heteroskedasticity) ARCH(q) model involves a q^{th}-order AR process on the conditional variance.

The GARCH model

Suppose now that instead of modelling through an AR process, we progress to an ARMA process:

$$\epsilon_t = \nu_t \sqrt{h_t},$$

where now

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i},$$

and h_t is the conditional variance of ϵ_t. This is known as a (generalised autoregressive conditional heteroskedasticity) GARCH(p, q) model. Clearly, this is a generalised version of ARCH(1) already seen, in which $p = 0$ and $q = 1$.

This model avoids a very high q in ARCH, in turn allowing more parsimonious modelling for the volatility. Also, GARCH is stationary when

$$\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i < 1.$$

As before, the ACF (of squared residuals) would be useful in determining q and p. Once we obtain an ARMA model for $\{y_t\}$, we may calculate the squares of the estimated residuals $\hat{\epsilon}_t$, from which we may derive the large sample variance of the residuals:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_t^2$$

as well as their sample ACF:

$$\rho(i) = \frac{\sum_{t=i+1}^{n} (\hat{\epsilon}_t^2 - \hat{\sigma}^2)(\hat{\epsilon}_{t-i}^2 - \hat{\sigma}^2)}{\sum_{t=1}^{n} (\hat{\epsilon}_t^2 - \hat{\sigma}^2)^2}.$$

We may also use these in a Ljung-Box statistic for k lags, which again has a χ_k^2 distribution under the null of no ARCH or GARCH.

A more formal test is the Lagrange Multiplier test:

1. Estimate AR(k) or a regression model with ordinary least squares.

2. Calculate $\hat{\epsilon}_t^2$ and regress it on a constant and q-lagged squared residuals. The model is not ARCH only if the constant alone is non-zero, and nR^2 of that regression converges to a χ_q^2 under the null of non-ARCH.[2]

Estimation may be done using maximum likelihood.

Remark: Thus far, we have focused on covariance stationarity. This is a weaker form of stationarity, in that we only require the first two moments to be constant regardless of the time space. A stronger condition is strict stationarity.

3.3 Strictly Stationary Models

3.3.1 Stationary (ARCH) model

Definition: A time series $\{y_t\}$ is said to be **strictly stationary** if its spatial law is invariant under translation, that is, if random vectors $(y_{t_1}, y_{t_2}, \ldots, y_{t_n})'$ and $(y_{t_1+c}, y_{t_2+c}, \ldots, y_{t_n+c})'$ have the same joint distribution for all sets of indices $\{t_1, t_2, \ldots, t_n\}$ and for $c \in \mathbb{Z}$ and $n > 0$.

Introduction and motivation

To begin with, consider the Markov case. As we saw earlier: $AR(1)$ has the form: $y_t = a_1 y_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma^2)$ and where $\mid a_1 \mid < 1$ implies stationarity. We may generalise this to

$$y_t - \mu = a_1(y_{t-1} - \mu) + \epsilon_t. \tag{3.12}$$

However, if y_t is *not normally distributed* then it is not straightforward to ensure that the sequence $\{y_t\}$ is strictly stationary (that is, stationary for all moments).

Pitt, Chatfield and Walker (2002, [94]) look at the construction of first-order stationary autoregressive models using latent variables. These marginal distributions are not Gaussian restricted, but may instead come from the exponential family. The models presented are density based and easily adaptable.

The method involves not the conditional density function but rather its expectation. Pitt, Chatfield and Walker (2002) discuss specifying a joint density (y_t, y_{t-1}) (or equivalently in the stationary case specifying the marginal density of y_t and the conditional - or transition - density of $(y_t|y_{t-1})$), aiming for a linear

[2] These ARCH and GARCH models are not to be confused with the ARCH and GARCH models to which we refer hereafter.

relationship between y_t and y_{t-1} with respect to the mean, that is:

$$\mathbb{E}(y_t|y_{t-1}) = a_1 y_{t-1} + (1 - a_1)\mu, \tag{3.13}$$

where $\mu = \int y f_Y(y)\,\mathrm{d}y$ and $f_Y(y)$ is the required stationary density of y_t.

We saw earlier that the autocorrelation function (ACF) of $AR(1)$ is ρ^s, if (3.13) holds (see (3.7)).

We define $f(y_t|y_{t-1})$ such that:

$$f_{Y_t}(y_t) = \int f_{Y_t|Y_{t-1}}(y_t|y_{t-1})\,\mathrm{d}y_{t-1}; \text{ and} \tag{3.14}$$

$$\int y_t f_{Y_t|Y_{t-1}}(y_t|y_{t-1})\,\mathrm{d}y_t = \rho y_{t-1} + (1-\rho)\mu. \tag{3.15}$$

The key to the approach, from which we will eventually form the basis of our own method, is to have a latent variable, say x_t, such that the $\{x_t\}$ represent a vector of *probabilities*; by contrast, the $\{y_t\}$ represent a vector of *observations.* We then update the definition of $f_{Y_t|\,Y_{t-1}}(y_t|y_{t-1})$ to

$$f_{Y_t|\,Y_{t-1}}(y_t|y_{t-1}) = \int f_1(y_t|x_t) f_2(x_t|y_{t-1})\,\mathrm{d}x_t, \tag{3.16}$$

which is a formal version of what is shown in Figure 3.2.

In summary, *we* define $f_{Y_t|X_t}(y_t|x_t)$, with $f_1(y_t|x_t) = f_{Y_t|X_t}(y_t|x_t)$ and $f_2(x_t|y_{t-1}) = f_{X_t|Y_{t-1}}(x_t|y_{t-1})$, and where

$$\begin{aligned} f_{Y_t}(y_t) &= \int f_{Y_t,X_t}(y_t, x_t)\,\mathrm{d}x_t \\ &\equiv \int f_{X_t}(x_t) f_{Y_t|X_t}(y_t|x_t)\,\mathrm{d}x_t. \end{aligned}$$

Note that using the above, we can show simply that

$$\begin{aligned} \int f_{Y_t|Y_{t-1}}(y_t|y_{t-1}) f_{Y_{t-1}}(y_{t-1})\,\mathrm{d}y_{t-1} &= \int f_{Y_t,Y_{t-1}}(y_t, y_{t-1})\,\mathrm{d}y_{t-1} \\ &= f_{Y_t}(y_t), \end{aligned}$$

as required.

Stationarity

From looking at (3.13), we infer that the a_1 acts like a weight function between the observation directly before and the mean of y_t. Therefore, stationarity, i.e. $|\,a_1\,| < 1$, implies that $\mathbb{E}(y_t|\,y_{t-1})$ cannot steer too widely away over a noticeable time period, as desired.

To sample from the transition density, we can simulate from the latent process $\{x_t\}$, that is $y_t|x_t \sim f_{Y_t|X_t}(y_t|x_t)$ and $x_t|y_{t-1} \sim f_{X_t|Y_{t-1}}(x_t|y_{t-1})$. Use of these two densities has links with the Gibbs sampler process.

Also, Pitt, Chatfield and Walker (2002) state it useful both for $f_{Y|X}$ to have the same form as f_Y and that $E(x|y) \propto y$, in order to satisfy (3.15). Other than this, what we decide to be $f_{X|Y}$ does not matter.

ACF

The consequent autocorrelation function takes the form ρ^s, for some $s \in \mathbb{N}$ with $|\rho| < 1$ for stationarity. This very form implies exponential decay, which seems suitable since we may argue that people's votes are less and less likely to depend on the state of the party further back in time, and more likely to based on recent happenings. Again, of course, this is an oversimplification.

Application to our scenario

Recall that the $\{x_t\}$ are latent variables, which effectively work 'behind the scenes' in our modelling. They are the probability of voting for the relevant party, represented by y, at time t. Typically, we would therefore want x to follow a continuous distribution, such as the beta distribution.

The method starts at x_1 which we allocate a (continuous) distribution. The dependency sequence then begins as shown in Figure 3.2. The underlying idea is that the theoretical probability x_t is used to determine the actual outcome y_t for each t, assuming stationarity of voting and constant poll time intervals. We have a simple process which alternates between the continuous distribution and discrete distribution with time. The benefit of this approach is that we have quite a natural arrangement, which progresses with time. Clearly though, stationarity is a likely oversimplification, since for example an effective campaign with this party compared with those of other parties would affect the proportion of votes. This obviously can work adversely too. Also, in reality, polls are not taken over constant time intervals. Nevertheless, it is a useful starting point provided that we ignore external (and internal) influences.

Use of latent variables

Figure 3.2 shows the relationship between the observed data and the latent data in the stationary ARCH model for the simplest case. We see here that a latent

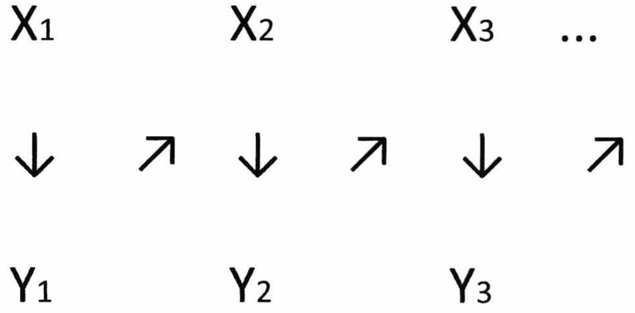


Figure 3.2: Relationship between observed and unobserved variables in the (simplest-case) stationary model.

variable is determined solely by the previous observation and that subsequently the forthcoming observation is influenced solely by it.

Example

A Beta-Binomial model: As we will eventually see, we will use this model for our election data. Let $y_t | \, x_t \sim Bin(r, \, x_t)$ and $x_t | \, y_{t-1} \sim Be(\alpha + y_{t-1}, \, \beta + r - y_{t-1})$, and also assume that r remains fixed.

Then,

$$\begin{aligned} \mathbb{E}(y_t \,|\, y_{t-1}) &= \mathbb{E}(\mathbb{E}(y_t \,|\, x_t) \,|\, y_{t-1}) \\ &= \mathbb{E}(r x_t \,|\, y_{t-1}) = r\mathbb{E}(x_t \,|\, y_{t-1}) \\ &= r\frac{\alpha + y_{t-1}}{\alpha+\beta+r} = r\frac{\alpha}{\alpha+\beta+r} + r\frac{y_{t-1}}{\alpha+\beta+r} \\ \Rightarrow a_1 &= \frac{r}{\alpha+\beta+r}, \end{aligned}$$

(where a_1 is as defined in (3.13)) which must imply stationarity, since α, β and $r > 0$.

Hence,

$$\begin{aligned} \mu(1 - a_1) &= \mu\frac{\alpha+\beta}{\alpha+\beta+r} = \frac{r\alpha}{\alpha+\beta+r} \\ \iff \mu &= \frac{r\alpha}{\alpha+\beta}; \end{aligned}$$

thus implying also that we expect future probabilities $\mathbb{E}(x_{t+f}) = \alpha/(\alpha+\beta)$, for $f \in \mathbb{N}$ some step in the future, i.e. the time series is stationary.

Remark: It is important to realise that, with such models, the dependence properties of the sequence are carried out only by the *latent* part of the model, $f_{X|Y}(x|\, y)$, thus enabling independence of the $f_Y(y)$. Such a property will aid our computation of the likelihood, as we will see later. For a fuller discussion of the theory and methodology, see Pitt, Chatfield and Walker (2002).

3.3.2 The stochastic volatility model

Introduction

In the easiest case, these models have the relationship:

$$y_t \sim f_{Y|X}(y_t|\, x_t) \quad \text{and} \quad x_t \sim f_{X|X}(x_t|\, x_{t-1}). \tag{3.17}$$

We may extend, of course, to consider any lag or a combination of lags, as well as multi parties. The idea here is that all the modelling of probabilities x_t is done in the background, in complete isolation from the actual data, y_t. Again then, we have x_t as unobserved data.

The conditional independence structure of the observations means that the marginal density of y is

$$f_Y(y) = \int f_{Y|X}(y|\, x) \cdot f_X(x)\, dx.$$

Pitt and Walker (2005, [95]) actually consider a slightly different version of (3.17), in which there is another step: the x_{t-1} does not *directly* generate the next x_t. Instead, there is an implementation of the method described in the previous section; that is to say, the x_{t-1} generates an **auxillary variable**, say z_t, which is then used to produce the next x_t. In our case, however, it will suffice to keep to our simplified version without the z_t. This is because the main point of the stochastic volatility models is to account for variation amongst time series data, which we will achieve by defining our $f_{X|X}(x_t|\, x_{t-1})$ such that the variance increases in direct proportion to the time difference between polls. See Example 1 in Section 3.4 for clarification.

Use of latent variables

Figure 3.3 shows the relationship between the observed data and the latent data in the stochastic volatility model for the simplest case. We see here that a latent variable is determined solely upon the previous latent variable and that subsequently the forthcoming observation is influenced solely by it. The observations play no role in influencing either the latent variables or other observations.

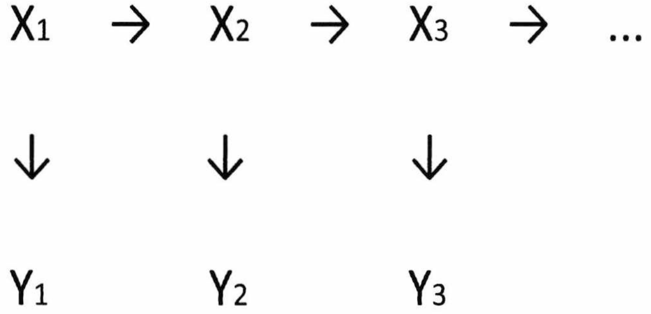


Figure 3.3: Relationship between observed and unobserved variables in the (simplest-case) stochastic volatility model.

3.3.3 The GARCH model

Introduction

The GARCH(1,1) model shown in 3.2.10 is one of the most widely-used models for modelling volatility. The observations can feed back directly into predicting the next observation, in contrast to the ARCH(1) model shown in 3.2.10, which is indirect. Pitt and Walker (2005) keep the marginal density of y_t and x_t both fixed and known. The method involves specifying a joint density

$$f_{Y,X,Z}(y,x,z) = f_{Y|X}(y|\,x) \cdot f_X(x) \cdot f_{Z|X}(z|\,x), \tag{3.18}$$

in which x is the latent mixing process, as before. Note also that $f_{Y|X,Z}(y|\,x,z) \equiv f_{Y|X}(y|\,x)$ so the observation is dependent only on x. The variable z is introduced to increase dependence, without affecting the marginal distribution of y. As before, the marginal density is simply

$$f_Y(y) = \int f_{Y|X}(y|\,x) \cdot f_X(x)\,\mathrm{d}x.$$

We want to create a Markov process $\{y_t, x_t\}$. We achieve this by first generating $x_1 \sim f_X(.)$ and use this to get $y_1 \sim f_{Y|X}(.|\,x_1)$ and $z_1 \sim f_{Z|X}(.|\,x_1)$. Next, for $t = 2, 3, \ldots, n$ we generate $x_t \sim f_{X|Y,Z}(.|\,y_{t-1}, z_{t-1}), y_t \sim f_{Y|X}(.|\,x_t)$ and $z_t \sim f_{Z|X}(.|\,x_t)$.

The resulting sequence $\{y_t, x_t, z_t\}$ for $t = 1, 2, \ldots, n$ is a Markov chain with invariant distribution (3.18). Note also that $f_{Y,X}(y, x)$ is the invariant distribution of $\{y_t, x_t, z_t\}$ for $t = 1, 2, \ldots, n$.

As with the stochastic volatility model, we will disregard the auxillary variable and focus simply on our latent x_t to carry out the necessary underlying modelling.

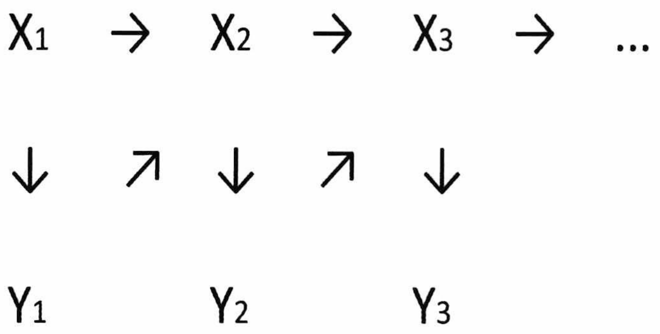


Figure 3.4: Relationship between observed and unobserved variables in the (simplest-case) GARCH model.

This means that we end up with the form $f_{X|X,Y}(x_t|\, x_{t-1}, y_{t-1})$. For illustrations of the more general type see Pitt and Walker (2005).

Use of latent variables

We define x_t as before. The sequence generates as shown in Figure[3] 3.4. We see here that a latent variable is determined both upon the previous latent variable and the previous observation, and that subsequently the forthcoming observation is influenced solely by it.

3.3.4 The higher-order stationary model

Introduction

Mixture Transition Distribution (MTD) models were introduced in Mena and Walker (2007, [83]). They are based on the (strictly-stationary) ARCH models already discussed in Section 3.3.1, but extend now to a non-Markov scenario. This lets us consider more lags. As before, having latent variables enables more flexibility in dependence structures.

Specification

Now let $Y^{[t,\,i]} = (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-i})$, the random lagged values down from time t to lag i, and similarly let $y^{[t,\,i]}$ denote observed points. Then, we define a MTD model $\{Y_t\}_{t\in\mathbb{N}}$ as strictly stationary with marginal density f_Y when $Y_1 \sim f_Y$ and

[3] These diagrams get complex very quickly as we increase the number of lags which we wish to model.

for all $t \geq 2$

$$f(y_t|y^{[t,t_p]}) = \sum_{k=1}^{t_p-1} w_k p(y_{t-k}, y_t) + (1 - \sum_{k=1}^{t_p-1} w_k) p(y_{t-t_p}, y_t), \tag{3.19}$$

where $t_p = (t-1) \wedge p$, $i \wedge p = \min\{i, p\}$, $\sum_{k=1}^{p-1} w_k \leq 1$ and the transition densities $p(y_k, .)$ take the form based on (3.16). For a proof of this, see Mena and Walker (2007, [83]).

We end up with

$$f(y_t|y^{[t,p]}) = \sum_{k=1}^{n} w_k p(y_{t-k}, y_t). \tag{3.20}$$

Estimation

We condition on a new latent variable, Z_t, taking values $1, \ldots, p$ with probabilities $w_1, w_2, \ldots, w_p$ respectively. Typically, it is assumed that

$$Pr(Z_t = z_t) = w_1^{z_{1t}} w_2^{z_{2t}} \ldots w_p^{z_{pt}}.$$

Obviously, we require that $\sum_{k=1}^{p} w_k = 1$. Hence, given $Z_t = (0, \ldots, 1, \ldots, 0)$, that is one in the k^{th} entry, (3.20) ultimately reduces to $p(y_{t-k}, y_t)$ for $k = 1, \ldots, p$.

In practice, it is convenient to work with a p-dimensional latent vector Z_{kt}, where Z_{kt} is binary (defined to be either one or zero according to whether the lagged value is k or not).

Given a sample $y = \{y_1, y_2, \ldots, y_n\}, n > p$, the logarithm of the augmented data likelihood is

$$l(\theta) = \ln L_{y,z}(\theta) = \sum_{k=1}^{p} \sum_{t=1}^{n} z_{kt}[\ln w_k + \ln p(y_{t-k}, y_t; \theta)],$$

where θ denotes all the parameters in the model. In theory, estimation of parameters including weights may be done using the Expectation-Maximisation (EM) algorithm. For an overview of this algorithm, see Appendix B and Dempster, Laird and Rubin (1977, [34]). Furthermore, Mena and Walker (2007) discuss the details of this in relation to the MTD model.

The problem is that such an integral as this may not even be available explicitly, or easy to compute. In such cases, it would be helpful, as before, to consider the latent variables $X = (X_1, X_2, \ldots, X_n)$, and so we have

$$\ln L_{x,y,z}(\theta) = \ln f_Y(y_1; \theta) + \sum_{k=1}^{p} \sum_{t=1}^{n} z_{kt}\{\ln w_k + \ln[f_{Y|X}(y_t|x_t; \theta)\, f_{X|Y}(x_t|y_{t-k}; \theta)]\}. \tag{3.21}$$

Example

A Dirichlet-multinomial model: Recall Example 2 showing the beta-binomial stationary ARCH model. Ultimately in this thesis, we will work with more than two parties in the modelling of our election data. We now show how Example 2 generalises, illustrating with the three-party and two-lag case, with the MTD model.

Firstly, with two lags we now have $AR(2)$: $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \epsilon_t$, where $|\, a_1 + a_2 \,| < 1$ must hold in order to imply stationarity.

Thus we must update equation (3.13) so that now

$$\mathbb{E}(y_t | y_{t-1},\, y_{t-2}) = a_1 y_{t-1} + a_2 y_{t-2} + (1 - a_1 - a_2)\mu,$$

where $\mu = \int y f_Y(y)\, dy$ and $f_Y(y)$ is the required stationary density of Y_t.

Now we have that $\mathbf{y}_t | \mathbf{x}_t \sim MN(r,\, \mathbf{x}_t)$ and $\mathbf{x}_t | \mathbf{y}_{t-j} \sim Dir(\alpha_1 + y_{1(t-j)},\, \alpha_2 + y_{2(t-j)},\, \alpha_3 + y_{3(t-j)})$, where $MN(r,\, \mathbf{x}_t)$ denotes the multinomial distribution with trial size r and probability vector of success $\mathbf{x}_t$ and where $Dir(\alpha_1 + y_{1(t-j)},\, \alpha_2 + y_{2(t-j)},\, \alpha_3 + y_{3(t-j)})$ denotes the Dirichlet distribution with shape parameters $\alpha_1 + y_{1(t-j)},\, \alpha_2 + y_{2(t-j)}$ and $\alpha_3 + y_{3(t-j)}$, for $j = 1, 2$ (the number of lags). We assume again that r remains fixed, and have vectors

$$\mathbf{y}_t = (y_{1(t)}\; y_{2(t)}\; y_{3(t)})'$$

with $y_{3(t)} = r - y_{1(t)} - y_{2(t)}$, $t = 1,\, 2,\, \ldots,\, n$ for n observations, and also

$$\mathbf{x}_t = (x_{1(t)}\; x_{2(t)}\; x_{3(t)})'$$

with $x_{3(t)} = 1 - x_{1(t)} - x_{2(t)}$.

Then, for $k = 1,\, 2,\, 3$,

$$\begin{aligned}
\mathbb{E}(y_{k(t)} \,|\, y_{k(t-1)},\, y_{k(t-2)}) &= \mathbb{E}(\mathbb{E}(y_{k(t)} \,|\, x_{k(t)}) \,|\, y_{k(t-1)},\, y_{k(t-2)}) \\
&= \mathbb{E}(r x_{k(t)} \,|\, y_{k(t-1)},\, y_{k(t-2)}) = r\mathbb{E}(x_{k(t)} \,|\, y_{k(t-1)},\, y_{k(t-2)}) \\
&= r \left\{ w \frac{\alpha_k + y_{k(t-1)}}{\sum_{m=1}^{3} \alpha_m + r} + (1 - w) \frac{\alpha_k + y_{k(t-2)}}{\sum_{m=1}^{3} \alpha_m + r} \right\} \\
&= r \left\{ \frac{\alpha_k}{\sum_{m=1}^{3} \alpha_m + r} + \frac{w y_{k(t-1)}}{\sum_{m=1}^{3} \alpha_m + r} + \frac{(1 - w) y_{k(t-2)}}{\sum_{m=1}^{3} \alpha_m + r} \right\} \\
\Rightarrow a_1 &= \frac{rw}{\sum_{m=1}^{3} \alpha_m + r} \quad \text{and} \quad a_2 = \frac{r(1 - w)}{\sum_{m=1}^{3} \alpha_m + r},
\end{aligned}$$

which must imply stationarity, since $\alpha_1, \alpha_2, \alpha_3$ and $r > 0$.

Hence,

$$\mu(1 - a_1 - a_2) = \mu \frac{\sum_{m=1}^{3} \alpha_m}{\sum_{m=1}^{3} \alpha_m + r} = \frac{r\alpha_k}{\sum_{m=1}^{3} \alpha_m + r}$$
$$\iff \mu = \frac{r\alpha_k}{\sum_{m=1}^{3} \alpha_m};$$

thus implying also that we expect future probabilities

$$\mathbb{E}(x_{k(t+f)}) = \frac{\alpha_k}{\sum_{m=1}^{3} \alpha_m},$$

for $f \in \mathbb{N}$ some step in the future, i.e. the time series is stationary.

Remark: Here we have seen how to extend the stationary model so as to consider multi-lags. We need not make such extensions for the stochastic volatility and GARCH models as doing so with them is more direct and straightforward, at least with the particular forms which we choose, as we will see in Chapter 6.

3.4 Adaptions to Strictly Stationary Models

With the strictly stationary models outlined, the conditional variance does not depend upon the time interval between observations. In order to account for *time variation*, it is possible to adapt the models such that the variance increases with the time interval.

Consider the conditional beta density

$$x_t | \, \xi_{t1}, \, \xi_{t-2} \sim Be(c_t\xi_{t1}, \, c_t\xi_{t2}),$$

in which $0 \leq \xi_{t1}, \, \xi_{t2} \leq 1$ and $\xi_{t1} + \xi_{t2} = 1$.

Our aim is to have c_t such that as the time interval tends to zero the conditional variance tends to zero, and as the time interval increases the conditional variance increases. One way would be to write

$$c_t = \frac{e^{-\phi\delta_t}}{1 - e^{-\phi\delta_t}},$$

in which δ_t denotes the time interval between y_t and y_{t-1} and ϕ is a positive parameter (constant) to be estimated. Here, as δ_t tends to zero c_t tends to infinity and as δ_t tends to infinity c_t tends to zero.

We now illustrate how we might apply this to the stochastic volatility and GARCH models discussed in the previous section. These examples will be of particular importance later.

Example 1: a stochastic volatility model

Let

$$y_t|\, x_t \sim Bin(m, x_t) \quad \text{and} \quad x_t|\, x_{t-1} \sim Be(c_t x_{t-1}, c_t(1 - x_{t-1}))$$

for $t = 2, \ldots, n$, where c_t is as defined above, m is the total number of trials (fixed for all t) and where x_1 is the starting point which we choose. Here, at each step t, it is the previous probability which accordingly splits the current c_t amongst the two shape parameters, thereby effectively determining the probability density function at that stage. The y_t are simply independent outputs. We end up with

$$\mathbb{V}ar(x_t|\, x_{t-1}) = \frac{x_{t-1}(1 - x_{t-1})}{c_t + 1}. \tag{3.22}$$

We revisit this in Chapter 6, 6.4.7.

Example 2: a GARCH model

Similarly to before, let $y_t|\, x_t \sim Bin(m, x_t)$ but now with

$$x_t|\, x_{t-1}, y_{t-1} \sim$$
$$Be\left(c_t \left(ux_{t-1} + (1-u)\frac{y_{t-1}}{m}\right), c_t \left(u(1 - x_{t-1}) + (1-u)\left(1 - \frac{y_{t-1}}{m}\right)\right)\right)$$

for $t = 2, \ldots, n$, where c_t is as defined above, m is the total number of trials (fixed for all t) and where x_1 is the starting point which we choose. This time we also have a weight $0 \leq u \leq 1$, and so end up including both the previous probability and corresponding actual outcome; with the latter we have converted it into a proportion via employing m, in order to make it comparative to the probability. Here, at each step t, it is both the previous probability *and* previous actual outcome which accordingly distribute the current c_t amongst the two shape parameters, thereby effectively determining the probability density function at that stage. To account for variation between events y_t we have

$$\mathbb{V}ar(x_t|\, x_{t-1}, y_{t-1}) = \frac{(ux_{t-1} + (1-u)\frac{y_{t-1}}{r_{t-1}})(u(1 - x_{t-1}) + (1-u)(1 - \frac{y_{t-1}}{r_{t-1}}))}{c_t + 1},$$

which we will analyse further in Chapter 6, 6.4.7.

3.5 Review

In summary, we have reviewed the theory and progress of time series in a general sense. Our account deliberately put focus on modelling and less emphasis on technical details and estimation. This is mainly because it is the *modelling* which we will be doing in the ensuing chapters. We firstly outlined the motivations for using time series analysis, and showed the general construction of time series models. We spent much thought on the covariance stationary aspect, discussing its tools which include the ACF and PACF, and also its models (AR, MA and ARMA). Next, we outlined methods of tackling nonstationarity of the mean, covering the separate techniques for predictable and stochastic trends; an important concept met here was differencing, thus leading us to the ARIMA model.

When choosing such models, we use intuition and graphs of the data, as well as tools such as the AIC and BIC. The next stage is the estimation of parameters of our chosen model given the data, for which the maximum likelihood method will play an important role, starting with the expression of the likelihood. Then, we need to carry out inference to check our assumptions made before fitting the model; tools include residual analysis and the F-test.

Having done this, of great interest is forecasting. We have outlined forecasting functions for the AR, MA and ARMA models, and shown how it is possible to forecast less simple models using the Holt-Winters algorithm. We also looked at volatility in the data, introducing the ARCH and GARCH models, which are designed to deal with this.

The latter section of the chapter switched to cover strict stationarity, which focused on newer models and all of which make use of latent variables. Firstly, we discussed in detail the stationary model, which uses certain time series theory akin to the former section. We extended this model so as to account for time variation between observations, for which we have the stochastic volatility model as well as the GARCH model, again analysing in detail their key properties. Also, we showed how to extend the basic stationary model so as to deal with more lags than just the previous one. Overall, the basic idea is that we have the stationary model, but tamper with the conditional variance of the probability component. This is to make the model more realistic and to account for the difference in time between polls, by conditioning on the time intervals; by contrast, the stationary model essentially assumes a unit time interval between each observation. It is these three types of model on which we now focus attention for the remainder of this thesis.

As we have said, time series has a wide range of applications in the real world.

It has been used for modelling election data, an example of which is in Harvey and Shephard (1990, [55]). However, in subsequent chapters we will introduce and analyse a new approach to modelling election data, in order first to interpret what the poll data before a general election say and then to predict using our findings and interpretations. We will see from plots in Chapter 7 how the development of poll voting per party with time offers itself naturally to time series analysis.

In the general breakdown shown in equation (3.2), we will focus on T_t and I_t from now; S_t may be dropped as the polls will not cover a long enough period of time to encounter seasonal variation, or at least we will assume this to be the case. We will employ the AR(p) model, starting with the AR(1) case, as part of one of our main general models. (Indeed, we could also look at the effect of introducing lags in the *errors* (MA and ARMA models); we make reference to this in Chapter 9 as possible further work.) Then, we will borrow ideas from the ARCH and GARCH theory already outlined, for the other nominated models. This means that we may skip the Box Jenkins model selection stage, and thus much of the diagnostic checking outlined. In fact, we will carry out specific model selection *as part of* our diagnostic checking, that is to say, we first fit the data to all of our candidate models and *then* assess which is the best.

Interestingly, Whiteley (1979, [108]) considered monthly opinion polls from 1947 to 1975, focusing on a two-party contest of Conservative versus Labour, with the main aim of *forecasting poll movements*. He fit AR(1), MA(1) and ARMA(1,1) models to each dataset (Conservative popularity and Labour popularity) and performed a complete Box-Jenkins analysis. The conclusion was that the ARMA(1,1) models fitted most adequately, thus implying also that the errors were white noise. The parameter estimates indicated that each model was stationary and invertible, and behaved similarly to each other. Forecasting of the next fifty months in the way which we outlined in Subection 3.2.9 showed relatively good outputs compared to what happened. In particular, forecasting of the polls around the time of each election was extracted to study this more closely. It might be important to consider seasonality here due to the wide time period covered, as opposed to our environment which is much narrower. This would be true not only in the modelling of historical polls but also in the subsequent forecasting. However, from results of an earlier spectral analysis (frequency-domain time series as opposed to time-domain) Whiteley (1979) argued that the latter part was only over a fifty-month period and not a significant enough length in which any seasonal cyclic behaviour was noticeable, at least given that specific data.

Chapter 4

Forecasting Methods

4.1 Introduction

Time series refers to data observed at different points in time. Time series forecasting then, is a forecasting method which uses a set of historical data to predict the likely outcome of a future event. Usually, these data are spaced equally over time (discrete time intervals, $t = 1, 2, \ldots$) and may represent anything from monthly profit figures to weekly electricity consumption. The underlying assumption is a *combination of a pattern and some random error.* Subsequently, the aim is to separate the pattern from the error by understanding the pattern's trend, its longer-term increase or decrease, as well as its seasonality, that is, the change caused by seasonal influences. In Chapter 3, we showed how all these may be expressed in a single general formula (see equation (3.2)). We also saw in Chapter 3 one way to perform time series forecasting, but note that the method outlined there catered specially for the relatively straightforward *autoregressive* model, and in a different (covariance stationary) context, which we will not end up using for our forecasting.

The forecasting of elections is extremely complicated. A large range of factors influences voting behaviour, and these factors differ from consituency to constituency. The method presented in Brown and Payne (1975, [23]) has since developed into a sophisticated approach to forecasting during election night. Practice has shown that the general method can provide accurate results before all seats have been declared. For this reason, our emphasis in this thesis is on pre-election night data obtained from opinion polls, and using our methodology in essentially a modification of the method in the Brown and Payne (1975) paper.

The following section discusses the most basic statistic, the poll of polls, used widely in the media to represent the electorate opinion about an upcoming general election. We outline a more elaborate method of modelling the opinion polls, designed to improve on the poll of polls. This approach lies in the framework of time series and simultaneously provides a forecast of the final outcome. Also, we provide a brief overview of other methods of forecasting which have been employed for election outcomes, such as the cube rule.

4.2 Poll of Polls

Objective: In order to summarise the information contained in all opinion polls conducted pre-election, the most known statistic currently available is the poll of polls.

Data: As the name suggests, the only data used are the recent opinion poll results conducted by various organisations.

The model: This simply comprises the **arithmetic means** of votes per party by *combining* all opinion polls.

Inference: Essentially, these averages imply the prediction of outcome, or at least how the public are likely to vote.

Discussion: The statistic fails to incorporate time; we have already mentioned that public opinion - especially more recently - is volatile. Ideally, we would like to treat the data as time series, in order to identify and track changes with time.

4.3 Cube Rule Model

4.3.1 Objective

This was first seriously used as an election forecasting method in 1950. The basic principle involves turning the total numbers of declared votes, at a particular point in time, per party into a forecast of the *final* outcome. It is designed for a two-party contest scenario.

4.3.2 Data

The primary source of information used here is voting data so far declared on election night per party.

4.3.3 The model

The method works as follows:

1. As each constituency declares the number of votes for each of Labour and Conservative, we are able to refine the final forecast, taking in more information, yet in a cumulative manner. Let L_j denote the number of votes for Labour in constituency j, C_j the number for Conservative and $T_j = L_j + C_j$, and suppose that there are 607 seats in the UK. Then we construct the cumulative number of votes per party known at some time, that is, $\sum_{j=1}^m L_j$, $\sum_{j=1}^m C_j$ and $\sum_{j=1}^m T_j$, in which m seats have declared.

2. First let $p = (\sum_{j=1}^m L_j)/(\sum_{j=1}^m T_j)$ and $q = 1 - p$. If we focus on forecasting the number of seats eventually won by Labour, La, with To the corresponding total number of seats, then the cube rule states that the ratio of Labour seats won to Conservative seats (Co) should be *proportional* to p^3/q^3, which may be expressed as follows:

$$\frac{La}{Co} > \frac{p^3}{q^3}.$$

Using this it is possible to derive

$$La > \frac{(\frac{p}{q})^3 To}{1 + (\frac{p}{q})^3}, \tag{4.1}$$

which boils down in our illustration to

$$La > \frac{607}{1 + \left(\frac{\sum_{j=1}^m C_j}{\sum_{j=1}^m L_j}\right)^3} = \frac{607p^3}{p^3 + q^3}, \tag{4.2}$$

and obviously $To = 607 = La + Co$.

4.3.4 Inference

As may be seen, this is a remarkably simple model; we see from (4.2) that the number of seats predicted to be won by a party is just (greater than) the total number of UK seats multiplied by a proportion. In the latter, all terms have the same power, and it is merely the relative proportion of votes which the party has won compared to what every party has won. If the power were one instead of three then clearly the forecast is just the total number of seats multiplied by the proportion of votes which the party has won (since $p + q = 1$), which makes sense if we were say interested in getting an intuitive, albeit a somewhat crude, forecast.

4.3.5 Discussion

The cube rule proved to be very accurate for elections from 1935 to 1970. A necessary assumption is that the variance of seats is approximately equal to the variance of votes in total, both of which are assumed to be Gaussian. However, it does carry some important setbacks. Firstly, we have no prior forecasts and have to wait for the first seat to be declared before we can provide any kind of forecast. The forecast is always only that of the final outcome and offers no forecast per undeclared seat. Also, the rule is designed for the now outdated UK scenario of a two-party contest. Although it would be straightforward to extend such a formula as (4.1) to cover more than two parties, there is no empirical evidence on any power being good enough for prediction, in the way that the cube is for the two-party case. This predicts less well the less that Labour and Conservative collectively account for in votes compared to the total number of votes.

4.4 Pollyvote Model

4.4.1 Objective

Cuzán, Armstrong and Jones (2004, [33]) apply what is known as the *combination principle* to election night forecasting. The model aims to give a combined forecast of the incumbent's share of the two-party vote.

4.4.2 Data

The method was used to forecast the 2004 US Presidential Election and came within 0.2 per cent of the actual result. The four forecasting methods combined were polls, prediction markets, regression models and expert surveys.

4.4.3 The model

- Predictions within the first three methods were combined, averaging recent polls, averaging the mean daily quotes for the previous week, and averaging results of the models.

- Then, the forecast vote was averaged across all four methods - the combined (averaged) forecasts of the polls, quotes and models plus the predictions of the experts panel - assigning *equal weights* to each:

$$\text{Pollyvote} = \frac{\text{mean}(polls) + \text{mean}(quotes) + \text{mean}(models) + experts}{4}$$

- This so-called 'Pollyvote' was updated at first once a week and then, as more polls were published and the election closer approaching, twice a week.

4.4.4 Inference

The argument made is that different methods are likely to have different biases meaning that their forecast errors would probably be *uncorrelated* and perhaps *offsetting*. Also, more information is used to steer the final forecasts. However, five methods should be the maximum number combined; after this point, every additional method accuracy normally improves, but at a lower rate. Further, previous work suggests that such combinations never harm forecasting accuracy and substantially reduces the risk of large forecast errors. As well as five methods maximum, different methods and/or datasets should be used, forecasts should be combined according to some predetermined procedure, and equal weights should be applied, unless there is strong prior evidence of varying accuracy of the components. Combining is ideal where forecast errors from the different methods have negative or zero correlation; otherwise, combining is still useful the further away from +1.0 the correlation coefficients are.

4.4.5 Discussion

This principle has been shown to give a lower forecast error than those obtained by each component method. This model is clearly simple, although just as we mentioned with the poll of polls averages only provide point estimates; they cannot identify outliers or the impact of shocks should they occur. A time series approach would be able to assess both. Also, again it would not be possible to perform an analysis per seat in the way that the BBC method (see below) can.

4.5 Whiteley's Forecasting Model

4.5.1 Objective

Whiteley (2005, [109]) introduced a statistical model which combined information from both shares of votes in the previous election and the most recent poll proportions. The motivation was to focus on forecasting the number of seats rather than the number of votes, since previous work has tended to try to predict the number of votes. Where there do not exist breakdowns by seat, modelling of the latter has been of restricted use, for it is the number of seats not the number of votes which determines the ultimate winner. This model also falls into the subject of time series; however, it covers an extremely wide time period of years rather than days and hours.

4.5.2 Data

Sources of information required are recent opinion poll results per party and previous election results in terms of number of seats won per party, that is, for all historical general elections.

4.5.3 The model

- Unfortunately, the estimation procedure is not spelt out fully, although the basic model is
$$S_{i(t)} = \alpha S_{i(t-1)}^{\beta_0} \prod_{j=1}^{k} P_{j(t)}^{\beta_j} \epsilon_i,$$
in which $S_{i(t)}$ is the seat share of party i at election t, $P_{j(t)}$ is party j's vote share in the poll taken before election t, α, β_0 and β_j are parameters to be estimated and error $\epsilon_i \sim N(0, \sigma^2)$.

- Obviously, we would want the $P_{j(t)}$ to be as close to the corresponding election as possible; Whiteley uses poll data conducted six months before each election.

- Whitely then states the following regression model:
$$\text{ln}S_{i(t)} = \text{ln}\alpha + \beta_0 \text{ln}S_{i(t-1)} + \sum_{j=1}^{k} \beta_j \text{ln}P_{j(t)} + \text{ln}\epsilon_i.$$

4.5.4 Inference

One may use maximum likelihood to obtain estimates, $\hat{\theta}$, for the θ, given the fifteen observations (previous election results and corresponding poll results). Then, as required, we would arrive at our prediction
$$\hat{S}_{i(16)} = \hat{\alpha}\, S_{i(15)}^{\hat{\beta}_0} \prod_{j=1}^{k} P_{j(16)}^{\hat{\beta}_j}.$$
Whiteley plots the actual values with the fitted values, in order to examine the accuracy of the model described. With regression models a standard inferential tools is the adjusted R^2 statistic. In addition, Whiteley performs χ^2 tests for serial correlation, normality and heteroskedasticity (variability). With regards to the forecast, Whiteley briefly discusses a predictive failure F-test.

4.5.5 Discussion

Due to poll data before 1945 being unreliable, this means that observations are seriously limited in quantity. At the time of writing, there were only fifteen observations. This would imply then the need for a parsimonious model, say. Using opinion polls instead, we will typically have more data available than this with which to work to obtain parameters.

Additionally, Whiteley includes dummy variables for the 1983 and 1987 elections to capture the effect of the divide in the Labour party in 1981 and the creation of the Social Democratic party. This had the potential to disrupt the usual relationships between seats and votes until the system balanced out overall with the creation of the Liberal Democrats in 1988.

This is a Markov model, whereby only the last election directly influences the next, although due to its basic construction may easily be extended. The model fit slightly well in forecasting the 2001 election, with absolute errors of 18, 24 and 13 seats for Labour, Conservative and Liberal Democrats, respectively. However, its level of simplicity means that we only get a total forecast as opposed to a breakdown by constituency. Ultimately, of course, this total is what we wish to forecast, but analysts from several areas are interested in breakdowns. Such breakdowns will be possible by our method. Also, this model makes no attempt to consider the probabilities of voting at any time, which we argue is crucial since it governs (and explains) the number of votes and effectively the number of seats won.

4.6 Electoral Calculus Model

4.6.1 Objective

Electoral Calculus ([38]) is a website which uses scientific analysis of opinion polls and electoral geography to predict what would happen if a general election were to happen tomorrow. Therefore, the focus is on a forecast given the current scenario. Prediction by seat is based on the latest opinion polls and so is continually updated.

4.6.2 Data

Sources of information for the prediction are the latest opinion polls available in addition to the previous election results per party per constituency.

4.6.3 The model

There are basically two models, the simplest of which is called the 'additive uniform national swing model' and the slightly more enhanced of which is called the 'transition model'.

The additive uniform national swing model

Here we try to estimate

$$A(k, j) = (P(k) - E(k)) + C(k, j),$$

in which: there are k parties and j constituencies; $A(k, j)$ is the *share* of support for party k in consituency j; $P(k)$ is the national share of support for party k currently (obtained solely using opinion polls); $E(k)$ is the national share of support for party k in the last election; and $C(k, j)$ is the share of support for party k in constituency j at the last election.

The transition model

This has two parts for the predicted support levels:

1. $A1(k, j) = C(k, j) \cdot \frac{P(k)}{E(k)}$ if party k's support declines; and
2. $A2(k, j) = C(k, j) + S(k) \cdot V(j)$ if party k's support increases,

in which

$$S(k) = \frac{\max(P(k) - E(k), 0)}{\sum_{i=1}^{n} \max(P(i) - E(i), 0)}$$

where n is the total number of constituencies and in which

$$V(j) = \sum_{i=1}^{n} C(k, i) \cdot \max\left(1 - \frac{P(i)}{E(i)}, 0\right).$$

Therefore in $A2(k, j)$, $S(k) \cdot V(j)$ is the product of how much party k has gained nationally relative to all gainers and the fraction of voters in seat j who have swung.

4.6.4 Inference

We can see that the simpler model uses the opinion polls in equal weight to the last election result. These are both national figures, which are assumed equal in importance to the individual figures. To get figures for current party strengths, the poll data results are employed as the substitute for not yet having the final election results. In the transition model parties who decline are assumed to do so multiplicatively, that is if national support drops by 10 per cent, then constituency support also drops by 10 per cent. Parties who increase in support gain voters from the declining parties, in proportion to the amount of that party's increase relative to all increasing parties. Obviously then, votes may only increase in any seat to the extent that declining parties have lost votes.

4.6.5 Discussion

Overall, the models have proven to be quite good when fit to past elections, despite some noted minor weaknesses in each method. We also get detailed breakdowns of the forecast by seat, as well as the overall result. However, we have no information on the theoretical side, the probabilities of voting, which we regard as useful and interesting in the modelling of election outcomes. [Note that the website has since improved on this model by also splitting the voters into strong and weak categories, which has meant slight adjustments to the methodology.]

4.7 Harvey and Shephard Model

4.7.1 Objective

Harvey and Shephard (1990, [55]) focus on the modelling of opinion polls in order to make a forecast of the final election, based primarily on poll data. The underlying motivation was to step up from the simple poll of polls *point* estimate. Their method is a multivariate structural time series approach.

4.7.2 Data

The only source of information used is the recent opinion poll results per party. In the paper, Harvey and Shephard model the opinion polls for each general election between October 1974 and 1987.

4.7.3 The model

The model is a continuous-time local level model. In its simplest (univariate) case,

$$y_t = \mu_t + \epsilon_t,$$

$$\mu_t = \mu_{t-1} + \eta_t,$$

in which $\{y_t\}$ is a sequence of the number of votes for one of any two parties in poll t, where $t = 1, 2, \ldots, n$ and n polls have been conducted. The **random walk** component μ_t is the **underlying level** of the process and errors $\epsilon_t \sim N(0, {\sigma_1}^2)$ and $\eta_t \sim N(0, {\sigma_2}^2)$ imply a two-staged **drift**.

Since this level varies with time, the optimal forecasts are constructed by discounting past observations. The method is an extension of exponentially-weighted moving averages, and the model uses the Kalman filter (see below) to handle the fact that observations arrive at irregularly-spaced intervals. It is relatively simple to extend this to the multivariate case.

Multivariate version

- Let $\mu(t)$ be an $n \times 1$ vector which evolves according to a multivariate Wiener (Brownian motion) process, such that $\mathrm{d}\mu(t)$ has uncorrelated Gaussian increments with zero mean and covariance matrix

$$\mathbb{E}\left[\left(\int_r^s \mathrm{d}\mu(t)\right)\left(\int_r^s \mathrm{d}\mu(t)\right)'\right] = (s-r)\mathbf{Q},$$

  for $s > r$ and where $\mathbf{Q}$ is a positive definite matrix. The elements of $\mu(t)$ represent the underlying levels of each of the n processes.

- Observations are made at discrete intervals, at times t_τ, $\tau = 1, 2, \ldots, t$. The intervals between the observations are

$$\delta_\tau = t_\tau - t_{\tau-1},$$

  for $\tau = 2, 3, \ldots, t$, so δ_τ will be zero if two observations occur at the same time.

- The discrete-time process corresponding to the points at which observations are made is the multivariate random walk

$$\mu_\tau = \mu_{\tau-1} + \eta_\tau,$$

where η_τ has a multivariate normal distribution with zero mean and covariance matrix

$$\mathbb{V}ar(\eta_\tau) = \delta_\tau \mathbf{Q}.$$

- Not all the elements of $\mu(t)$ need to be observed at a particular time, t_τ. Furthermore, those which are observed are assumed to be subjected to measurement error. If y_τ denotes the $n_\tau \times 1$ vector of observable random variables at time t_τ, then

  $$y_\tau = \mathbf{Z}_\tau \mu_\tau + \epsilon_\tau,$$

  for $\tau = 2, 3, \ldots, t$, where $\mathbf{Z}_\tau$ is an $n_\tau \times n$ selection matrix of zeros and ones, which picks out the appropriate elements of μ_τ. The $n_\tau \times 1$ vector of disturbances is assumed to have a multivariate normal distribution with zero mean and covariance matrix

  $$\mathbb{V}ar(\epsilon_\tau) = \mathbf{\Sigma}_\tau.$$

- For given values of the positive definite matrices $\mathbf{Q}$ and $\mathbf{\Sigma}_\tau$, the Kalman filter yields the minimum mean square error estimators (MMSEs) of the state μ_τ based on the information at time t_τ, while MMSEs of μ_τ based on all the information in the sample can be obtained by smoothing. Predictions of future observations, together with their mean square errors (MSEs) can also be made by the Kalman filter. Finally, maximum likelihood estimators of the unknown parameters may be computed via the prediction error decomposition. Note that if $n_1 = n$ then the Kalman filter may be initiated with an estimated state vector of y_1 and a MSE matrix $\mathbf{\Sigma}_1$.

- Harvey and Shephard impose a **design effect**, ϕ, (assumed constant) to act as an 'inflating' factor for $\mathbb{V}ar(\epsilon_\tau)$, so that $\mathbb{V}ar(\epsilon_\tau)$ is updated now to

  $$\mathbb{V}ar(\epsilon_\tau) = \phi \mathbf{\Sigma}_\tau.$$

  Fuller detail of implementing sampling theory in the model is given in Harvey and Shephard (1990).

4.7.4 Inference

- Maximum likelihood is used to estimate the unknown parameters. Given that $\mathbf{\Sigma}_\tau$ is taken to be known, the model has, as unknown parameters, ϕ and the $n(n+1)/2$ distinct elements of $\mathbf{Q}$. The constraint that $\mathbf{Q}$ must be positive definite is imposed by working with the lower triangular matrix $\mathbf{A}$ such that $\mathbf{A}'\mathbf{A} = \mathbf{Q}$.

- $n_\tau = n$ in general, so that $\mathbf{Z}_\tau$ is the identity matrix.

- Fitted values, $\hat{\mu}_{jt|t}$, for the model are found by the Kalman filter recursions below. The starting values are $\hat{\mu}_{1|1} = y_1$ and $\mathbf{P}_{1|1} = \hat{\phi}\hat{\mathbf{\Sigma}}_1$, where μ_1 is replaced by y_1. For $\tau = 2, 3, \ldots, t$ we have

  1. $\hat{\mu}_{\tau|\tau-1} = \hat{\mu}_{\tau-1|\tau-1}$
  2. $\mathbf{P}_{\tau|\tau-1} = \mathbf{P}_{\tau-1|\tau-1} + \delta_\tau \hat{\mathbf{Q}}$
  3. $\hat{\mu}_{\tau|\tau} = \hat{\mu}_{\tau|\tau-1} + \mathbf{P}_{\tau|\tau-1}\mathbf{Z}_\tau'\mathbf{F}_\tau^{-1}(y_\tau - \mathbf{Z}_\tau\hat{\mu}_{\tau|\tau-1})$
  4. $\mathbf{P}_{\tau|\tau} = \mathbf{P}_{\tau|\tau-1} - \mathbf{P}_{\tau|\tau-1}\mathbf{Z}_\tau'\mathbf{F}_\tau^{-1}\mathbf{Z}_\tau\mathbf{P}_{\tau|\tau-1}$,

  where $\mathbf{F}_\tau = \mathbf{Z}_\tau\mathbf{P}_{\tau|\tau-1}\mathbf{Z}_\tau' + \hat{\phi}\hat{\mathbf{\Sigma}}_\tau$.

- Diagnostic checks are undertaken on the one-step ahead prediction errors of the model. These are transformed using a Cholesky decomposition of the prediction error covariance matrix, so that if the model were true and parameters estimated exactly then they would independent, each with a standard normal distribution. Checks made include a nonparametric test for serial correlation, an F-test for heteroskedacity in the residuals, and skewness and kurtosis tests for normality.

- Mean-squared errors are calculated. Also, the sum of squares of the forecast errors, SSE, is used to measure the overall accuracy of the forecasts; the values are compared with comparative figures from the opinion polls of major polling companies.

4.7.5 Discussion

Findings and tests show that the model predicts well, and that the gain from it, rather than the poll of polls, is likely to be higher during the early and middle stages of the campaign, when the data are less frequent.

The overall idea makes sense, since the effects are incremental over the short term, yet more noticeable over the longer term. We would expect this kind of behaviour with poll data, as we would probably not expect high fluctuations. In the event that we do, then it is possible to control the shocks via the variances of both ϵ_t and η_t. Similarly, we will have models which also consider time differences.

The model presented is Markovian, such that the authors put the case forward of having a model which places more weight on recent observations and less so on previous ones. We will explore this case by looking at multi-lags, and demonstrate how in theory our models can cater easily to involve them. Note that it should be possible to extend the method of Harvey and Shephard to the non-Markov case.

The model is based primarily on poll data, but makes use neither of election night data nor exit poll data in arriving at its forecast, but instead relies almost solely on the results of polls in the run up. This does enable a forecast to be made well before the night itself and thus assuming that the results are accurate, or modelling acceptable enough, would be of great interest to psephologists and politicians. However, since we will also model opinion polls initially we will be able to provide such an 'early' estimate. We will therefore make such a comparison in Chapter 7 with real pre-election data. Our overall approach then goes on firstly to use exit poll data and secondly to channel through election night, letting our model evolve as the seats are declared.

This multivariate structural time series model assumes the normality of poll voting and the authors perform diagnostic tests to accept such a null, but this does not mean that this is the best or most suitable model. We will depart from the common Gaussian assumption and look at models which appear to suit the *types* of data, that is to say, the probabilities of voting as well as the numbers of votes in polls.

Assumptions are made both that the error (ϵ_t) is purely due to sampling, and independent through time. Smith (1978, [104]) makes an interesting point that the second assumption would in practice be false, as most opinion polls are based on a master sample of constituencies, which then act as a panel of primary sampling units. Harvey and Shephard argue that nevertheless the assumption still enables a good first approximation, and that the effect would not be too influential.

There is no slope component in the model, and so it consistently underestimates (overestimates) a series containing a strong upwards (downwards) trend. The magnitude of this error is dependent upon the strength of the trend compared to the estimated sampling error. If more opinion polls are conducted then the lack of slope becomes less important. In contrast, we consider models which could track such trends.

4.8 Review

This chapter has reviewed forecasting techniques which have been proposed for election outcomes. We began by looking at the simple poll of polls estimate, which provides little scope for deeper analysis, contrast to what psephologists are ultimately interested in performing. In response, in 1990 Harvey and Shephard presented a more sophisticated model in the field of time series. Here, the focus is on modelling the opinion poll data in order to declare from them a forecast. These two approaches are prior forecasts. In the remainder of the chapter, we broadly outlined some other approaches which have been applied, the most known of which is the cube rule.

We will define and illustrate our own method in Chapter 8. We look also at the BBC's regression-based method, as well as suggested enhancements by Brown and Payne (1975 and 1984) and Brown, Firth and Payne (1999), in which the focus now is on updating forecasts on election night, as and when seats are declared. Some weight is also placed on more historical information, in particular exit poll data. First though we must concentrate on how to estimate the parameters of our types of model; as we have already stressed, the majority of our work focuses on this as opposed to developing a completely new forecasting method.

Chapter 5

Hierarchical Likelihood

5.1 Introduction

In this chapter we discuss parameter estimation. In Chapter 3 we specified models whose likelihoods are based on latent variables. This effectively means that we cannot use the standard estimation methods such as direct maximum likelihood. The EM algorithm (see Appendix B) is designed to deal with finding maximum likelihood estimates (MLEs) in the presence of missing data. However, with the type of models which we are using, the integrals which we obtain from the expectation stage are difficult if not impossible to solve, apart from a special case, which we will see later. An alternative approach would be maximisation using Markov Chain Monte Carlo (MCMC). The methodology is detailed in Geyer and Thompson (1992, [47]), and involves simulating ergodic Markov chains having equilibrium distributions in the model. From one realisation of such a Markov chain, a Monte Carlo approximant to the likelihood function is obtained, and the parameter value (if it exists) maximising this function approximates the MLE.

Lee and Nelder (1996, [69]) introduced an alternative method of estimation, called the hierarchical likelihood, or **h-likelihood**. Chapter 5 is devoted to the theory and application of the h-likelihood, as developed in Lee and Nelder (1996, [69]), and discussed in Lee and Nelder (2005, [68]), Lee, Nelder and Noh (2007, [71]) and in detail in Lee, Nelder and Pawitan (2006, [72]).

First, though, it would be useful to review some basic theory on generalised linear models (GLMs). This leads nicely into *hierarchical* GLMs (HGLMs), in which there is dependence among two types of variables, one of which is unobserved. There is a wealth of material on HGLMs in Lee and Nelder (1996 and 2005), including inferential techniques and illustrations. Then, we consider the h-likelihood itself, which is used to estimate parameters of HGLMs, studying the properties of these estimators. Briefly, the method has similar features to that of both maximum likelihood and the EM algorithm; see Appendix B for outlines of each.

Under appropriate conditions, maximising the h-likelihood gives us fixed effect estimators which are *asymptotically* equivalent to those obtained from the marginal likelihood; at the same time, we get random effect estimates which are asymptotically the best unbiased predictors.

5.2 Generalised Linear Models

GLMs were introduced in Nelder and Wedderburn (1972, [88]) to unify various statistical models under one framework. Part of the motivation was to enable a general method for estimation. It is a generalisation of ordinary least-squares regression, in which the data y are assumed to be normally distributed, thus $y \in (-\infty, \infty)$ and so $\mathbb{E}(y)$ can be any real number. However, if the data were instead to come from a Poisson distribution, say, then $y \in \mathbb{N}_0$ and we must cater for the restriction that $\mathbb{E}(y) > 0$.

Let y denote the outcomes, or response, such that y follows a distribution within the *exponential family*. Distributions from this family, including the normal, binomial and Poisson, have a probability density (or mass) function of the form:

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{a(y)b(\theta) - c(\theta)}{m(\phi)} + d(y, \phi)\right\}, \tag{5.1}$$

where

- θ is called the **canonical parameter**;
- ϕ is called the **dispersion parameter**, which is usually known and related to $\mathbb{V}ar(y)$; and
- a, b, c, d and m are known.

We have that

$$\begin{aligned}\mathbb{E}(y) &= c'(\theta), \\ \mathbb{V}ar(y) &= \phi c''(\theta).\end{aligned}$$

We will define $\mathbf{A}\beta$ as the **linear predictor**, that is, a linear combination of the unknown parameters β, in which $\mathbf{A}$ are the explanatory variables $(A_1, A_2, \ldots)$. Also, let

$$\mu = \mathbb{E}(y). \tag{5.2}$$

The key to the theory is that $\mathbf{A}\beta$ and (5.2) are connected by the **link function**, g, such that

$$\eta = g(\mu) = \mathbf{A}\beta.$$

It is then a matter of estimating the parameters β, using a tool such as maximum likelihood. Chapter 2 of Lee and Nelder (2006) includes the necessary assumptions and further properties of GLMs.

5.2.1 Example

We verify that the binomial distribution, $y \sim Bin(r, p)$, is a member of the exponential family. We have that

$$\begin{aligned} Pr(y = y) &= \binom{r}{y} p^y (1-p)^{r-y} \\ &= \exp\left\{ \ln\binom{r}{y} + y \cdot \ln\left(\frac{p}{1-p}\right) + r \cdot \ln(1-p) \right\} \\ &= \exp\left\{ \frac{y \cdot \ln\left(\frac{p}{1-p}\right) - (-r \cdot \ln(1-p))}{1} + \ln\binom{r}{y} \right\}, \end{aligned}$$

where $a(y) = y$, $b(\theta) = \ln\left(\frac{p}{1-p}\right)$, $c(\theta) = -r{\cdot}\ln(1-p)$, $d(y,\, \phi) = \ln\binom{r}{y}$ and $m(\phi) = 1$. Also, we have that $p = \exp(\theta)/(1+\exp(\theta))$ so that $c(\theta) = r{\cdot}\ln(1+\exp(\theta))$.

We may rewrite $c(\theta)$ as

$$c(\theta) = r{\cdot}\ln\left(\frac{1}{1 - \frac{e^\theta}{1+e^\theta}}\right) = r{\cdot}\ln(1+e^\theta).$$

Then,

$$c'(\theta) = \frac{r{\cdot}e^\theta}{1+e^\theta} = r{\cdot}p,$$

which we know to be the mean of a binomial distribution. We may follow a very similar procedure with $c''(\theta)$ in order to obtain $\mathbb{V}ar(y) = r{\cdot}p{\cdot}(1-p)$.

Here the link function is

$$g(p) = \ln\left(\frac{p}{1-p}\right).$$

This distribution will become important to us later.

5.3 Hierarchical Generalised Linear Models

Definition: If we keep y as the response but now introduce x as an *unobserved* random component then we update (5.1) so that we get a log-likelihood of the form

$$l(\theta, \phi;\, y|\, x) = \frac{a(y)b(\theta) - c(\theta)}{m(\phi)} + d(y, \phi), \tag{5.3}$$

where

- we write μ' for the conditional mean of y given x such that

  $$\eta' = g(\mu') = \eta + v \tag{5.4}$$

  is the link function for the GLM, which describes the conditional distribution of $y|\, x$;

- $\eta = \mathbf{A}\beta$ as for a GLM; and where

- $v = v(x)$ for some strictly monotonic function of x.

Notice how in η' we have a fixed effects part (in η) and a dispersion part (in v) to describe the overdispersion, each of which requires modelling. Note that Lee and Nelder (1996) mention an estimation method using score equations. These equations require v rather than x, as v can assume any real number whereas x usually has range restrictions, which may cause problems in convergence. For the method which we will eventually adopt we only need to work with x alone.

Here, x *has its own assigned distribution.* Lee and Nelder (1996) state that the distribution of x (or equivalently v) is best decided by the properties of the data or the purposes of inference, and to try simple forms of random effect estimators. It is often assumed that x is normally distributed, which is unlikely to be always the case. Furthermore, Lee and Nelder (1996) concentrate on distributions which are conjugate to that defined for the response, y, although this is not essential. Then, we call the resultant model a **conjugate HGLM**; Lee and Nelder (1996) give a formal definition of this, which we summarise in the next section.

5.4 H-Likelihood

Definition: We define the h-likelihood, h, as

$$h = l(\theta,\, \phi;\, y|\, v) + l(\alpha;\, v), \tag{5.5}$$

- where $l(\alpha;\, v)$ is the logarithm of the density function for unobserved v with parameter vector α; and
- $l(\theta,\, \phi;\, y|\, v)$ is that for observed y given v.

Remember from earlier that the random component v is a scale on which the random effect x occurs linearly in the linear predictor, but we may derive the h-likelihood from the density functions of x and $y|\, x$. We will later work with x directly rather than a transformation of it. Lee, Nelder and Pawitan (2006, [72]) detail how to go about choosing v, thereby implying that v is not necessarily unique.

Therefore, h is the logarithm of the joint density function for v and y. Clearly h is not an orthodox likelihood in the sense that the x are *latent* and not observed. The h-likelihood is a generalisation of Henderson's joint likelihood, developed for normal models with random effects (Henderson, 1975, [57]). In deriving via h the 'equivalent' to maximum likelihood estimates (MLEs), we obtain estimators which we will call **maximum h-likelihood estimates** (MHLEs) by the process described below.

Process

The basic idea is as follows. We first choose an initial value for $\beta^{(0)} = \tilde{\beta}^{(0)}$, where $\tilde{\beta}$ denotes the MHLE of β.

Then, for $j = 0,\, 1,\, 2,\, \ldots,\, N$, such that at the N^{th}-iteration the system converges:

1. Fix $\beta^{(j)}$ at $\tilde{\beta}^{(j)}$. Then maximise h by solving

$$\frac{\partial h}{\partial v} = 0 = \tilde{v}^{(j+1)}. \tag{5.6}$$

2. Fix $v^{(j+1)}$ at $\tilde{v}^{(j+1)}$. Then maximise h by solving

$$\frac{\partial h}{\partial \beta} = 0 = \tilde{\beta}^{(j+1)}. \tag{5.7}$$

Features

- As with MLEs, the MHLEs are invariant with respect to the transformation v; in other words, if instead of (5.6) we used $\partial h/\partial x = 0$ then we would end up with the same random effect estimate. See Lee and Nelder (1996) for more.

- If we integrate out these random effects from h then we get the marginal distribution of the observed response. However, this integration will often be difficult if not impossible, and will be hard to work with subsequently. The key advantage of the h-likelihood is that we avoid the integration which is necessary when the marginal likelihood is used; furthermore, h is easily available. Random effects may also be of interest if inference is focused on individual responses; the double-step structure h enables us to analyse the former as well as the latter.

- For estimation of dispersion parameters an **adjusted profile h-likelihood** provides the required generalisation of restricted maximum likelihood. For inferences, various test statistics are developed in Lee and Nelder (1996), which include the scaled deviance test for the goodness of fit and a model selection criterion for choosing between various dispersion models.

5.4.1 Conjugate HGLMs

Let the response be y_{ij}, for $i = 1, 2, \ldots, t$ (where t is the number of groups) and $j = 1, 2, \ldots, n_i$, and furthermore let $n = \sum_{i=1}^{t} n_i$.

Now consider the canonical link model

$$\theta'_{ij} = \theta_{ij} + v_i,$$

in which $\theta'_{ij} = \theta(\mu'_{ij})$, $\theta_{ij} = \theta(\mu_{ij})$ and $v_i = \theta(u_i)$.

Then, using the definition of h given earlier, we have that

$$\frac{\partial h}{\partial \beta_k} = \sum_{ij} \frac{(y_{ij} - \mu'_{ij})x_{Aij}}{\phi}. \tag{5.8}$$

Assume that the kernel of $l(\alpha; v)$ has the form

$$\sum_i (a_1(\alpha)v_i - a_2(\alpha)b(v_i)),$$

where a_1 and a_2 are functions of the dispersion parameters, α.

The prior is specified for the random component v only; we need not specify priors for β, ϕ or α.

The kernel of h then is

$$\sum_{ij} \frac{\theta' y - b(\theta')}{\phi} + \sum_i (a_1(\alpha)v - a_2(\alpha)b(v)).$$

Since

$$\frac{\partial b(\theta(\mu))}{\partial \theta} = \mu,$$

and

$$\frac{\partial b(v)}{\partial v} = u,$$

we have that

$$\frac{\partial h}{\partial v_i} = \frac{\sum_j (y_{ij} - \mu'_{ij}) + \phi a_1(\alpha)}{\phi} - a_2(\alpha)u_i.$$

Setting this equal to 0 and then solving for u gives

$$\hat{u}_i = \frac{\sum_j y_{ij} - \sum_j \mu'_{ij} + \phi a_1(\alpha)}{\phi a_2(\alpha)}. \tag{5.9}$$

If

$$\mathbb{E}(u) = \frac{a_1(\alpha)}{a_2(\alpha)}$$

and the fixed effects have an intercept term then from (5.8) and (5.9),

$$\frac{\sum_{i=1}^t \hat{u}_i}{t} = \frac{a_1(\alpha)}{a_2(\alpha)},$$

analogously to the result for residuals in normal linear models.

5.4.2 Example: the beta-binomial model

Let

$$y|\, u \sim Bin(m,\, \pi')$$

so that

$$\mu' = m\pi'.$$

We have the following conjugate HGLM ($\theta'_{ij} = \theta_{ij} + v_i$):

$$\theta' = \ln\left(\frac{\pi'}{1-\pi'}\right),$$

$$v_i = \ln\left(\frac{u_i}{1-u_i}\right)$$

and

$$\theta = \ln\left(\frac{\pi}{1-\pi}\right) = \mathbf{A}\beta.$$

We also let

$$u \sim Be(\alpha_1,\, \alpha_2).$$

Then, we have that

$$l(\alpha;\, v) = \sum \left(\alpha_1 v_i - (\alpha_1 + \alpha_2)\ln\left(\frac{1}{1-u_i}\right) - \ln B(\alpha_1,\, \alpha_2) \right),$$

in which $a_1(\alpha) = \alpha_1$ and $a_2(\alpha) = \alpha_1 + \alpha_2$.

By equation (5.8) the MHL equations for β are

$$\frac{\partial h}{\partial \beta_k} = \sum_{ij} (y_{ij} - m_{ij}\pi'_{ij})A_{kij} = 0. \tag{5.10}$$

By equation (5.9) the MHL equations for u are

$$\hat{u}_i = \frac{\sum_j y_{ij} - \sum_j \mu'_{ij} + \alpha_1}{\alpha_1 + \alpha_2}. \tag{5.11}$$

When $\alpha_1/\alpha_2 \rightarrow 1$ and $\alpha_1 \rightarrow \infty$, $\hat{u}_i \rightarrow 1/2$, i.e. $\pi' \rightarrow \pi$. When $\alpha_1, \alpha_2 \rightarrow 0$, $\sum_j y_{ij} = \sum_j \mu'_{ij} = \sum_j m_{ij}\pi'_{ij}$ for all i.

Since

$$\pi' = \frac{\pi u}{\pi u + (1-\pi)(1-u)},$$

$\mathbb{E}(y) = \mathbb{E}(m\pi') \neq \mu = m\pi$, so inference on the marginal mean may not be easy. In general, an explicit form of marginal likelihood for beta-binomial models is not available.

However, with no fixed effects and only random effects,

$$\theta'_{ij} = \ln\left(\frac{u_i}{1-u_i}\right),$$

and the resulting marginal distribution is beta-binomial.

Then, (5.11) becomes

$$\hat{u}_i = \frac{\sum_j y_{ij} + \alpha_1}{\alpha_1 + \alpha_2 + \sum_j m_{ij}} = \mathbb{E}(u_i|\, y).$$

Lee and Nelder (1996) have more examples to illustrate the idea, particularly conjugate HGLMS such as the Poisson-gamma and gamma-inverse gamma models.

5.5 Properties of Estimators

It is important that the MHLEs are useful to us, that is, similar in behaviour to MLEs, which have attractive properties. Lee and Nelder (1996) make the necessary proofs, the most essential of which are shown below.

5.5.1 Asymptotically best unbiased prediction

Firstly, we require that our MHLE for the unobserved component is reliable. The EM algorithm is a way of finding the MLE in the presence of missing data, such as unobserved components. In the EM algorithm, the unobserved component would be dealt with by taking its expectation. Therefore, it would be helpful if the MHLE for the random component were close in value to what we expect it to be.

Let $v(x)$ denote some transformation of unobserved variable x. Then $\mathbb{E}(v(x)|\, y)$ is the best unbiased predictor for $v(x)$. It is shown in Lee and Nelder (1996), and covered below, that as long as v is strictly monotonic then asymptotically

$$\tilde{v} \to \mathbb{E}(v|\, y),$$

where $\tilde{v}$ is the MHLE of v and further that

$$v|\, y \sim N(\tilde{v},\, D^{-1}), \tag{5.12}$$

i.e. $\tilde{v} \approx \mathbb{E}(v|\, y)$, in which D is a diagonal matrix such that the i^{th} element

$$D_i = -\frac{\partial^2 h}{\partial {v_i}^2}|_{v=\tilde{v}}.$$

However, it is essential also that for all i

$$D_i^{-1} = O_P\left(\frac{1}{n}\right), \tag{5.13}$$

where n is the sample size and $\sum_{i=1}^{t} n_i = n$ as we defined in 5.4.1.

It is found that (5.13) holds if the number of groups t remains the same but the within group sample sizes $n_i \to \infty$ at the same rate.

Proof

Lee and Nelder (1996) consider the relationship

$$\tilde{h}_i \propto \tilde{L}_i + \tilde{l}(\beta, \phi, \alpha;\, v_i|\, y),$$

where

- h_i is the h-likelihood of v_i and y_{ij};
- L_i is the marginal likelihood of y_{ij}; and
- $l(\beta, \phi, \alpha;\, v_i|\, y)$ is the conditional likelihood of v_i given y.

Using the power series expansion,

$$\exp(h_i) = \exp(\tilde{h}_i\{1 + c_3(v_i - \tilde{v}_i)^3 + c_4(v_i - \tilde{v}_i)^4 + \ldots\}),$$

where c_3 and c_4 are coefficients with order $O_P(n)$, Liu and Pierce (1993, [79]) showed that

$$\exp(L_i) = \exp(\tilde{L}_i\{1 + O_p(n^{-1})\}).$$

Therefore, we may show that

$$\exp(l_i(\beta, \phi, \alpha;\, v_i|\, y)) = \exp(\tilde{l}_i(\beta, \phi, \alpha;\, v_i|\, y)\times \\ \{1 + c_3(v_i - \tilde{v}_i)^3 + c_4(v_i - \tilde{v}_i)^4 + \ldots + O_P(n^{-1})\}),$$

where $\tilde{l}_i(\beta, \phi, \alpha;\, v_i|\, y)$ is the log-likelihood of the normal density (5.12).

Therefore,

$$\mathbb{E}(v_i|\, y) = \tilde{v}_i + (c_3 - \tilde{v}_i c_4)\mathbb{E}^*(v_i - \tilde{v}_i)^4 + O_P(n^{-1}) = \tilde{v}_i + O_P(n^{-1}),$$

where $\mathbb{E}^*$ is an expectation with respect to the normal density (5.12), which proves that

$$\tilde{v}_i = \mathbb{E}(v_i|\, y) + O_P(n^{-1}),$$

as required. □

Illustration

Consider the following beta-binomial model:

$$y_i|\, x_i \sim Bin(r_i,\, x_i)$$
$$x_i \sim Be(\alpha,\, \beta),$$

where $j = 1$, which means both that we may drop this subscript and that now $n_i = n$. Hence $i = 1, 2, \ldots, n$, so that there are n observations. Further details of the binomial and beta models are provided in Chapter 6, 6.2.

Since the mode of the beta distribution is $(\alpha-1)/(\alpha+\beta-2)$, when we apply it to our model above the explicit form for estimating the MHLE of x_i, $\tilde{x}_i$, is

$$\tilde{x}_i = \frac{\alpha-1+y_i}{\alpha+\beta-2+r_i}.$$

Also, the mean of the beta distribution is $\alpha/(\alpha+\beta)$.

According to Lee and Nelder (1996), we need to check that ${D_i}^{-1} = O_P(1/n)$.

The h-likelihood, h, is

$$h \propto \sum_{i=1}^{n} (\alpha-1+y_i)\text{ln}x_i + (\beta-1+r_i-y_i)\text{ln}(1-x_i).$$

Then,

$$\frac{\partial h}{\partial x_i} = \frac{\alpha-1+y_i}{x_i} - \frac{\beta-1+r_i-y_i}{1-x_i}$$
$$\Longrightarrow -\frac{\partial^2 h}{\partial {x_i}^2}|_{\tilde{x}_i} = \frac{\alpha-1+y_i}{\tilde{x}_i^2} + \frac{\beta-1+r_i-y_i}{(1-\tilde{x}_i)^2} = D_i.$$

Therefore for ${D_i}^{-1} = O_P(1/n)$ to hold it is necessary for D_i to increase with n.

Now if we let

$$c_0 = \frac{1}{\tilde{x}_i^2}(\alpha-1) + \frac{1}{(1-\tilde{x}_i)^2}(\beta-1),$$

which is a constant value, then we may rewrite D_i as

$$\begin{aligned} D_i =& c_0 + \frac{y_i}{\tilde{x}_i^2} + \frac{r_i-y_i}{(1-\tilde{x}_i)^2} \\ =& c_0 + \frac{y_i(1-2\tilde{x}_i)+r_i\tilde{x}_i}{\tilde{x}_i^2(1-\tilde{x}_i)^2}. \end{aligned}$$

Thus, as long as we have that $r_i = O_P(n)$ the result holds, since $\tilde{x}_i$ is a constant and y_i will increase with r_i.

5.5.2 Asymptotic efficiency for $\tilde{\beta}$

MLEs are appealing to us for many reasons, which are summarised in Appendix B; therefore, the closer the MHLEs are to the MLEs the better, and the more useful they will be for inference purposes. We now prove that the MHLE $\tilde{\beta}$ is asymptotically equivalent to the MLE $\hat{\beta}$.

Proof

Consider the Taylor series expansion

$$\frac{\partial h}{\partial \beta_k} = \frac{\partial h}{\partial \beta_k}|_{v=\tilde{v}} + A_1(v-\tilde{v}) + (v-\tilde{v})'\frac{A_2}{2!}(v-\tilde{v}) + \dots,$$

where

$$A_1 = \left(\frac{\partial}{\partial \beta_k}\right)\left(\frac{\partial h}{\partial v}\right)|_{v=\tilde{v}} \text{ and } A_2 = \left(\frac{\partial}{\partial \beta_k}\right)\left(\frac{\partial^2 h}{\partial v^2}\right)|_{v=\tilde{v}}.$$

Since

$$\tilde{v} = \mathbb{E}(v|\,y) + O_P\left(\frac{1}{n}\right)$$

and, similarly,

$$\mathbb{V}ar(v|\,y) = O_P\left(\frac{1}{n}\right)$$

then we have that

$$\mathbb{E}\left(\frac{\partial h}{\partial \beta_k}|\,y\right) = \frac{\partial h}{\partial \beta_k}|_{v=\tilde{v}} + O_P(1),$$

provided that $A_1/n = O_P(1) = A_2/n$.

Now, the marginal ML equation for β becomes

$$\mathbb{E}\left(\frac{\partial h}{\partial \beta_k}|\,y, \hat{\beta}^{(p)}\right) = \frac{\partial h}{\partial \beta_k}|_{v=\hat{v}} + O_P(1).$$

Since

$$\frac{\partial^2 h}{\partial {\beta_k}^2} = O_P(n),$$

we can show that the difference between the marginal ML and MHL estimator for β is of the order $O_P(n^{-1})$, using the inversion of series; see Barndorff-Nielsen and Cox (1989, [4]). □

Illustration

Again, we consider the beta-binomial model

$$y_i \sim Bin(r_i,\, x_i)$$
$$x_i \sim Be(\alpha,\, \beta),$$

for $i = 1, 2, \dots, n$, so that there are n observations. Now, we compare the performance of the h-likelihood in terms of getting parameter estimates with that of the EM algorithm, the latter of which provides MLEs. Given the previous two proofs, and particularly the second, we would expect the MHLEs to converge to the MLEs as n increases.

We now consider various simulations of increasing size (fifty different simulation datasets for each case). The process involves first simulating data given the chosen values for α and β and then maximising the likelihood of the model given the simulated data, first using the EM algorithm and then using the h-likelihood procedure. Just for simplicity, we will keep r_i fixed for all i. We will in fact set $r_i = n$ due to what we found in 5.5.1.

Full details of the h-likelihood and EM algorithm methods for the beta-binomial model are given in Chapter 6, when applied in a more complicated time-series scenario (see 6.5). Briefly, for the h-likelihood we first maximise $\{x_i\}$ explicitly given fixed values for α and β, and then we switch to using the Newton-Raphson method to update and maximise both α and β given the most recent values for $\{x_i\}$. We simply repeat this two-stage procedure until the convergence of α, β and the latent $\{x_i\}$.

n	mean MLE	mean MHLE	absolute error
50	$\hat{\alpha}$=2.2345 $\hat{\beta}$=5.8825	$\tilde{\alpha}$=2.3976 $\tilde{\beta}$=6.7025	+0.1631 +0.8200
100	$\hat{\alpha}$=2.4016 $\hat{\beta}$=6.5217	$\tilde{\alpha}$=2.5225 $\tilde{\beta}$=7.0888	+0.1509 +0.5671
500	$\hat{\alpha}$=1.9777 $\hat{\beta}$=4.7723	$\tilde{\alpha}$=1.9782 $\tilde{\beta}$=4.8016	+0.0005 +0.0293
1000	$\hat{\alpha}$=2.0221 $\hat{\beta}$=5.0287	$\tilde{\alpha}$=2.0236 $\tilde{\beta}$=5.0345	+0.0015 +0.0058
5000	$\hat{\alpha}$=1.9284 $\hat{\beta}$=4.8613	$\tilde{\alpha}$=1.9286 $\tilde{\beta}$=4.8629	+0.0002 +0.0016
10000	$\hat{\alpha}$=2.0007 $\hat{\beta}$=5.0680	$\tilde{\alpha}$=2.0008 $\tilde{\beta}$=5.0683	+0.0001 +0.0003

Table 5.1: EM algorithm versus h-likelihood for simulations of probabilities from $Be(2, 5)$.

Remarks: From looking at both tables we can see that as n increases the absolute errors generally get smaller, as required. By the time we get to $n = 500$

the (mean) MHLEs are reasonably accurate, and when $n = 10000$ they are very close to the MLEs (to two decimal places).

There are some other interesting observations. The MHLEs, in all instance but one above, give larger values than the MLEs. Also, importantly, the MHLEs (and MLEs) are close in value to the starting values allocated, to indicate the reliability of the underlying h-likelihood method.

Lee and Nelder (1996) have other proofs which collectively justify the use of MHLEs in the absence of MLEs. These include a proof for using particular estimates of the covariance matrix of MHLEs, and furthermore the justification for using maximum *adjusted profile* h-likelihood estimators. In our context, we are more concerned with the former of these and not the latter, since we will not have many different *kinds* of parameter (but usually three: one set for determining party strength, one for identifying stronger time lags and one related to the time interval between polls). The estimates mentioned are useful when we make inferences about the realised or sample values of our latent data, x. Due to the proofs in Section 6.6 of the next chapter, our proof of the estimates for the covariance matrix would follow through that shown in Lee and Nelder (1996).

5.6 Review

To summarise this chapter, we recalled the fundamental theory behind GLMs. From this, we were able to extend to HGLMs, which involve the use of unobserved variables. These unobserved variables actually have their own distributions, thus enabling more underlying flexibility overall. Lee and Nelder (1996) specified a method of estimation called the h-likelihood method, which may be used to provide estimators called MHLEs. These are asymptotically equivalent to MLEs, the latter of which we would use in the absence of latent data; MHLEs also perform well enough to use them instead of MLEs. The basic idea is to have a double maximisation step within each iteration, the first of which maximises the latent data having fixed the parameters temporarily and the second of which maximises the parameters given the recently maximised latent data. This happens until convergence, at which point we obtain our MHLEs. In the following two chapters we will apply this method to a specific case, namely, a time series setting, in order to get MHLEs to help describe election voting.

Chapter 6

The General Model

6.1 Introduction

To summarise so far, chapters have focused on reviewing various aspects: the UK general election, statistical election forecasting methods, time series concepts and the h-likelihood method. From this point, we apply the knowledge gained from these chapters to define and illustrate throroughly our modelling of election data.

This chapter begins by the specification of our time series model in a general sense, including justifications as to why we have chosen this form given the poll data. We then focus on outlining each specific model, which falls firstly into the first-order and higher-order cases and secondly into three different types of probability model; this involves their specification as well as basic properties. The following part of the chapter focuses on the estimation of parameters via h-likelihood, in which we specify the form of the likelihood and then describe how to obtain the parameters given the specific model, as the method differs from model to model. We show a convenient instance where the integrals from the EM algorithm compute tidily for our simplest model; this enables a useful comparison of the h-likelihood with the EM algorithm. It is important to check that the properties proven in Lee and Nelder (1996, [69]) follow through in our time series context; the next section shows the necessary proofs. Following this, we move on to our more complex models, the volatility models, and study their features using simulations.

Once we obtain the parameter estimates we need to infer about them. The next section is devoted to this, showing how to derive approximate confidence intervals via approximate correlation matrices. Finally, subsequent inference tools - comprising model selection, hypothesis testing and simulation analysis - are outlined.

6.2 General Model

6.2.1 Binomial/multinomial distributions

Recall that for some discrete variable y we write $y \sim Bin(n,p)$, for which n is the total number of trials and p is the fixed probability of success, and have the probability mass function:

$$Pr(y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y} & 0 \leq p \leq 1, (y,n) \in \mathbb{N}, y \leq n \\ 0 & \text{otherwise.} \end{cases}$$

Let us assess the assumptions when adopting the binomial distribution in the context of our scenario:

1. A fixed number of trials - for any t, this will certainly be true, as it is simply the total number of votes made (for all parties) in the t^{th} poll.

2. Independence of events - when thinking of general elections, it is an essential requirement that the person voting does so purely on his own opinion and choice, and with no influence of another person's vote. The extent to which bias creeps in here may have an impact; however, it is an assumption which is commonly accepted in election poll analysis. Similarly, we will assume that one poll does not influence another, which in reality is unlikely as once a poll result is published some dependence upon the errors in successive polls may be induced.

3. Only two outcomes - to achieve this, we must focus interest only on one party and have an 'otherwise' category; here, 'otherwise' may be either simply the remainder of votes, that is, the combined total number of votes for all parties other than the party of interest, or another party of interest, that is, where we temporarily neglect the remainder of the data and focus just on data for our two parties, by re-scaling. For instance, if for some general election we were interested in simply comparing the performances of Labour and Conservative, and no other parties, then we would simply take all collected data concerning that election for these two parties. We would then - without loss of generality - choose the number of votes for one of these parties to be our random variable, say Labour.

4. A constant probability of success - we are effectively saying that at the very time when the polls were collected, the chance of voting for Labour was fixed per individual. Clearly then, we are, one may say, restricted by

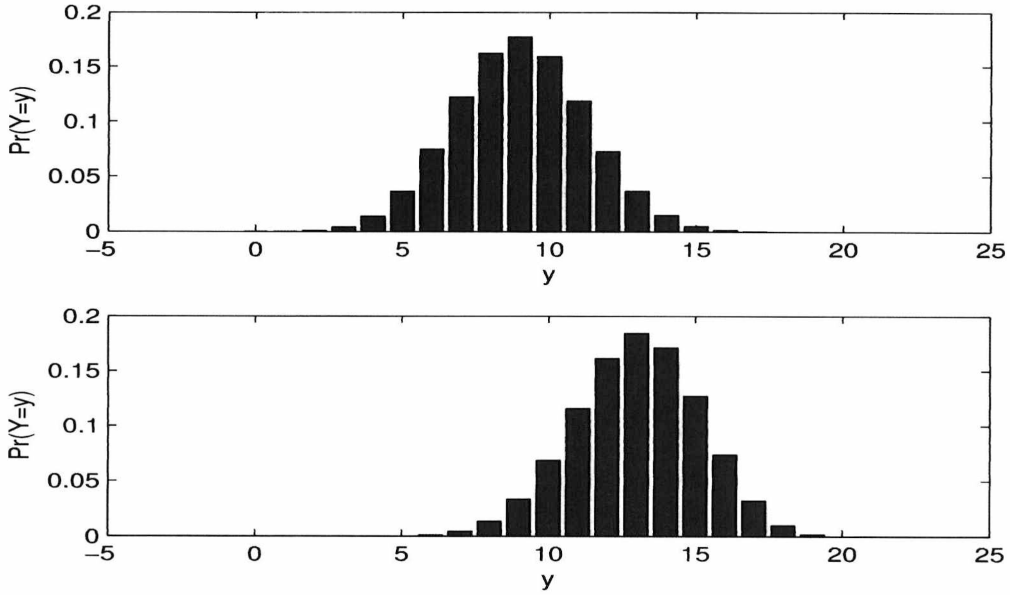


Figure 6.1: Typical spread of $Bin(n, p)$ with simple examples $Bin(20, 0.45)$ (upper) and $Bin(20, 0.65)$ (lower).

operating within discrete time; a continuous-time model would improve the malleability of the probability with respect to time. However, the fact that we are letting this probability vary between polls gives us some relief from this limitation, that is, we are sequentially modelling the probability of voting implicitly in the overall model.

Figure 6.1 illustrates the typical way in which values are distributed when modelled by the binomial distribution. As we can see, it looks quite symmetrical and bell shaped given the kinds of values of p in which we will be interested (roughly between 0.25-0.75). Recall that $\mathbb{E}(y) = np$ (9 and 13 respectively for our examples), so practically this will mean that for a given poll we would expect np to vote for the party represented by y. Also, $\mathbb{V}ar(y) = np(1 - p)$ (4.95 and 4.55 respectively for our examples).

From this then, it seems justifiable to assume the binomial distribution in modelling the number of votes for a specific party at poll t. Goodness of fit to chosen datasets will be looked at later.

Extension

In order to deal with more than just two parties, we must extend the binomial distribution to the multinomial distribution.

For some column vector $\mathbf{y}$ of discrete variables $(y_1, y_2, \ldots, y_k)'$, in which there are k variables, we will write $\mathbf{y} \sim MN(n, p_1, p_2, \ldots, p_k)$, for which n is the total number of trials and p_j, for $j = 1, 2, \ldots, k$ is the fixed probability of success for the j^{th} outcome. Then, the probability mass function is:

$$Pr(y_1, y_2, \ldots, y_k) = \begin{cases} \frac{n!}{y_1! y_2! \ldots y_k!} p_1^{y_1} p_2^{y_2} \ldots p_k^{y_k} & 0 \leq p_j \leq 1, (y_j, n) \in \mathbb{N}, y_j \leq n \\ 0 & \text{otherwise,} \end{cases}$$

where $\sum_{j=1}^{k} p_j = 1$ and $\sum_{j=1}^{k} y_j = n$.

The assumptions outlined earlier follow through, except that there are now not only two outcomes but k.

6.2.2 Beta/Dirichlet distributions

We are working with probabilities, so obviously require the distribution to lie in [0, 1].[1]

Let $f(x) \sim Be(\alpha, \beta)$ for $\alpha, \beta > 0$. Therefore we have the probability density function:

$$f(x) = \begin{cases} x^{\alpha-1}(1-x)^{\beta-1}\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta)) & x \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

Typically, we would expect individual probabilities of voting for the major parties (Labour, Conservative, Liberal Democrats, Scottish and Remainder) to be far enough away from either 0 or 1. Also, it is unlikely that our shape parameters will be equivalent (for that would imply equal strength of parties); neither will they be small in value, such as below 1, as will be seen later. Therefore, we have probability density functions of the form shown in Figure 6.2. Note here that the actual shape parameters given are very small, compared to what we will see with our model. However, these graphs illustrate the general shape of what we will assume the probabilities of voting to look like.

Location and dispersion

$$\text{mode}(x) = \frac{\alpha - 1}{\alpha + \beta - 2}, \quad \mathbb{E}(x) = \frac{\alpha}{\alpha + \beta}, \quad \mathbb{V}\text{ar}(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

[1] Note that frequently in this thesis we refer to functions as f, but this is not the same f throughout. We clarify the specific form as and when necessary; here, for instance, f is a beta density.

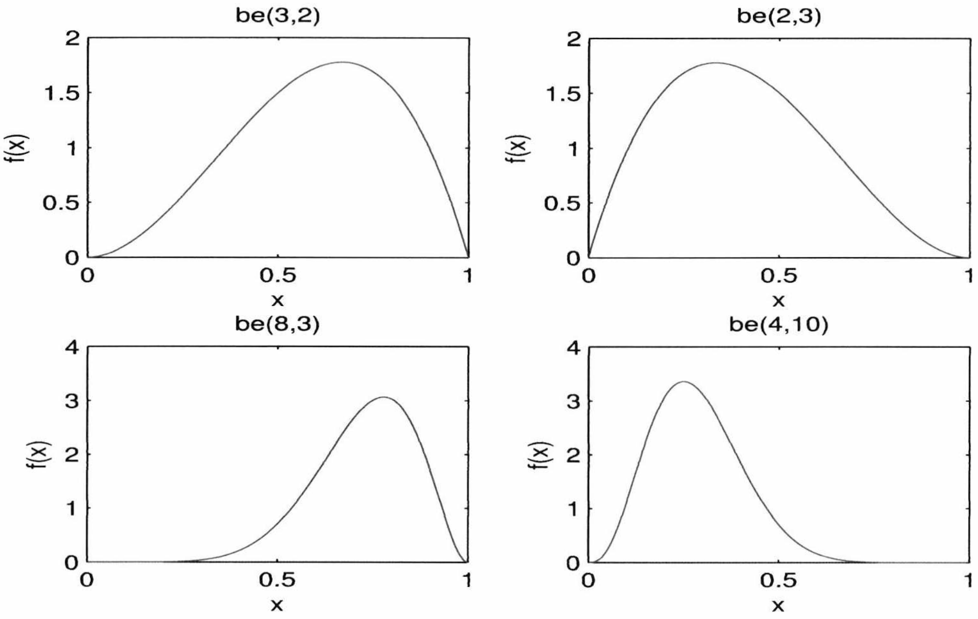


Figure 6.2: Typical forms for $Be(\alpha, \beta)$ when $\alpha > \beta$ (left graphs) and $\alpha < \beta$ (right graphs).

Extension

In order to deal with more than just two parties, we must extend the beta distribution to the Dirichlet distribution.

Let $f(\mathbf{x}) \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_k)$ for $\alpha_1, \alpha_2, \ldots, \alpha_k > 0$, in which there are k variables and let $j = 1, 2, \ldots, k$. Therefore we have the probability density function:

$$f(x_1, x_2, \ldots, x_k) = \begin{cases} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \ldots x_k^{\alpha_k - 1} \Gamma(\sum_{j=1}^k \alpha_j)/(\prod_{j=1}^k \Gamma(a_j)) & x_j \in [0,1] \\ 0 & \text{otherwise,} \end{cases}$$

where $\sum_{j=1}^k x_j = 1$.

6.2.3 Model specification

First, consider the simplest two-party contest, say, Conservative versus Labour: the Markov case. The first step is to define

$$y_t | \, x_t \sim Bin(r_t, \, x_t), \tag{6.1}$$

for $t = 1, 2, \ldots, n$, where y_t is the number of votes for a party of our choice in the election poll conducted at time t, r_t is the corresponding total number of votes in that poll, x_t is the probability of voting for that party at the time of the poll, for each individual voter in the UK, and n is the number of election polls recorded.

The second step involves x_t, which has its own model. x_t will be dependent upon previous historical information, represented by ξ_{t-1}, such that

$$x_t|\, \xi_{t-1} \sim Be(a_t,\, b_t), \tag{6.2}$$

where $a_t = a_t(\xi_{t-1})$ and $b_t = b_t(\xi_{t-1})$.

The time series modelling involves having some x_1, which is the starting point. From this, x_1 generates y_1, x_2 generates y_2 and so on, up to y_n. What we are specifically interested in here is how to generate the $\{x_t\}$ in order to be able to model the $\{y_t\}$, that is, we require particular probability models for each poll. These model the probabilities of voting for parties of interest at the time of, a given poll. Recall Chapter 3, in which we considered three particular types of probability model: the strictly stationary ARCH model, the stochastic volatility model and the GARCH model, the latter two of which are nonstationary. Each type of model considers different historical information, and this is the focus for the rest of this chapter.

More generally, to deal with more than two parties, we extend to a multinomial distribution, that is,

$$\mathbf{y_t}|\, \mathbf{x_t} \sim MN(r_t,\, [\mathbf{x_t}]), \tag{6.3}$$

where $\mathbf{y_t}$ is $p \times 1$ column vector $(y_{1(t)},\, y_{2(t)},\, \ldots,\, y_{p(t)})'$, $\mathbf{x_t}$ is $p \times 1$ column vector $(x_{1(t)}\; x_{2(t)}\; \cdots\; x_{p(t)})'$, $[\mathbf{x_t}] = \{x_{1(t)}, x_{2(t)}, \ldots, x_{p(t)}\}$ and p is the total number of parties of interest. This multinomial distribution is just simply a generalised version of (6.1) to enable us to consider more parties than just two.

Correspondingly, we need to extend the beta model to a Dirichlet model, that is,

$$\mathbf{x_{(t)}}|\, \xi_{\mathbf{t-1}} \sim Dir([a_{k(t)}(\xi_{k(t-1)})]), \tag{6.4}$$

with $[a_{k(t)}] = \{a_{1(t)}, a_{2(t)}, \ldots, a_{p(t)}\}$ and p parties. Again, the Dirichlet distribution is a generalisation of the beta simplification, thus enabling us to progress to consider as many parties as we wish.

For computation purposes, we only have to worry about $p-1$ variables for both x and y, since we know that for some t, $\sum_{k=1}^{p} x_{k(t)} = 1$ and $\sum_{k=1}^{p} y_{k(t)} = r_t$.

We aim to deal with three parties. Then, $p = 3$ in (6.3) and (6.4).

Brown and Payne (1986, [25]) apply a Dirichlet-multinomial model to aggregate voting data, concentrating on transition probabilities for *movements* between the options available to a voter at each election. In their modelling, they make use of ecological regression techniques.

6.2.4 Latent variables

In all our models, we have $\mathbf{x}$, which is a matrix of probabilities of voting for individual parties at each poll. They are actually latent variables, which work 'behind the scenes' to assist the time series modelling, in that they provide an underlying flexibility. By this, we mean that at each poll we have a unique probability of voting for a particular party, which is used to model the actual number of votes, y_t, for that party in that poll, yet after we have generated all our $\mathbf{y}$ the $\mathbf{x}$ are effectively phased out. In the modelling, as has been seen, the $\mathbf{x}$ have their own continuous distributions which change within the progressive time series scenario.

Below, we consider the actual models which we will consider fitting to our datasets and investigate the simulations of. First, we focus on the first-order (Markov) models in the two-party and three-party cases, and then on the higher-order (two and three lag) models in the two-party and three-party cases. We specify the beta/Dirichlet shape parameters as well as the mean and variances. The methodology was covered in Chapter 3 (3.3).

6.3 First-order Models

There are three cases which we will deal with:

- For the ARCH model $\xi_{t-1} = y_{t-1}$, that is, the previous observation.
- For the SV model $\xi_{t-1} = x_{t-1}$, that is, the previous probability.
- For the GARCH model $\xi_{t-1} = (x_{t-1}, y_{t-1})$, that is, the previous probability and observation, combined in a weight function.

These follow through similarly for the multivariate cases.

KENT
TEMPLEMAN
LIBRARY
UNIVERSITY

6.3.1 ARCH (stationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.1.

- MODEL 1 - ARCH(2,1)

$$a_t = \alpha + y_{t-1}, \;\; b_t = \beta + r_{t-1} - y_{t-1} \qquad t = 2, \ldots, n,$$
$$a_1 = \alpha, \;\; b_1 = \beta,$$

with

$$\mathbb{E}(x_t | \, y_{t-1}) = \frac{\alpha + y_{t-1}}{\alpha + \beta + r_{t-1}} \quad \text{and}$$
$$\mathbb{V}ar(x_t | \, y_{t-1}) = \frac{(\alpha + y_{t-1})(\beta + r_{t-1} - y_{t-1})}{(\alpha + \beta + r_{t-1})^2(\alpha + \beta + r_{t-1} + 1)},$$
$$\mathbb{E}(x_1) = \frac{\alpha}{\alpha + \beta} \quad \text{and}$$
$$\mathbb{V}ar(x_1) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

- MODEL 2 - ARCH(3,1)

$$a_{j(t)} = \alpha_j + y_{j(t-1)}, \qquad t = 2, \ldots, n,$$
$$a_{j(1)} = \alpha_j \qquad j = 1, 2, 3,$$

with

$$\mathbb{E}(x_{j(t)} | \, y_{j(t-1)}) = \frac{\alpha_j + y_{j(t-1)}}{\sum_{i=1}^{3} \alpha_i + r_{t-1}} \quad \text{and}$$
$$\mathbb{V}ar(x_{j(t)} | \, y_{j(t-1)}) = \frac{(\alpha_j + y_{j(t-1)})(\sum_{i=1}^{3}\{\alpha_i\} + r_{t-1} - (\alpha_j + y_{j(t-1)}))}{(\sum_{i=1}^{3}\{\alpha_i\} + r_{t-1})^2(\sum_{i=1}^{3}\{\alpha_i\} + r_{t-1} + 1)}.$$

By their construction, we can see that these are particularly simple models in that the shape parameters comprise an additive combination of the parameter representing party strength and the poll vote for that party.

6.3.2 SV (nonstationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.2 and 3.4.

- MODEL 3 - SV(2,1)

$$\begin{aligned}&a_t = c_t x_{t-1}, \;\; b_t = c_t(1 - x_{t-1})),\\ &\text{where } c_t = \exp(-\phi\delta_t)/(1 - \exp(-\phi\delta_t)) \qquad t = 2, \dots, n\\ &(a_1,\, b_1 \text{ are our hypothetical choices})\end{aligned}$$

and with

$$\begin{aligned}\mathbb{E}(x_t|\, x_{t-1}) &= x_{t-1} \quad \text{and}\\ \mathbb{V}ar(x_t|\, x_{t-1}) &= \frac{x_{t-1}(1 - x_{t-1})}{c_t + 1}.\end{aligned}$$

- MODEL 4 - SV(3,1)

$$a_{j(t)} = c_t x_{j(t-1)} \qquad j = 1, 2, 3,$$

where $c_t = \exp(-\phi\delta_t)/(1 - \exp(-\phi\delta_t))$ for $t = 2, \dots, n$, where $a_{j(1)}$ are our hypothetical choices and with

$$\begin{aligned}\mathbb{E}(x_{j(t)}|\, x_{j(t-1)}) &= x_{j(t-1)} \quad \text{and}\\ \mathbb{V}ar(x_{j(t)}|\, x_{j(t-1)}) &= \frac{x_{j(t-1)}(1 - x_{j(t-1)})}{c_t + 1}.\end{aligned}$$

With these models, the probability of voting at poll $t-1$ distributes the value of c_t amongst the shape parameters. Notice how the conditional mean is just the previous probability of voting, and how the conditional variance increases with δ_t, the time interval between polls, thus giving us the nonstationarity aspect:

$$\lim_{\delta_t \to 0} \mathbb{V}ar(x_t|\, x_{t-1}) = 0$$

whereas

$$\lim_{\delta_t \to \infty} \mathbb{V}ar(x_t|\, x_{t-1}) = x_{t-1}(1 - x_{t-1}).$$

By contrast, in the ARCH models the comparative variance is *not* a function of δ_t.

6.3.3 GARCH (nonstationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.3 and 3.4.

- MODEL 5 - GARCH(2,1)

$$\begin{aligned}
&a_t = c_t(ux_{t-1} + (1-u)y_{t-1}/r_{t-1})),\\
&b_t = c_t(u(1-x_{t-1}) + (1-u)((1-y_{t-1}/r_{t-1}))),\\
&\text{where } c_t = \exp(-\phi\delta_t)/(1-\exp(-\phi\delta_t)) \qquad t = 2,\ldots,n\\
&(a_1,\ b_1 \text{ are our hypothetical choices})
\end{aligned}$$

and with

$$\mathbb{E}(x_t|\, x_{t-1}, y_{t-1}) = \left(ux_{t-1} + (1-u)\frac{y_{t-1}}{r_{t-1}}\right) \quad \text{and}$$

$$\mathbb{V}ar(x_t|\, x_{t-1}, y_{t-1}) =$$

$$\frac{\left(ux_{t-1} + (1-u)\frac{y_{t-1}}{r_{t-1}}\right)\left(u(1-x_{t-1}) + (1-u)\left(1-\frac{y_{t-1}}{r_{t-1}}\right)\right)}{c_t+1}.$$

- MODEL 6 - GARCH(3,1)

$$\begin{aligned}
&a_{j(t)} = c_t(ux_{j(t-1)} + (1-u)y_{j(t-1)}/r_{t-1}) \qquad j = 1,2,3,\\
&\text{where } c_t = \exp(-\phi\delta_t)/(1-\exp(-\phi\delta_t)) \qquad t = 2,\ldots,n\\
&(a_{j(1)} \text{ are our hypothetical choices})
\end{aligned}$$

with weight $0 \leq u \leq 1$ and with

$$\begin{aligned}
\mathbb{E}(x_{j(t)}|\, x_{j(t-1)}, y_{j(t-1)}) &= ux_{j(t-1)} + (1-u)y_{j(t-1)}/r_{t-1} \quad \text{and}\\
\mathbb{V}ar(x_{j(t)}|\, x_{j(t-1)}, y_{j(t-1)}) &= \frac{(\mathbb{E}(x_{j(t)}|\,\cdot))(1-\mathbb{E}(x_{j(t)}|\,\cdot))}{c_t+1}.
\end{aligned}$$

With these models, the probability of voting at poll $t-1$ *and* the number of votes at poll $t-1$ distribute the value of c_t amongst the shape parameters. The extent to which one dominates over the other in influencing c_t depends on the weight u. As with the stochastic volatility models, notice how the conditional variance increases with δ_t, the time interval between polls, thus giving us the nonstationarity aspect as opposed to with the ARCH models.

6.4 Higher-order Models

x_t will now be dependent upon more previous historical information, represented by ξ_{t-l}, in which l is the number of lags, such that

$$x_t|\, [\xi_{t-l}] \sim Be(a_t([\xi_{t-l}]),\ b_t([\xi_{t-l}])), \tag{6.5}$$

where $[\xi_{t-l}] = \{\xi_{t-1},\, \xi_{t-2},\, \ldots,\, \xi_{t-l}\}$. The Dirichlet extension is the same idea.

Second Order:

There are three cases which we will deal with:

- For the ARCH model $\xi_{t-1} = y_{t-1}$ and $\xi_{t-2} = y_{t-2}$.
- For the SV model $\xi_{t-1} = x_{t-1}$ and $\xi_{t-2} = x_{t-2}$.
- For the GARCH model $\xi_{t-1} = (x_{t-1},\, y_{t-1})$ and $\xi_{t-2} = (x_{t-2},\, y_{t-2})$.

These follow through similarly for the multivariate cases.

6.4.1 ARCH (stationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.4.

- MODEL 7 - (MTD)ARCH(2,2)

$$\begin{aligned}
&a_t = \alpha + y_{t-2},\;\; b_t = \beta + r_{t-2} - y_{t-2} && t = 3, \ldots, n,\\
&a_t = \alpha + y_{t-1},\;\; b_t = \beta + r_{t-1} - y_{t-1} && t = 2, \ldots, n,\\
&a_1 = \alpha,\;\; b_1 = \beta, &&
\end{aligned}$$

which are used in

$$\begin{aligned}
&wf(x_t|\, y_{t-1}) + (1-w)f(x_t|\, y_{t-2}), && t = 3, \ldots, n,\\
&f(x_2|\, y_1), &&
\end{aligned}$$

where $0 \leq w \leq 1$.

$$\begin{aligned}
\mathbb{E}(x_t|\, y_{t-k}) &= \frac{\alpha + y_{t-k}}{\alpha + \beta + r_{t-k}} \quad k = 1, 2, \;\text{ and}\\
\mathbb{V}ar(x_t|\, y_{t-k}) &= \frac{(\alpha + y_{t-k})(\beta + r_{t-k} - y_{t-k})}{(\alpha + \beta + r_{t-k})^2(\alpha + \beta + r_{t-k} + 1)} \quad k = 1, 2,\\
\mathbb{E}(x_1) &= \frac{\alpha}{\alpha + \beta} \quad \text{and}\\
\mathbb{V}ar(x_1) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.
\end{aligned}$$

- MODEL 8 - (MTD)ARCH(3,2)

  Here, we have the same components as with ARCH(3,1), and additionally

$$a_{j(t)} = \alpha_j + y_{j(t-2)} \qquad j = 1, 2, 3, \quad t = 3, \ldots, n,$$

all of which we combine using a weight parameter, w (as discussed in Chapter 4, 3.3.4), that is,

$$\begin{aligned} & wf(\mathbf{x}_t|\,\mathbf{y}_{t-1}) + (1-w)f(\mathbf{x}_t|\,\mathbf{y}_{t-2}), \qquad t = 3,\dots,n, \\ & f(\mathbf{x}_2|\,\mathbf{y}_1), \end{aligned}$$

where $0 \leq w \leq 1$.

$$\mathbb{E}(x_{j(t)}|\,y_{j(t-k)}) = \frac{\alpha_j + y_{j(t-k)}}{\sum_{i=1}^{3}\alpha_i + r_{t-k}} \quad k = 1,2 \text{ and}$$

$$\mathbb{V}ar(x_{j(t)}|\,y_{j(t-k)}) = \frac{(\alpha_j + y_{j(t-k)})(\sum_{i=1}^{3}\{\alpha_i\} + r_{t-k} - (\alpha_j + y_{j(t-k)}))}{(\sum_{i=1}^{3}\{\alpha_i\} + r_{t-k})^2(\sum_{i=1}^{3}\{\alpha_i\} + r_{t-k} + 1)}.$$

By their construction, we can see that these are particularly simple models in that the shape parameters comprise an additive combination of the parameter representing party strength and the poll vote for that party. This is true for each of the two lags. Then, the two are combined additively using a weight function w, to provide the complete probability model.

6.4.2 SV (nonstationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.2 and 3.4.

- MODEL 9 - SV(2,2)

$$\begin{aligned} a_t &= wc_{(1)t}x_{t-1} + (1-w)c_{(2)t}x_{t-2}, \\ b_t &= wc_{(1)t}(1-x_{t-1}) + (1-w)c_{(2)t}(1-x_{t-2}) && t = 3,\dots,n, \\ a_2 &= c_2x_1, \quad b_2 = c_2(1-x_1), \\ \text{where } c_{(1)t} &= \exp(-\phi\delta_{(1)t})/(1-\exp(-\phi\delta_{(1)t})) && t = 2,\dots,n \\ \text{and } c_{(2)t} &= \exp(-\phi\delta_{(2)t})/(1-\exp(-\phi\delta_{(2)t})) && t = 3,\dots,n \end{aligned}$$

(a_1, b_1 are our hypothetical choices).

$$\mathbb{E}(x_t|\,[\xi_{t-2}]) = wx_{t-1} + (1-w)x_{t-2} \text{ and}$$

$$\mathbb{V}ar(x_t|\,[\xi_{t-2}]) = \frac{(wx_{t-1} + (1-w)x_{t-2})(1-(wx_{t-1} + (1-w)x_{t-2}))}{wc_{1(t)} + (1-w)c_{2(t)} + 1}.$$

- MODEL 10 - SV(3,2)

$$\begin{aligned} a_{j(t)} &= wc_{1(t)}x_{j(t-1)} + (1-w)c_{2(t)}x_{j(t-2)} && j = 1,2,3, \quad t = 3,\dots,n, \\ a_{j(2)} &= c_{1(2)}x_{j(1)}, \end{aligned}$$

where $c_{i(t)} = \exp(-\phi\delta_{i(t)})/(1-\exp(-\phi\delta_{i(t)}))$ for $t = 2,\ldots,n$, $i = 1, 2$, and where $a_{j(1)}$ are our hypothetical choices. We have now introduced to the stochastic volatility model a weight w such that $0 \leq w \leq 1$.

$$\begin{aligned}
&\mathbb{E}(x_{j(t)}|\,[\xi_{j(t-2)}]) = wx_{j(t-1)} + (1-w)x_{j(t-2)} \quad \text{and} \\
&\mathbb{V}ar(x_{j(t)}|\,[\xi_{j(t-2)}]) = \\
&\qquad \frac{(wx_{j(t-1)} + (1-w)x_{j(t-2)})(1-(wx_{j(t-1)} + (1-w)x_{j(t-2)}))}{wc_{1(t)} + (1-w)c_{2(t)} + 1}.
\end{aligned}$$

With these models, the probability of voting at polls $t-1$ and $t-2$ distribute the value of c_t amongst the shape parameters. The weight w determines which of the two dominates. As with the Markov cases, the conditional variance increases with δ_t, the time interval between polls, thus giving us the nonstationarity aspect, as opposed to the ARCH models in which the comparative variance is *not* a function of δ_t.

6.4.3 GARCH (nonstationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.3 and 3.4.

- MODEL 11 - GARCH(2,2)

$$\begin{aligned}
a_t &= wc_{(1)t}(ux_{t-1} + (1-u)y_{t-1}/r_{t-1}) + \\
&\qquad (1-w)c_{(2)t}(ux_{t-2} + (1-u)y_{t-2}/r_{t-2}), \\
b_t &= wc_{(1)t}(1-(u(1-x_{t-1}) + (1-u)(1-y_{t-1}/r_{t-1}))) + \\
&\quad (1-w)c_{(2)t}(1-(u(1-x_{t-2}) + (1-u)(1-y_{t-2}/r_{t-2}))) \qquad t = 3,\ldots,n, \\
a_2 &= c_{(1)2}(ux_1 + (1-u)y_1/r_1)), \\
b_2 &= c_{(1)2}(u(1-x_1) + (1-u)((1-y_1/r_1))), \\
&\text{where } c_{(1)t} = \exp(-\phi\delta_{(1)t})/(1-\exp(-\phi\delta_{(1)t})) \qquad t = 2,\ldots,n \\
&\quad \text{and } c_{(2)t} = \exp(-\phi\delta_{(2)t})/(1-\exp(-\phi\delta_{(2)t})) \qquad t = 3,\ldots,n
\end{aligned}$$

(a_1, b_1 are our hypothetical choices).

$$\begin{aligned}
\mathbb{E}(x_t|\,[\xi_{t-2}]) &= w\left(ux_{t-1} + (1-u)\frac{y_{t-1}}{r_{t-1}}\right) + (1-w)\left(ux_{t-2} + (1-u)\frac{y_{t-2}}{r_{t-2}}\right) \\
\mathbb{V}ar(x_t|\,[\xi_{t-2}]) &= \frac{\mathbb{E}(x_t|\,[\xi_{t-2}])(1-\mathbb{E}(x_t|\,[\xi_{t-2}]))}{wc_{1(t)} + (1-w)c_{2(t)} + 1}.
\end{aligned}$$

- MODEL 12 - GARCH(3,2)

$$a_{j(t)} = \sum_{i=1}^{2} wc_{i(t)}(ux_{j(t-i)} + (1-u)y_{j(t-i)}/r_{t-i}) \qquad t = 3, \ldots, n,$$

$$a_{j(2)} = c_{i(2)}(ux_{j(1)} + (1-u)y_{j(1)}/r_1) \quad j = 1, 2, 3,$$

where $c_{i(t)} = \exp(-\phi\delta_{i(t)})/(1-\exp(-\phi\delta_{i(t)})) \qquad t = 2, \ldots, n$

($a_{j(1)}$ are our hypothetical choices),

with weights $0 \leq u \leq 1$ and $0 \leq w \leq 1$.

$$\mathbb{E}(x_{j(t)}|\, [\xi_{j(t-2)}]) = w\left(ux_{j(t-1)} + (1-u)\frac{y_{j(t-1)}}{r_{t-1}}\right) + (1-w)\left(ux_{j(t-2)} + (1-u)\frac{y_{j(t-2)}}{r_{t-2}}\right)$$

$$\mathbb{V}ar(x_{j(t)}|\, [\xi_{j(t-2)}]) = \frac{(\mathbb{E}(x_{j(t)}|\, [\xi_{j(t-2)}]))(1-\mathbb{E}(X_{j(t)}|\, [\xi_{j(t-2)}]))}{wc_{1(t)} + (1-w)c_{2(t)} + 1}.$$

With these models, the probability of voting at polls $t-1$ and $t-2$ *as well as* the number of votes at polls $t-1$ and $t-2$ distribute the value of c_t amongst the shape parameters. The extent to which one dominates over the other in influencing c_t depends on the weight u. As with the stochastic volatility models, notice how the conditional variance increases with δ_t, the time interval between polls, thus giving us the nonstationarity aspect as opposed to with the ARCH models.

Third Order:

There are three cases which we will deal with:

- For the ARCH model $\xi_{t-1} = y_{t-1}$, $\xi_{t-2} = y_{t-2}$ and $\xi_{t-3} = y_{t-3}$.
- For the SV model $\xi_{t-1} = x_{t-1}$, $\xi_{t-2} = x_{t-2}$ and $\xi_{t-3} = x_{t-3}$.
- For the GARCH model $\xi_{t-1} = (x_{t-1}, y_{t-1})$, $\xi_{t-2} = (x_{t-2}, y_{t-2})$ and $\xi_{t-3} = (x_{t-3}, y_{t-3})$.

These follow through similarly for the multivariate cases.

6.4.4 ARCH (stationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.4.

- MODEL 13 - (MTD)ARCH(2,3)

$$\begin{aligned}
&a_t = \alpha + y_{t-3}, \;\; b_t = \beta + r_{t-3} - y_{t-3} && t = 4, \ldots, n,\\
&a_t = \alpha + y_{t-2}, \;\; b_t = \beta + r_{t-2} - y_{t-2} && t = 3, \ldots, n,\\
&a_t = \alpha + y_{t-1}, \;\; b_t = \beta + r_{t-1} - y_{t-1} && t = 2, \ldots, n,\\
&a_1 = \alpha, \;\; b_1 = \beta,
\end{aligned}$$

all of which we combine using weight parameters, w_1 and w_2 (as discussed in Chapter 4, 3.3.4), that is,

$$\begin{aligned}
&w_1 f(x_t|\, y_{t-1}) + w_2 f(x_t|\, y_{t-2}) + (1 - w_1 - w_2) f(x_t|\, y_{t-3}) && t = 4, \ldots, n,\\
&w_1 f(x_3|\, y_2) + (1 - w_1) f(x_3|\, y_1),\\
&f(x_2|\, y_1),
\end{aligned}$$

where $0 \leq w_1,\, w_2 \leq 1$.

$$\begin{aligned}
\mathbb{E}(x_t|\, y_{t-k}) &= \frac{\alpha + y_{t-k}}{\alpha + \beta + r_{t-k}} \quad k = 1, 2, 3, \quad \text{and}\\
\mathbb{V}ar(x_t|\, y_{t-k}) &= \frac{(\alpha + y_{t-k})(\beta + r_{t-k} - y_{t-k})}{(\alpha + \beta + r_{t-k})^2(\alpha + \beta + r_{t-k} + 1)} \quad k = 1, 2, 3,\\
\mathbb{E}(x_1) &= \frac{\alpha}{\alpha + \beta} \quad \text{and}\\
\mathbb{V}ar(x_1) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.
\end{aligned}$$

- MODEL 14 - (MTD)ARCH(3,3)
  Here, we have the same components as with ARCH(3,2), and additionally

$$a_{j(t)} = \alpha_j + y_{j(t-3)} \qquad j = 1, 2, 3, \quad t = 4, \ldots, n,$$

all of which we combine using weight parameters, w_1 and w_2 (into the format outlined in Chapter 4, 3.3.4), that is,

$$\begin{aligned}
&w_1 f(\mathbf{x}_t|\, \mathbf{y}_{t-1}) + w_2 f(\mathbf{x}_t|\, \mathbf{y}_{t-2}) + (1 - w_1 - w_2) f(\mathbf{x}_t|\, \mathbf{y}_{t-3}), && t = 4, \ldots, n,\\
&w_1 f(\mathbf{x}_3|\, \mathbf{y}_2) + (1 - w_1) f(\mathbf{x}_3|\, \mathbf{y}_1),\\
&f(\mathbf{x}_2|\, \mathbf{y}_1),
\end{aligned}$$

where $0 \leq w_1,\ w_2 \leq 1$.

$$\mathbb{E}(x_t|\, y_{t-k}) = \frac{\alpha + y_{t-k}}{\alpha+\beta+r_{t-k}} \quad k = 1, 2, 3, \quad \text{and}$$

$$\mathbb{V}ar(x_t|\, y_{t-k}) = \frac{(\alpha + y_{t-k})(\beta + r_{t-k} - y_{t-k})}{(\alpha+\beta+r_{t-k})^2(\alpha+\beta+r_{t-k}+1)} \quad k = 1, 2, 3,$$

$$\mathbb{E}(x_1) = \frac{\alpha}{\alpha+\beta} \quad \text{and}$$

$$\mathbb{V}ar(x_1) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

By their construction, we can see that these are particularly simple models in that the shape parameters comprise an additive combination of the parameter representing party strength and the poll vote for that party. This is true for each of the three lags. Then, the three are combined additively using weight functions w_1 and w_2, to provide the complete probability model.

6.4.5 SV (nonstationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.2 and 3.4.

- MODEL 15 - SV(2,3)

$$a_t = \sum_{k=1}^{3} w_k c_{(k)t} x_{t-k}, \quad b_t = \sum_{k=1}^{3} w_k c_{(k)t}(1 - x_{t-k}) \quad t = 4, \ldots, n,$$

$$a_3 = w_1 c_{(1)3} x_2 + (1 - w_1) c_{(2)3} x_1,$$

$$b_3 = w_1 c_{(1)3}(1 - x_2) + (1 - w_1) c_{(2)3}(1 - x_1),$$

$$a_2 = c_2 x_1, \quad b_2 = c_2(1 - x_1),$$

$$\text{where}\, c_{(1)t} = \exp(-\phi\delta_{(1)t})/(1 - \exp(-\phi\delta_{(1)t})) \qquad t = 2, \ldots, n,$$

$$c_{(2)t} = \exp(-\phi\delta_{(2)t})/(1 - \exp(-\phi\delta_{(2)t})) \qquad t = 3, \ldots, n,$$

$$c_{(3)t} = \exp(-\phi\delta_{(3)t})/(1 - \exp(-\phi\delta_{(3)t})) \qquad t = 4, \ldots, n$$

and $w_3 = 1 - w_1 - w_2$

(a_1, b_1 are our hypothetical choices).

$$\mathbb{E}(x_t|\, [\xi_{t-3}]) = \sum_{k=1}^{3} w_k x_{t-k} \quad \text{and}$$

$$\mathbb{V}ar(x_t|\, [\xi_{t-3}]) = \frac{(\sum_{k=1}^{3} w_k x_{t-k})(1 - \sum_{k=1}^{3} w_k x_{t-k})}{\sum_{k=1}^{3} w_k c_{k(t)} + 1}.$$

- MODEL 16 - SV(3,3)

$$\begin{aligned} a_{j(t)} &= \sum_{i=1}^{3} w_i c_{i(t)} x_{j(t-i)} \qquad j = 1,2,3, \quad t = 4,\ldots,n, \\ a_{j(3)} &= w_1 c_{1(3)} x_{j(2)} + (1-w_1) c_{1(3)} x_{j(2)}, \\ a_{j(2)} &= c_{1(2)} x_{j(1)}, \end{aligned}$$

where $c_{i(t)} = \exp(-\phi\delta_{i(t)})/(1-\exp(-\phi\delta_{i(t)}))$ for $t = 2,\ldots,n$, and where $a_{j(1)}$ are our hypothetical choices. We have the weights w_1 and w_2 such that $0 \leq w_1 + w_2 \leq 1$.

$$\begin{aligned} \mathbb{E}(x_{j(t)}|\,[\xi_{j(t-3)}]) &= \sum_{k=1}^{3} w_k x_{j(t-k)} \quad \text{and} \\ \mathbb{V}ar(x_{j(t)}|\,[\xi_{j(t-3)}]) &= \frac{(\sum_{k=1}^{3} w_k x_{j(t-k)})(1-\sum_{k=1}^{3} w_k x_{j(t-k)})}{\sum_{w=1}^{3} w_k c_{k(t)} + 1}. \end{aligned}$$

With these models, the probability of voting at polls $t-1$, $t-2$ and $t-3$ distribute the value of c_t amongst the shape parameters. The weights w_1 and w_2 determine which of the three lags dominate. As with the Markov cases, the conditional variance increases with δ_t, the time interval between polls, thus giving us the nonstationarity aspect, as opposed to the ARCH models.

6.4.6 GARCH (nonstationary) models

The first model is for two parties and the second for three parties. These are based on the theory presented in Chapter 3, 3.3.3 and 3.4.

- MODEL 17 - GARCH(2,3)

$$
\begin{aligned}
a_t &= \sum_{k=1}^{3} w_k c_{(k)t}(ux_{t-k} + (1-u)y_{t-k}/r_{t-k}),\\
b_t &= \sum_{k=1}^{3} w_k c_{(k)t}(u(1-x_{t-k}) + (1-u)(1-y_{t-k}/r_{t-k})) \quad t = 4,\dots,n,\\
a_3 &= w_1 c_{(1)3}(ux_2 + (1-u)y_2/r_2) + (1-w_1)c_{(2)3}(ux_1 + (1-u)y_1/r_1),\\
b_3 &= w_1 c_{(1)3}(1-(u(1-x_2) + (1-u)(1-y_2/r_2)))+\\
&\qquad (1-w_1)c_{(2)3}(1-(u(1-x_1) + (1-u)(1-y_1/r_1))),\\
a_2 &= c_{(1)2}(ux_1 + (1-u)y_1/r_1)),\\
b_2 &= c_{(1)2}(u(1-x_1) + (1-u)((1-y_1/r_1))),\\
\text{where } c_{(1)t} &= \exp(-\phi\delta_{(1)t})/(1-\exp(-\phi\delta_{(1)t})) \qquad t = 2,\dots,n,\\
c_{(2)t} &= \exp(-\phi\delta_{(2)t})/(1-\exp(-\phi\delta_{(2)t})) \qquad t = 3,\dots,n,\\
c_{(3)t} &= \exp(-\phi\delta_{(3)t})/(1-\exp(-\phi\delta_{(3)t})) \qquad t = 4,\dots,n
\end{aligned}
$$

and $w_3 = 1 - w_1 - w_2$

(a_1, b_1 are our hypothetical choices).

$$
\begin{aligned}
\mathbb{E}(x_t|\,[\xi_{t-3}]) &= \sum_{k=1}^{3} w_k \left(ux_{t-k} + (1-u)\frac{y_{t-k}}{r_{t-k}}\right)\\
\mathbb{V}ar(x_t|\,[\xi_{t-3}]) &= \frac{\mathbb{E}(x_t|\,[\xi_{t-3}])(1-\mathbb{E}(x_t|\,[\xi_{t-3}]))}{\sum_{k=1}^{3} w_k c_{k(t)} + 1}.
\end{aligned}
$$

- MODEL 18 - GARCH(3,3)

$$
\begin{aligned}
a_{j(t)} &= \sum_{i=1}^{3} w_i c_{i(t)}(ux_{j(t-i)} + (1-u)y_{j(t-i)}/r_{t-i}) \qquad t = 4,\dots,n,\\
a_{j(3)} &= w_1 c_{1(3)}(ux_{j(2)} + (1-u)y_{j(2)}/r_2)\\
&\qquad + (1-w_1)c_{2(3)}(ux_{j(1)} + (1-u)y_{j(1)}/r_1),\\
a_{j(2)} &= c_{1(2)}(ux_{j(1)} + (1-u)y_{j(1)}/r_1), \quad j = 1,2,3,
\end{aligned}
$$

where for $t = 2,\dots,n$, $c_{i(t)} = \exp(-\phi\delta_{i(t)})/(1-\exp(-\phi\delta_{i(t)}))$

($a_{j(1)}$ are our hypothetical choices),

with weights $0 \le u \le 1$ and $0 \le w_1 + w_2 \le 1$.

$$
\begin{aligned}
\mathbb{E}(x_{j(t)}|\,[\xi_{j(t-3)}]) &= \sum_{k=1}^{3} w_k(ux_{j(t-k)} + (1-u)y_{j(t-k)}/r_{t-k})\\
\mathbb{V}ar(x_{j(t)}|\,[\xi_{j(t-3)}]) &= \frac{(\mathbb{E}(x_{j(t)}|\,[\xi_{j(t-3)}]))(1-\mathbb{E}(x_{j(t)}|\,[\xi_{j(t-3)}]))}{\sum_{k=1}^{3} w_k c_{k(t)} + 1}.
\end{aligned}
$$

With these models, the probability of voting at polls $t-1$, $t-2$ and $t-3$ *as well as* the number of votes at polls $t-1$, $t-2$ and $t-3$ distribute the value of c_t amongst the shape parameters. The extent to which one dominates over the other in influencing c_t depends on the weight u. As with the stochastic volatility models, notice how the conditional variance increases with δ_t, the time interval between polls, thus giving us the nonstationarity aspect as opposed to with the ARCH models.

Remark: It should now be clear from the above how, in principle, it is easy to generalise these models to consider *any* number of lags and parties, but how computation in practice gets messy very easily.

6.4.7 Interpretation of parameters

Overall, we have *four* types of parameter: α, ϕ, u and $\mathbf{w}$, and so will from now write $\theta = (\alpha, \mathbf{w})'$ for our ARCH models, $\theta = (\phi, \mathbf{w})'$ for our stochastic volatility models and $\theta = (\phi, u, \mathbf{w})'$ for our GARCH models.

1. α is employed in the ARCH models and governs the strength of the *parties*; thus the higher the value of α_k relative to the others in α the stronger party k is.

2. ϕ is employed as part of the stochastic volatility models in $c_t = \exp(-\phi\delta_t)/(1-\exp(-\phi\delta_t))$ whose distribution is shown in Fig. 6.3 . Primarily, this function is used because of its implications for the variance increasing in direct proportion to δ_t as explained earlier.

3. $\mathbf{w}$ is a weight function which determines the dominance of the *lags*. Therefore, the higher the value of w for the corresponding lag the more influential this lag is.

4. u is a weight function which determines to what extent the data tend towards the previous probabilities x compared to the observations y in y/r.

6.4.8 Practical use

The end state of the general model introduced and discussed in this thesis enables consideration of *up to three parties and up to three lags.* At least in principle, it is not difficult to extend our models to consider as large a selection of parties as we are interested in. The same can be said for the number of lags. However, the

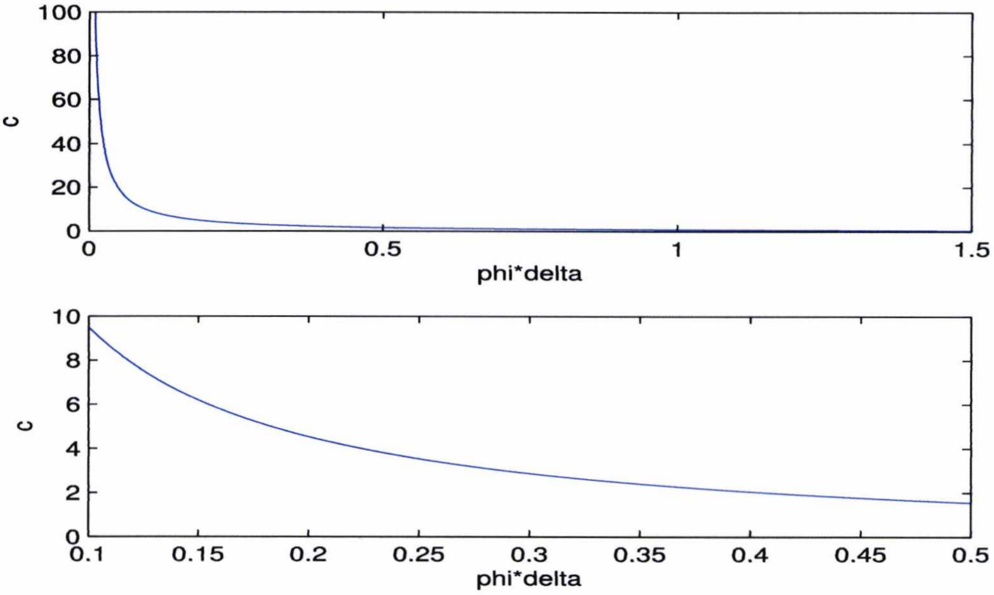


Figure 6.3: Distribution of c_t showing rapid exponential decay.

way in which the programs have been written makes it cumbersome somewhat to extend to a greater number of parties and lags, and more efficiently-written programs would be ideal. The mathematics involved becomes repetitive, and so only poses a slight complication as we increase the number of parties and/or number of lags in time. Nevertheless, in practice, in terms of parties analysts are generally interested only in the dominant three. For the UK, of course these have proven in recent times to be Labour, Conservative and Liberal Democrats. Indeed, as we saw in Chapter 4, Harvey and Shephard (1990, [55]) go only so far as to analyse its model for three parties. Also, the model in Harvey and Shephard (1990) only has a Markovian approach. It may be argued, therefore, that the level of sophistication already developed here will suffice for analysis and discussion.

6.5 Estimation

6.5.1 General idea

The h-likelihood

Consider the univariate case. A simple way to write the likelihood, L, of the general model is

$$L(\theta,\, \mathbf{x}\,;\, \mathbf{y}, \mathbf{r},\, \delta) = \prod_{t=1}^{n} f(y_t|\, x_t) \cdot \prod_{t=l+1}^{n} \{f(x_t|\, [\xi_{t-l}]; \theta, \mathbf{x})\}$$
$$\cdot f(x_l|\, [\xi_{t-l-1}]; \theta, \mathbf{x}) \cdot f(x_{l-1}|\, [\xi_{t-l-2}]; \theta, \mathbf{x}) \cdot \,\ldots\, \cdot f(x_2|\, [\xi_1]; \theta, \mathbf{x}) \cdot f(x_1; \theta, \mathbf{x}).$$

Note that, in the above, we express the likelihood as a function of the *unknown* parameters/variables, for it is these which we are concerned with estimating and not the y, r and δ, which are *known*. The latter point allows us to follow most of the estimation method outlined in the previous chapter, as if it were outside a time series context.

Now let $h = \ln(L)$. Then we have

$$h \propto \sum_{t=k+1}^{n} \{\ln(f(y_t|\, x_t)) + \ln(f(x_t|\, [\xi_{t-k}]; \theta))\}, \tag{6.6}$$

where k is the number of lags. Note that we have no unknown components in our expression for $y_t|\, x_t$, which simplifies our estimation, as we only need to focus on $x_t|$'*history*'. Note also that, in contrast or extension to (5.5), we now have $x|\, y$ as opposed to just x. Obviously though, the y are given and known when mentioned here, as we are in a time series setting now, and thus may be treated as constant values. The multivariate case is a trivial extension of the univariate case.

General procedure

We follow the method described in the previous chapter, whereby we iterate between

$$\frac{\partial h}{\partial x_t}|_{\theta=\theta^{(j)}} = 0 \Rightarrow {x_t}^{(j)},\ \forall\, t \tag{6.7}$$

and

$$\frac{\partial h}{\partial \theta}\,|_{\mathbf{x}_t=\mathbf{x}_t{}^{(j)}} = 0 \Rightarrow \theta^{(j+1)}, \tag{6.8}$$

for $j = 0, 1, 2, \ldots,$ until we have convergence between j and $j+1$, at which point we obtain our maximum hierarchical likelihood estimate (MHLE) vector $\tilde{\theta}$, in addition to our (latent) maximised probabilities $\{\tilde{x}_t\}$ of voting.

The parameter estimates for θ are *always* obtained using the Newton-Raphson method, which we will shortly outline. However, the latent variable estimates are obtained in different ways depending upon the type of model. For the ARCH models they may be maximised *explicitly* as shown in the following subsection; for the volatility models they are maximised using the *Newton-Raphson* method.

6.5.2 Estimation of ARCH models

Part 1

1. The parameters θ are temporarily fixed.

2. We then require the maximum value of the likelihood given the remaining unknown component, which is the $\{x_t\}$.

3. We know from 6.2.2 that the mode of a beta distribution, $\tilde{x}_t$, is
$$\tilde{x}_t = \frac{a_t - 1}{a_t + b_t - 2}, \tag{6.9}$$
which we obtain for $t = 1, 2, \ldots, n$.

4. Once we put our fixed values within θ into (6.9) we have completed the first part *within one iterative step.*

Part 2

1. Now the $\{x_t\}$ are temporarily fixed at $\{\tilde{x}_t\}$.

2. We then require the maximum value of the likelihood given the remaining component, which is θ.

3. We use the iterative Newton-Raphson method in order to get a converged maximum, $\tilde{\theta}$, for θ. Briefly, this states that, with $\theta = (\theta_1, \theta_2, \ldots, \theta_d)'$, that is d parameters, and starting at $j = 0$,
$$\theta^{(j+1)} = \theta^{(j)} - \mathbf{A}^{-1}(\theta^{(j)})\mathbf{g}(\theta^{(j)}),$$
where
$$\mathbf{A} = \begin{pmatrix} \frac{\partial^2 h}{\partial {\theta_1}^2} & \frac{\partial^2 h}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 h}{\partial \theta_1 \partial \theta_d} \\ \frac{\partial^2 h}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 h}{\partial {\theta_2}^2} & \cdots & \frac{\partial^2 h}{\partial \theta_2 \partial \theta_d} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 h}{\partial \theta_d \partial \theta_1} & \frac{\partial^2 h}{\partial \theta_d \partial \theta_2} & \cdots & \frac{\partial^2 h}{\partial {\theta_d}^2} \end{pmatrix}$$

and

$$\mathbf{g} = \left(\frac{\partial h}{\partial \theta_1}, \frac{\partial h}{\partial \theta_2}, \ldots, \frac{\partial h}{\partial \theta_d} \right)'.$$

4. Then we have completed the second part within an iterative step.

We simply repeat the above procedure several times, using the most recent converged solutions as the next starting values, until convergence of both components.

MTD models: For the MTD (higher-order) models we must slightly add to this procedure, since we have introduced new dummy variables, $\{\mathbf{z}_t\}$, (see Chapter 3, 3.3 for details) into the likelihood in order to make computation easier. Consider the simplest MTD model, ARCH(2,2) for illustration purposes.

1. The parameters, θ, and $\{z_t\}$ are temporarily fixed.

2. Recall that the $\{z_t\}$ are attached to the lag weight w, having a value of 1 if the first lag dominates and 0 if the second lag dominates, for each poll. Consequently, the partial derivatives $\partial h/\partial w$ and $\partial^2 h/\partial w^2$ are purely functions of $w|\,\mathbf{z}$. Hence we now maximise $w|\,\mathbf{z}$, again using Newton-Raphson.

3. We then require the maximum value of the likelihood given $\{x_t\}$, which involves carrying out Part 1, with the only difference of the $\{z_t\}$ governing which lag is used:
$$\tilde{x}_t = \frac{\alpha + y_t + y_{t-1} - 1}{\alpha + \beta + r_t + r_{t-1} - 2},$$
if $z_t = 1$ and
$$\tilde{x}_t = \frac{\alpha + y_t + y_{t-2} - 1}{\alpha + \beta + r_t + r_{t-2} - 2},$$
if $z_t = 0$, which we obtain for $t = 3, 4, \ldots, n$.

4. Once we put our fixed values within θ into the $\{\tilde{x}_t\}$ the $\{x_t\}$ are temporarily fixed.

5. We then require the maximum value of the likelihood given the remaining component, which is θ. This is Part 2 above, with the only difference of the pre-determined $\{z_t\}$ governing which lag is used and appearing in the partial derivatives.

6. The next task is to revise the $\{z_t\}$:

$$\text{If } f(x_t|\, y_{t-1}) > f(x_t|\, y_{t-2}) \text{ then } z_t = 1.$$
$$\text{If } f(x_t|\, y_{t-1}) < f(x_t|\, y_{t-2}) \text{ then } z_t = 0.$$

7. In summary, so far we have updated each component: w, $\{x_t\}$, θ and $\{z_t\}$, and so have completed the first iterative step.

We repeat the above several times, using the most recent converged solutions as the next starting values, until convergence of all components.

EM algorithm estimation

The EM algorithm is a common tool used to produce maximum likelihood estimates in the presence of missing data. In the algorithm we:

1. obtain conditional expectations of missing data given parameters which maximise the current likelihood (E-step); and

2. seek parameters which maximise the revised likelihood given the latest expected values for the missing data (M-step),

until we have convergence of both.

Within each iteration of the h-likelihood maximisation, we are alternating between:

1. finding latent variables which maximise the current h-likelihood given fixed parameters; and

2. finding parameters which maximise the revised h-likelihood given the recently obtained latent variables

until we have convergence of both.

There is clear similarity of the two methods. The similarities are that we treat our latent data as missing data, and we have an iterative algorithm which alternates in two stages. The M-step is essentially the same for both methods. The E-step will be replaced by another M-step; the main difference is that now we do not work with what we expect the missing data to be, but instead with values for these data which will maximise the likelihood given the fixed parameters. In the following section we will verify that the two estimators have an asymptotically negligible difference as well as an asymptotically common variance. The basis for this proof was shown in Chapter 5. Also, since the MLE is asymptotically most efficient and invariant, then so is the MHLE.

The main problem with the EM algorithm is that it is based on integrals, which are often difficult if not impossible to evaluate. In our case, the integration gets complicated very easily and quickly as we add to the sophistication of our model. Nevertheless, we can - in the simplest (Markov ARCH) case - obtain the necessary integrals and so forth, in order to derive the MLE vector $\hat{\theta} = (\alpha, \beta)$'. Later, we will compare the output of a method which has been proven to give the (local if not global) MLE with the corresponding MHLE obtained from our own h-likelihood method. The closer they are the more valid and therefore useful they will be. Now we outline the mathematics to enable this.

Estimating ARCH(2,1) parameters via the EM algorithm: We have

$$\begin{aligned} h \propto \sum_{t=2}^{n} \{ &(y_t + \alpha + y_{t-1} - 1)\ln(x_t) + (r_t - y_t + \beta + r_{t-1} - y_{t-1} - 1)\ln(1 - x_t) + \\ &\ln B(\alpha + y_{t-1}, \beta + r_{t-1} - y_{t-1})\} + \\ &(y_1 + \alpha - 1)\ln(x_1) + (r_1 - y_1 + \beta - 1)\ln(1 - x_1) + \ln B(\alpha, \beta). \end{aligned}$$

Of course, if X_t were known then we simply proceed using a regular optimisation method such as the Newton-Raphson method. The X_t creates the problem in that it is latent data; so we may resolve with the EM algorithm by treating the X_t as if it were missing data. The E-step of the EM algorithm (see Appendix B) implies that we must evaluate both[2]:

$$\int_0^1 \ln(x_t) Be(\alpha + y_{t-1} + y_t, \beta + r_{t-1} - y_{t-1} + r_t - y_t; x_t | \, y_{t-1}) \, \mathrm{d}x_t \text{ and}$$

$$\int_0^1 \ln(1 - x_t) Be(\alpha + y_{t-1} + y_t, \beta + r_{t-1} - y_{t-1} + r_t - y_t; x_t | \, y_{t-1}) \, \mathrm{d}x_t.$$

Fortunately, these compute rather neatly into

$$\begin{aligned} &\Psi(\alpha + y_{t-1} + y_t) - \Psi(\alpha + \beta + r_{t-1} + r_t) \qquad \text{and} \\ &\Psi(\beta + r_{t-1} - y_{t-1} + r_t - y_t) - \Psi(\alpha + \beta + r_{t-1} + r_t) \end{aligned}$$

respectively, in which $\Psi(u)$ is the digamma function[3] for some variable u.

What we must do now is choose our starting point, say, $\theta^{(1)}$ (which is effectively $\hat{\theta}^{(1)}$). We then substitute this into our formulae above and we have essentially 'augmented' our likelihood, at least within the first iteration. It is just a straightforward implementation of the Newton-Raphson method to find $\hat{\theta}^{(2)}$ (the M-step). We re-evaluate the expected values for each x_t now (to repeat the E-step for the next iteration), maximise this and so on until we have convergence of θ between some step j' and $j' + 1$, as required.

[2] We have similar cases for the $t = 1$ special case.

[3] $\Psi(u) = \mathrm{d}(\ln(\Gamma(u)))/\mathrm{d}u$.

Example

Here, we demonstrate step-by-step how to go about obtaining the MHLEs for our simplest ARCH model, ARCH(2,1). Consider again Parts 1 and 2 at the start of this subsection.

Recall that

$$\begin{aligned} &y_t|\, x_t \sim Bin(r_t,\, x_t) && t = 1,\, 2,\, \ldots,\, n \\ &x_t|\, y_{t-1} \sim Be(\alpha + y_{t-1},\, \beta + r_{t-1} - y_{t-1}) && t = 2,\, 3,\, \ldots,\, n \\ &x_1 \sim Be(\alpha,\, \beta). && \end{aligned}$$

Then, we have the likelihood, which may be written as

$$\begin{aligned} L(\theta, \mathbf{x}|\, \mathbf{y}, \mathbf{r}) \propto & \frac{{x_1}^{y_1+\alpha-1}(1-x_1)^{r_1-y_1+\beta-1}}{B(\alpha,\, \beta)} \times \\ & \prod_{t=2}^{n} \left\{ \frac{{x_t}^{y_t+\alpha+y_{t-1}-1}(1-x_t)^{r_t-y_t+\beta+r_{t-1}-y_{t-1}-1}}{B(\alpha + y_{t-1},\, \beta + r_{t-1} - y_{t-1})} \right\}. \end{aligned}$$

From this, we derive h by simply taking logs, so that

$$\begin{aligned} h \propto\ & \alpha \text{ln} x_1 + \beta \text{ln}(1-x_1) + \text{ln}\Gamma(\alpha+\beta) - \text{ln}\Gamma(\alpha) - \text{ln}\Gamma(\beta) + \\ & \sum_{t=2}^{n} \{\alpha \text{ln} x_t + \beta \text{ln}(1-x_t) + \text{ln}\Gamma(\alpha+\beta+r_{t-1}) - \text{ln}\Gamma(\alpha+y_{t-1}) - \\ & \qquad \text{ln}\Gamma(\beta + r_{t-1} - y_{t-1})\}. \end{aligned}$$

Part 1:

1. We start (at $k = 0$) by fixing $\theta = (\alpha,\, \beta)'$ at $\theta^{(\mathbf{k})} = (\alpha^{(k)},\, \beta^{(k)})'$.

2. We now maximise the likelihood given the unknown (latent) component, $\{x_t\}$, to be estimated, $\{\tilde{x}_t\}$. This involves solving $\partial L/\partial x_t = 0$ for x_t in each case. Applying (6.9), our estimates our

$$\begin{aligned} \tilde{x}_1 &= \frac{y_1 + \alpha - 1}{r_1 + \alpha + \beta - 2} \\ \tilde{x}_t &= \frac{y_t + \alpha + y_{t-1} - 1}{r_t + \alpha + \beta + r_{t-1} - 2} && t = 2,\, 3,\, \ldots,\, n. \end{aligned}$$

3. Then, we insert our values of $\theta^{(\mathbf{k})}$ into the above, and thus have completed Part 1 within one iteration, to get $\{\tilde{x}_t^{(k)}\}$.

Part 2:
Now the $\{x_t\}$ are temporarily fixed at $\{\tilde{x}_t^{(k)}\}$, and we focus on maximising L given the unknown θ.

For the Newton-Raphson method, we now require all the partial derivatives[4]:

$$\frac{\partial h}{\partial \alpha} = \ln x_1 + \Psi(\alpha+\beta) - \Psi(\alpha) + \sum_{t=2}^{n}\{\ln x_t + \Psi(\alpha+\beta+r_{t-1}) - \Psi(\alpha+y_{t-1})\}$$

$$\frac{\partial h}{\partial \beta} = \ln(1-x_1) + \Psi(\alpha+\beta) - \Psi(\beta) + \sum_{t=2}^{n}\{\ln(1-x_t) + \Psi(\alpha+\beta+r_{t-1}) - \Psi(\beta+r_{t-1}-y_{t-1})\}$$

$$\frac{\partial^2 h}{\partial \alpha^2} = \Psi'(\alpha+\beta) - \Psi'(\alpha) + \sum_{t=2}^{n}\{\Psi'(\alpha+\beta+r_{t-1}) - \Psi'(\alpha+y_{t-1})\}$$

$$\frac{\partial^2 h}{\partial \beta^2} = \Psi'(\alpha+\beta) - \Psi'(\beta) + \sum_{t=2}^{n}\{\Psi'(\alpha+\beta+r_{t-1}) - \Psi'(\beta+r_{t-1}-y_{t-1})\}$$

$$\frac{\partial^2 h}{\partial \alpha\beta} = \Psi'(\alpha+\beta) + \sum_{t=2}^{n}\{\Psi'(\alpha+\beta+r_{t-1})\}$$

We put the above into

$$\begin{pmatrix}\alpha^{(j+1)}\\ \beta^{(j+1)}\end{pmatrix} = \begin{pmatrix}\alpha^{(j)}\\ \beta^{(j)}\end{pmatrix} - \frac{1}{|A|}\begin{pmatrix}\partial^2 h/\partial\beta^2 & -\partial^2 h/\partial\alpha\beta\\ -\partial^2 h/\partial\alpha\beta & \partial^2 h/\partial\alpha^2\end{pmatrix}\begin{pmatrix}\partial h/\partial\alpha\\ \partial h/\partial\beta\end{pmatrix},$$

for $j = 0, 1, 2, \ldots$, in which all partial derivatives are evaluated at $\theta^{(j)}$ and where

$$|A| = \frac{\partial^2 h}{\partial\alpha^2}\cdot\frac{\partial^2 h}{\partial\beta^2} - \left[\frac{\partial^2 h}{\partial\alpha\beta}\right]^2.$$

We insert our values of $\{\tilde{x}_t^{(k)}\}$ into the above, and subsequently run the algorithm. Following convergence, we have completed Part 2 for $k = 0$.

We simply repeat the two parts, for $k = 1, 2, \ldots,$ until convergence of both known and unknown components. Note that the values for α and β which we get will be typically large, thereby leading to similar values for the x_t as the iterations progress. Despite this, they do evolve noticeably and so we argue are worth modelling in the way in which we have chosen.

[4] As before, $\Psi(u)$ is the digamma function and $\Psi'(u)$ its derivative with respect to u.

Although the basic method is similar, this process becomes tedious and complicated quite easily as we move on to models with more parties and lags. In Appendix A, we show the MATLAB code for the ARCH(2,3) MTD model, which is considerably more complex.

6.5.3 Estimation of volatility models

Since the stochastic volatility and GARCH models use previous probabilities as historical information we slightly need to modify Part 1 from 6.5.2.

1. The parameters θ are temporarily fixed.

2. We then require the maximum value of the likelihood given the remaining unknown component, which is the $\{x_t\}$.

3. We first use (6.9) to estimate the most recent probability, x_n, denoted by $\tilde{x}_n$:
$$\tilde{x}_n = \frac{a_n - 1}{a_n + b_n - 2}.$$

4. From now we work backwards. First we maximise $x_{n-1}|\,\tilde{x}_n$, denoted by $\tilde{x}_{n-1}$, and filling the rest of the $\mathbf{x}$ vector with the last known values of $x_1,\, x_2,\, \ldots,\, x_{n-2}$. We illustrate with an example. With SV(2,1) (see 6.3) we arrive at the log-likelihood:
$$h \propto \sum_{t=2}^{n} \{c_t(x_{t-1}\ln x_t + (1-x_{t-1})\ln(1-x_t)) - \ln\Gamma(c_t x_{t-1}) - \ln\Gamma(c_t(1-x_{t-1}))\}.$$
After we fill the probability vector as explained above, the remaining unknown X_{n-1} appears only in
$$c_n x_{n-1}(\ln x_n - \ln(1-x_n)) - \ln\Gamma(c_n x_{n-1}) - \ln\Gamma(c_n(1-x_{n-1}))$$
from the n^{th} term and
$$c_{n-1}x_{n-2}(\ln x_{n-1} - \ln(1-x_{n-1}))$$
from the $(n-1)^{th}$ term.

   Therefore, we proceed applying the Newton-Raphson method discussed earlier to h to find $\tilde{x}_{n-1}$.

5. We repeat a similar process, up to and including the final calculation of maximising $x_1|\,\tilde{x}_n,\, \tilde{x}_{n-1},\, \ldots,\, \tilde{x}_2$.

6. Once we put our fixed values within θ into each $\tilde{x}_t$ we have completed the first part within one iterative step.

Example code demonstrating the general procedure, and the volatility and MTD special cases is given in Appendix A.

Example

Here, we demonstrate step-by-step how to go about obtaining the MHLEs for our simplest volatility model, SV(2,1). Consider again Parts 1 and 2.

Recall that

$$\begin{aligned} y_t|\, x_t &\sim Bin(r_t,\, x_t) & t &= 1,\, 2,\, \ldots,\, n \\ x_t|\, x_{t-1} &\sim Be(c_t x_{t-1},\, c_t(1-x_{t-1})) & t &= 2,\, 3,\, \ldots,\, n, \end{aligned}$$

where x_1 is a chosen value and $c_t = \exp(-\phi\delta_t)/(1-\exp(-\phi\delta_t))$.

Then, we have the likelihood, which may be written as

$$L(\theta, \mathbf{x}|\, \mathbf{y}, \mathbf{r}, \delta) \propto \prod_{t=2}^{n} \left\{ \frac{{x_t}^{y_t + c_t x_{t-1} - 1}(1-x_t)^{r_t - y_t + c_t(1-x_{t-1})-1}}{B(c_t x_{t-1},\, c_t(1-x_{t-1}))} \right\}.$$

From this, we derive h by simply taking logs, so that

$$h \propto \sum_{t=2}^{n} \{c_t x_{t-1} \ln x_t + c_t(1-x_{t-1})\ln(1-x_t) + \ln\Gamma(c_t) - \ln\Gamma(c_t x_{t-1}) - \ln\Gamma(c_t(1-x_{t-1}))\}.$$

Part 1:

1. We start (at $k=0$) by fixing ϕ at $\phi^{(k)}$.

2. We now maximise the likelihood given the unknown (latent) component, $\{x_t\}$, to be estimated, $\{\tilde{x}_t\}$. We saw that with the ARCH model this simply involved solving $\partial L/\partial x_t = 0$ for x_t to get an explicit estimate. However, now since the historical information is previous probabilities the $\{x_t\}$ appear in powers of terms, effectively meaning that we cannot obtain an explicit estimate.

3. First we use (6.9) to estimate the most recent probability, x_n, denoted by $\tilde{x}_n$, which is explicit:
$$\tilde{x}_n = \frac{y_n + c_n x_{n-1} - 1}{r_n + c_n - 2},$$
where we start with some suitable *initial* choice of values for the vector $\mathbf{x} = (x_1,\, x_2,\, \ldots,\, x_n)'$.

4. From now we work backwards. First we maximise $x_{n-1}|\,\tilde{x}_n$, denoted by $\tilde{x}_{n-1}$, and filling the rest of the $\mathbf{x}$ vector with the last known values of $x_1,\ x_2,\ \ldots,\ x_{n-2}$.

   As we saw earlier, the unknown x_{n-1} appears only in

   $$c_n x_{n-1}(\ln x_n - \ln(1-x_n)) - \ln\Gamma(c_n x_{n-1}) - \ln\Gamma(c_n(1-x_{n-1}))$$

   from the n^{th} term and

   $$c_{n-1}x_{n-2}(\ln x_{n-1} - \ln(1-x_{n-1}))$$

   from the $(n-1)^{th}$ term.

   Therefore, we proceed applying the Newton-Raphson method discussed earlier to h to find $\tilde{x}_{n-1}$. The derivatives[5] necessary are

   $$\frac{\mathrm{d}h}{\mathrm{d}x_{n-1}} = c_n(\ln(x_n) - \ln(1-x_n) - \Psi(c_n x_{n-1}) + \Psi(c_n(1-x_{n-1})))+ \\ \frac{y_n + c_{n-1}x_{n-2} - 1}{x_{n-1}} - \frac{r_n - y_n + c_{n-1}(1-x_{n-2}) - 1}{1-x_{n-1}}$$

   and

   $$\frac{\mathrm{d}^2h}{\mathrm{d}x_{n-1}^2} = -{c_n}^2\Psi'(c_n x_{n-1})) - {c_n}^2\Psi'(c_n(1-x_{n-1}))- \\ \frac{y_n + c_{n-1}x_{n-2} - 1}{x_{n-1}^2} - \frac{r_n - y_n + c_{n-1}(1-x_{n-2}) - 1}{(1-x_{n-1})^2},$$

   which we put the above into the Newton-Raphson algorithm in *one* dimension:

   $${x_{n-1}}^{(j+1)} = {x_{n-1}}^{(j)} - \frac{\partial h/\partial x_{n-1}}{\partial^2 h/\partial {x_{n-1}}^2}$$

   for $j = 0,\ 1,\ 2,\ \ldots$, in which the derivatives are evaluated at ${x_{n-1}}^{(j)}$.

   We insert our values of c_n, c_{n-1}, y_n, r_n and x_{n-2}, as well as $\phi^{(k)}$, into the above, and subsequently run the algorithm. Following convergence, we have found $\tilde{x}_{n-1}$.

5. We repeat a similar process to find each $\tilde{x}_t$, up to and including the final calculation of maximising $x_1|\,\tilde{x}_n,\ \tilde{x}_{n-1},\ \ldots,\ \tilde{x}_2$.

6. Then, we have completed Part 1 within one iteration, to get $\{\tilde{x}_t^{(k)}\}$.

[5] As before, $\Psi(u)$ is the digamma function and $\Psi'(u)$ its derivative with respect to u.

Part 2:
Now the $\{x_t\}$ are temporarily fixed at $\{\tilde{x}_t^{(k)}\}$, and we focus on maximising L given the unknown ϕ.

For the Newton-Raphson method, we now require the following derivatives:

$$\frac{\mathrm{d}h}{\mathrm{d}\phi} = \sum_{t=2}^{n} \frac{\mathrm{d}c_t}{\mathrm{d}\phi}\{x_{t-1}\ln(x_t) + (1-x_{t-1})\ln(1-x_t) + \Psi(c_t) - x_{t-1}\Psi(c_t x_{t-1}) - (1-x_{t-1})\Psi(c_t(1-x_{t-1}))\}$$

$$\frac{\mathrm{d}^2h}{\mathrm{d}\phi^2} = \sum_{t=2}^{n} \frac{\mathrm{d}^2c_t}{\mathrm{d}\phi^2}\{x_{t-1}\ln(x_t) + (1-x_{t-1})\ln(1-x_t) + \Psi(c_t) - x_{t-1}\Psi(c_t x_{t-1}) - (1-x_{t-1})\Psi(c_t(1-x_{t-1}))\} + \left(\frac{\mathrm{d}c_t}{\mathrm{d}\phi}\right)^2 \{\Psi'(c_t) - x_{t-1}^2\Psi'(c_t x_{t-1}) - (1-x_{t-1})^2\Psi'(c_t(1-x_{t-1}))\}$$

We put the above into the Newton-Raphson algorithm in one dimension

$$\phi^{(j+1)} = \phi^{(j)} - \frac{\partial h/\partial\phi}{\partial^2 h/\partial\phi^2}$$

for $j = 0, 1, 2, \ldots$, in which the derivatives are evaluated at $\phi^{(j)}$.

We insert our values of $\{\tilde{x}_t^{(k)}\}$ into the above, and subsequently run the algorithm. Following convergence, we have completed Part 2 for $k = 0$.

We simply repeat the two parts, for $k = 1, 2, \ldots,$ until convergence of both known and unknown components.

Although the basic method is similar, this process becomes tedious and complicated quite easily as we move on to models with more parties and lags. In Appendix A, we show the MATLAB code for the SV(3,2) model, which is considerably more complex, particularly in terms of estimating the $\{x_t\}$, when the Newton Raphson algorithm increases in dimension.

6.6 Properties of Estimators in ARCH Models

Before following our approach to obtain alternatives to MLEs, we need to justify the usefulness of the MHLEs. Our proof is similar to that provided in Lee and Nelder (1996, [69]) and outlined in Chapter 5 (5.5), except that ours is in a time series context. We must check that our MHLE is asymptotically efficient.

6.6.1 Asymptotic best unbiased predictors

It is simplest to work with the two-party and Markov case for the sake of clarity; the principle follows through smoothly for the multi-party and multi-lag case.

First, let

$$D_t = -\frac{\partial^2 h}{\partial {x_t}^2}$$

evaluated at $x_t = \tilde{x}_t$. Also, let ω_t denote the complete set of data at poll t, be it observed and unobserved, *except* for the total number of polls (r_t); the r_t are singled out as they will play a crucial role in the justification. Note that g_1 and g_2 are positive linear functions.

Equation (6.6) boils down neatly to the form:

$$\begin{aligned} h &\propto \sum_{t=1}^{n} (g_1(\omega_t;\, \theta) \cdot ln(x_t) + g_2(\omega_t, r_t;\, \theta) \cdot ln(1-x_t)) \\ \Longrightarrow \frac{\partial h}{\partial x_t} &= \frac{g_1(\omega_t;\, \theta)}{x_t} - \frac{g_2(\omega_t, r_t;\, \theta)}{1-x_t} \\ \Longrightarrow -\frac{\partial^2 h}{\partial {x_t}^2} &= \frac{g_1(\omega_t;\, \theta)}{{x_t}^2} + \frac{g_2(\omega_t, r_t;\, \theta)}{(1-x_t)^2}. \end{aligned}$$

Therefore, ${D_t}^{-1}$ may be written as:

$$\frac{1}{g_1(\omega_t;\, \theta)\frac{1}{\tilde{x}_t^2} + g_2(\omega_t,\, r_t;\, \theta)\frac{1}{(1-\tilde{x}_t)^2}} = O_p\left(\frac{1}{r_t}\right),$$

since the bigger r_t is the larger the denominator becomes, assuming we fix θ. We will make use of this important result shortly[6].

Lee and Nelder (1996) include a discussion about how $\mathbb{E}(v(x)|\, y)$ is the best unbiased predictor (BUP) for x, where $v(x)$ is some strictly monotonic increasing transformation of x, and has the minimum mean squared error of prediction. It then goes on to prove that, under appropriate conditions, $v(\tilde{x}) \longrightarrow \mathbb{E}(v(x)|\, y)$, so that the MHLEs for v are *asymptotically* BUPs. We have a simplification here in that we work with x and not some transformation of it, i.e. $v(x) = x$, which of course is strictly monotonic increasing.

[6] Note that here $\lim_{r_t \to \infty} O_p(1/r_t) = 0$.

What we must show now is (see Chapter 5, 5.5 for details) that

$$\begin{aligned}\mathbb{E}(x_t|\, y) &= \tilde{x}_t + O_p(1/r_t) \qquad \text{and}\\ \mathbb{V}ar(x_t|\, y) &= {D_t}^{-1}(1 + O_p(1/r_t))\\ &\equiv O_p(1/r_t),\end{aligned}$$

such that, as is shown in Lee and Nelder (1996), asymptotically,

$$x_t|\, y \sim N(\tilde{x}_t, {D_t}^{-1}),$$

where ${D_t}^{-1} = O_p(1/r_t)$ for all t (already shown above).

The above is easy to show, since

$$\tilde{x}_t = \frac{g_1}{g_1 + g_2} \to 0 \quad \text{as} \quad r_t \to \infty,$$

remembering that g_2 is a function of r_t unlike g_1, and similarly

$$\mathbb{E}(x_t|\, y) = \frac{a_{1(t)}}{a_{1(t)} + a_{2(t)}} \to 0 \quad \text{as} \quad r_t \to \infty,$$

where $a_{1(t)}$ and $a_{2(t)}$ are the shape parameters of the beta distribution as shown in (6.2), and since the latter parameter is a function of r_t unlike the former.

Using the same idea, it is trivial that

$$\mathbb{V}ar(x_t|\, y) = \frac{a_{1(t)}a_{2(t)}}{(a_{1(t)} + a_{2(t)})^2(a_{1(t)} + a_{2(t)} + 1)} \to 0 \quad \text{as} \quad r_t \to \infty,$$

as required.

6.6.2 Asymptotic efficiency

Chapter 5 (5.5) outlines the general proof to ensure that

$$|\,\hat{\theta} - \tilde{\theta}\,| = O_p\left(\frac{1}{r_t}\right).$$

Therefore, all we must do is verify that each stage holds in our case. This essentially means that all we must check is that

$$\frac{1}{r_t}\left(\frac{\partial}{\partial\alpha}\right)\left(\frac{\partial h}{\partial x_t}\right)|_{x=\tilde{x}} = O_P(1), \tag{6.10}$$

$$\frac{1}{r_t}\left(\frac{\partial}{\partial\alpha}\right)\left(\frac{\partial^2 h}{\partial {x_t}^2}\right)|_{x=\tilde{x}} = O_P(1) \quad \text{and} \tag{6.11}$$

$$\frac{\partial^2 h}{\partial\alpha^2} = O_P(r_t). \tag{6.12}$$

Firstly,

$$
\begin{aligned}
\frac{\partial h}{\partial x_t} &= \sum_{t=1}^{n}\left(\frac{g_1}{x_t}+\frac{g_2}{1-x_t}\right)\\
\Rightarrow \frac{1}{r_t}\left(\frac{\partial}{\partial\alpha}\right)\left(\frac{\partial h}{\partial x_t}\right)|_{x=\tilde{x}} &= \frac{1}{r_t}\sum_{t=1}^{n}\frac{1}{\tilde{x}_t}\\
\Rightarrow \frac{1}{r_t}\left(\frac{\partial}{\partial\alpha}\right)\left(\frac{\partial^2 h}{\partial {x_t}^2}\right)|_{x=\tilde{x}} &= -\frac{1}{r_t}\sum_{t=1}^{n}\frac{1}{\tilde{x}_t^2}
\end{aligned}
$$

Hence in order to have these have order $O_P(1)$ we require that

$$r_t = O_P(n). \tag{6.13}$$

Finally[7],

$$
\begin{aligned}
\frac{\partial^2 h}{\partial\alpha^2} &= \frac{\partial}{\partial\alpha}\left(\sum_{t=1}^{n}\left(\ln(x_t)+\Psi(a_{1(t)}+a_{2(t)})-\Psi(a_{1(t)})-\Psi(a_{2(t)})\right)\right)\\
&= \sum_{t=1}^{n}\frac{\partial}{\partial\alpha}\left(\Psi(a_{1(t)}+a_{2(t)})-\Psi(a_{1(t)})-\Psi(a_{2(t)})\right)\\
&= \sum_{t=1}^{n}\left(\Psi'(a_{1(t)}+a_{2(t)})-\Psi'(a_{1(t)})-\Psi'(a_{2(t)})\right)\\
&= O_P(r_t),
\end{aligned}
$$

since in the stationary case we have that $a_{2(t)}$ is a function of r_t. □

We have now checked that under appropriate conditions, including (6.13), our MHL estimator is asymptotically equivalent to the ML estimator. Although we have shown it only for our simplest case, similar results will follow through for our other types of model, due to the way in which we define them. Indeed, if ${D_t}^{-1} = O_P(1/r_t)$ then we may easily modify our results to give asymptotic properties of the MHLEs for a HGLM with more than one x (random and unobserved) component. This essentially means when we start to consider more parties than just two. For this equation to extend generally, Lee and Nelder (1996) explain that we require the total number of random effects, n, to remain fixed. This will clearly be true when they are probabilities of voting per poll; each individual poll sample size r_t is independent of the n, which is also the total number of polls recorded.

[7] Here, $\Psi(u) = \mathrm{d}(\ln(\Gamma(u)))/\mathrm{d}u$ (digamma function) and $\Psi'(u) = \mathrm{d}^2(\ln(\Gamma(u)))/\mathrm{d}u^2$ (trigamma function) for some variable u.

Lee and Nelder (1996) discuss the instance when the total number of random effects increases with r_t; then, the asymptotic results are different from what we have shown above, although still quite similar.

6.6.3 Dispersion and interval estimates

Asymptotic results

Simply accepting our MHLEs at face value is not enough - it is important to examine the reliability of them, particularly because of our new method of estimation. We typically do this by calculating their standard errors, correlations and confidence intervals. With MLEs, $\hat{\theta}$, we have that, asymptotically,

$$\hat{\theta} \sim N_d(\theta, \mathbf{J_1}^{-\mathbf{1}}(\theta)),$$

where the j-k^{th} element of $\mathbf{J_1}^{-\mathbf{1}}(\theta)$ is either

$$-\mathbb{E}\left(\frac{\partial^2 l}{\partial\theta_j \partial\theta_k}\right) \qquad \text{or} \qquad -\frac{\partial^2 l}{\partial\theta_j \partial\theta_k}$$

and d is the dimension or number of parameters θ. We would use $\mathbf{J_1}^{-\mathbf{1}}(\hat{\theta})$ as an approximation to $\mathbf{J_1}^{-\mathbf{1}}(\theta)$, our covariance matrix, and from this may derive our correlation matrix.

Lee and Nelder (1996) extend this to MHLEs and so we have that consequently

$$\tilde{\theta} \sim N_d(\theta, \mathbf{J_2}^{-\mathbf{1}}(\theta)),$$

where the j-k^{th} element of $\mathbf{J_2}^{-\mathbf{1}}(\theta)$ is either

$$-\mathbb{E}\left(\frac{\partial^2 h}{\partial\theta_j \partial\theta_k}\right) \qquad \text{or} \qquad -\frac{\partial^2 h}{\partial\theta_j \partial\theta_k}.$$

Henceforth, a 99% confidence interval may be found using

$$\tilde{\theta}_j \pm 2.5758 \cdot s.e.(\tilde{\theta}_j),$$

in which the standard error $s.e.(\tilde{\theta}_j)$ is either

$$\sqrt{-\mathbb{E}\left(\frac{\partial^2 h}{\partial{\theta_j}^2}\right)} \qquad \text{or} \qquad \sqrt{\left(-\frac{\partial^2 h}{\partial{\theta_j}^2}\right)}.$$

For our analysis we will use the observed Fisher information matrix, that is, minus the Hessian matrix.

Reparameterisation

Note that in our non-Markov stochastic volatility models we reparameterised the lag weights, in order that MATLAB could search in the correct space (of between 0 and 1, whereas MATLAB typically performs a global search within the space $(-\infty, \infty)$).

Specifically with SV(2,2) and SV(3,2) we had

$$w = \frac{e^{v}}{1+e^{v}}$$

and with SV(2,3) and SV(3,3) we had

$$w_k = \frac{e^{v_k}}{1+e^{v_1}+e^{v_2}},$$

for $k = 1,\, 2$.

This means that we need to apply the **delta method** in order to work out the actual estimated variances and covariances. We briefly review the delta method, as applied to our case.

A necessary assumption is that the variances of v, v_1 and v_2 are small, which they are for our examples. Then for $k = 1,\, 2$ we have that

$$\begin{aligned}
\mathbb{V}ar(w) &\approx \left(\frac{\mathrm{d}w}{\mathrm{d}v}\right)^2 \mathbb{V}ar(v),\\
\mathbb{V}ar(w_k) &\approx \left(\frac{\partial w_k}{\partial v_1}\right)^2 \mathbb{V}ar(v_1) + \left(\frac{\partial w_k}{\partial v_2}\right)^2 \mathbb{V}ar(v_2) + 2\left(\frac{\partial w_k}{\partial v_1}\right)\left(\frac{\partial w_k}{\partial v_2}\right)\mathbb{C}ov(v_1,\, v_2),\\
\mathbb{C}ov(w_1,\, w_2) &\approx \left(\frac{\partial w_1}{\partial v_1}\right)\left(\frac{\partial w_2}{\partial v_1}\right)\mathbb{V}ar(v_1) + \left(\frac{\partial w_1}{\partial v_2}\right)\left(\frac{\partial w_2}{\partial v_2}\right)\mathbb{V}ar(v_2) +\\
&\quad \left[\left(\frac{\partial w_1}{\partial v_1}\right)\left(\frac{\partial w_2}{\partial v_2}\right) + \left(\frac{\partial w_1}{\partial v_2}\right)\left(\frac{\partial w_2}{\partial v_1}\right)\right]\mathbb{C}ov(v_1,\, v_2).
\end{aligned}$$

6.6.4 Comparison with the EM algorithm

ARCH(2,1): $\alpha = 500, \beta = 450$

Let us consider simulated data y_t of different lengths n. We first generate the pseudo data (with latent probabilities) and then fit this to ARCH(2,1), to obtain the MHLE vector $\tilde{\theta}$. Since the EM algorithm is manageable in this simplest case, we will obtain the corresponding MLE to enable comparison. Table 6.1 summarises the results.

$r=n$	mean($\hat{\theta}$)	mean($\tilde{\theta}$)	absolute error	relative error (%)
100	$\hat{\alpha}$=589.6 $\hat{\beta}$=502.7	$\tilde{\alpha}$=599.6 $\tilde{\beta}$=508.8	+10.0 +6.1	1.7 1.2
1500	$\hat{\alpha}$=547.8 $\hat{\beta}$=485.7	$\tilde{\alpha}$=553.1 $\tilde{\beta}$=489.8	+5.3 +4.1	1.0 0.8
5000	$\hat{\alpha}$=525.5 $\hat{\beta}$=472.3	$\tilde{\alpha}$=525.6 $\tilde{\beta}$=472.4	+0.1 +0.1	0.0 0.0
10000	$\hat{\alpha}$=516.8 $\hat{\beta}$=460.0	$\tilde{\alpha}$=516.8 $\tilde{\beta}$=460.0	+0.0 +0.0	0.0 0.0

Table 6.1: EM algorithm versus h-likelihood for simulations when $r = n$

Remarks: We may assert that, as the sample size increases with the number of votes (1500), the absolute errors of corresponding parameter estimates decrease. We can see that the h-likelihood MHLEs tend to overestimate the actual MLE values, but that the errors relative to the actual parameter values are reasonably small. In our practical terms, we only have a small number of polls, i.e. value for n, and so when we eventually obtain our MHLEs we would expect the MLEs to be smaller. Put another way, from the MHLE values we are able to estimate roughly what the MLE values are. We do not need to worry about slight discrepancies, particularly because we are interested in the proportions of the parameters, which will be almost identical for MHLEs and MLEs (to about three or four significant figures, which should be more than adequate for us). Absolute errors for α and β also become similar as n increases. It was shown in Section 6.6 that the two parameter vectors (MLE and MHLE) are asymptotically equivalent (in terms of r_t) and the results in Table 6.1 help to illustrate this. Recall equation (6.13), which implies that for our $r = 1500$ we would theoretically need about 1500 polls to have this equivalence, or for r to increase with n.

From looking at Table 6.1 we can see that when $r = n$ is smaller, say around 100, the MHLEs are less accurate than when $r = n$ gets larger. This is exactly what we expect, since equation (6.13) is an *asymptotic* result. The simulations of size 1500 by contrast is more accurate, albeit less so than those of size 5000, and 1500 is roughly the location of the number of votes collected for a typical poll. The results for $n = r = 5000$ help to confirm the necessary relationship (6.13).

Interestingly, we found that when we made $r = n$ much larger, the MHLEs started to underestimate (rather than always overestimate) the corresponding MLEs, even though these differences were negligible. We can see an instance of this behaviour in Table 6.1 when $r = n = 5000$, but it also happened commonly when we tried trials as far as sizes $r = n = 10000$, 20000 and 25000.

Clearly though, in practice there are far fewer polls than this conducted before the main election. More realistic is a value around 30 to 50. Lee and Nelder (1996) state that the procedures developed may be reliable and useful as an approximate inference even in the worst situations, such as here when $n \ll r_t$.

Small vector: In the modelling of election voting, a threat to the application of our overall method is the limited number of polls conducted before the election. This restriction is commonly known amongst those who have worked in election modelling and predicting, and may explain the reason for the absence of time series modelling, and prediction, of election data on the whole. Lewis-Beck (2005, [74]) states that this has precluded 'sophisticated time series work or elaborately specified equations.' Indeed, with smaller data vectors, say of length thirty or less, we find that our models mostly do not converge to give the MHLE vector. What happens by stark contrast is that we typically have divergence of the parameters as we iteratively seek the latent vectors which maximise the h-likelihood. This is certainly the case, for example, when applying our models to the general election of October 1974, for which there were only fifteen polls recorded and suitable for use before the final outcome. Our specific models, nevertheless, perform as required provided that we enlarge the vector to thirty or forty observations, say. Hence we need a sufficiently-large vector of poll recordings for our programs to get up and running.

To relieve this, one approach might be to redefine what data we make use of. Instead of only using final poll counts, we might also include midway poll counts. This would effectively double the number of recordings available for our modelling. Obtaining this, quite possibly unpublished, data may not be easy though, and even if we had it there may be implications on their quality and therefore reliability. For instance, the organisations taking the polls may perform certain data cleansing on the final poll recording which they publish. We would need to assess this carefully before proceeding in this way.

6.6.5 Typical simulations

Figures 6.4-6.9 show time-series plots of simulated data for each type of model, ARCH, stochastic volatility and GARCH, respectively, each showing progression which is typical of the model behaviour. The first three are for two parties and one lag and the second three for three parties and two lags.

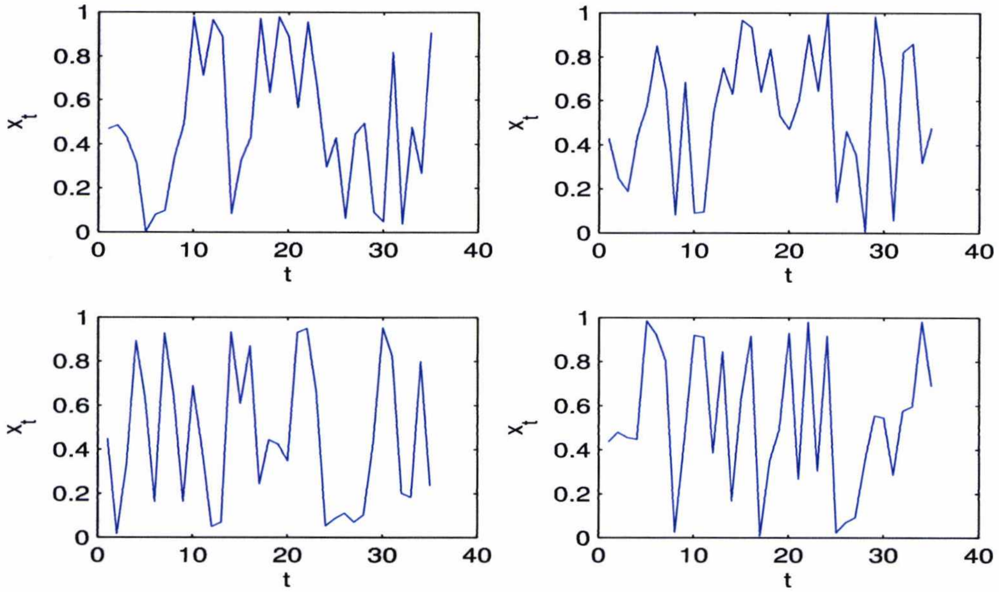


Figure 6.4: Simulations of x_t in ARCH(2,1) where $r_t = 1500, \alpha = 250, \beta = 290, n = 35$.

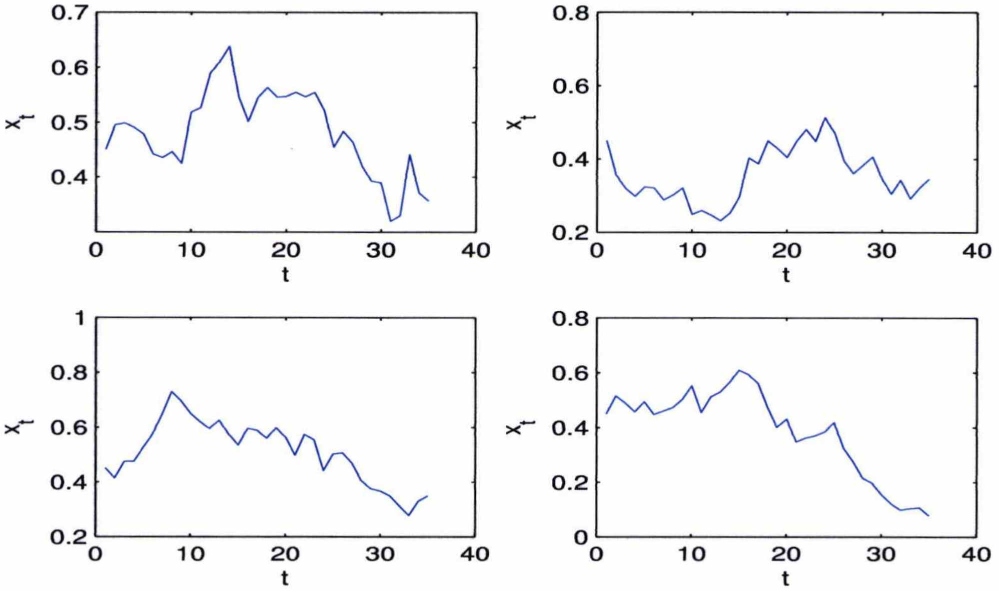


Figure 6.5: Simulations of x_t in SV(2,1) where $r_t = 1500, \phi = 0.01, n = 35, x_1 = 0.45$.

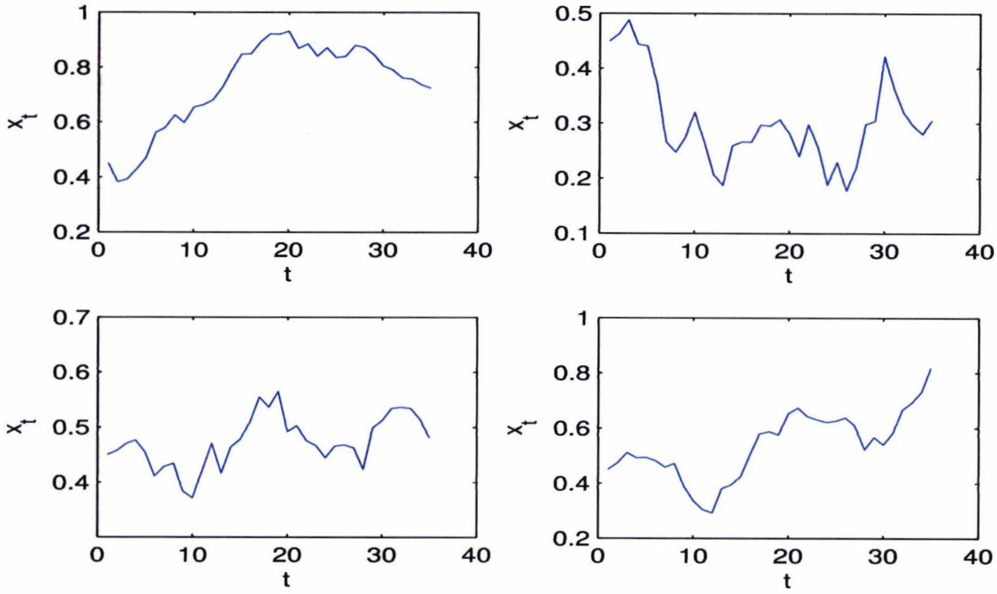


Figure 6.6: Simulations of x_t in GARCH(2,1) where $r_t = 1500, \phi = 0.01, u = 0.9, n = 35, x_1 = 0.45$.

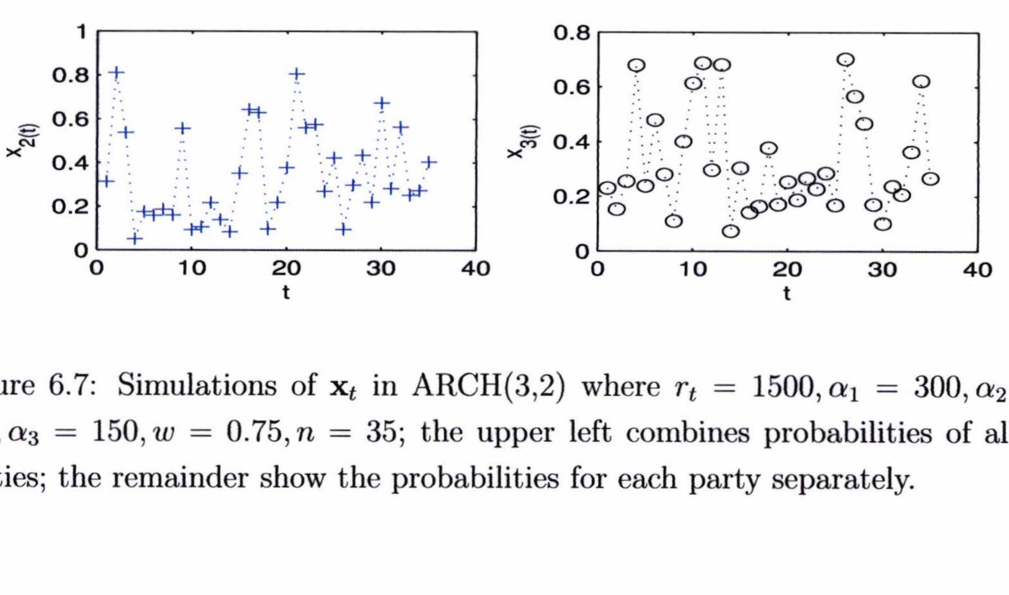


Figure 6.7: Simulations of $\mathbf{x}_t$ in ARCH(3,2) where $r_t = 1500, \alpha_1 = 300, \alpha_2 = 200, \alpha_3 = 150, w = 0.75, n = 35$; the upper left combines probabilities of all 3 parties; the remainder show the probabilities for each party separately.

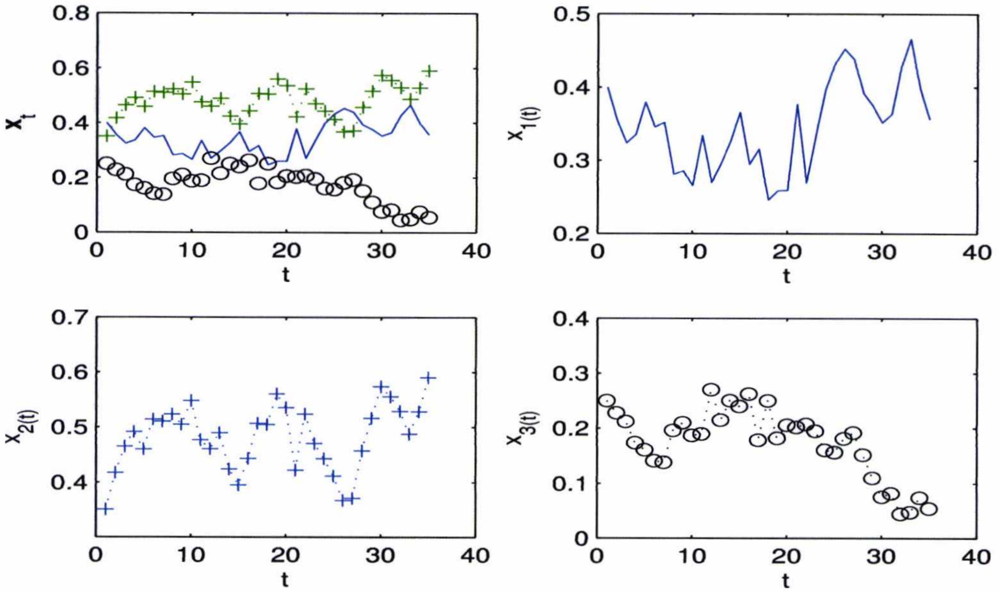


Figure 6.8: Simulations of $\mathbf{x}_t$ in SV(3,2) where $r_t = 1500, \phi = 0.01, n = 35, w = 0.75, x_{1(1)} = 0.40, x_{2(1)} = 0.35$; the upper left combines probabilities of all 3 parties; the remainder show the probabilities for each party separately.

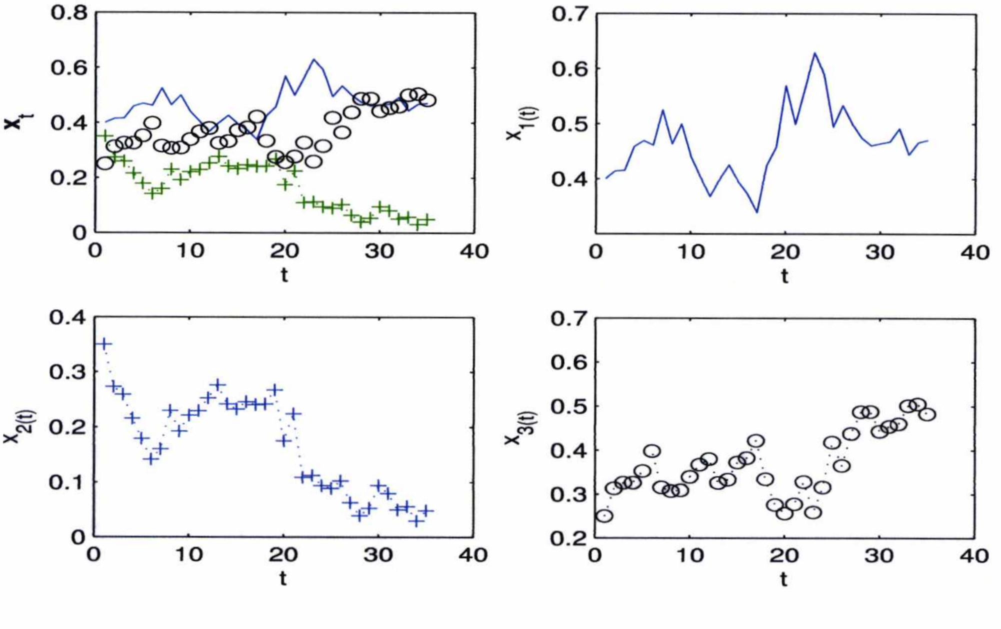


Figure 6.9: Simulations of $\mathbf{x}_t$ in GARCH(3,2) where $r_t = 1500, \phi = 0.01, n = 35, u = 0.6, w = 0.75, x_{1(1)} = 0.40, x_{2(1)} = 0.35$; the upper left combines probabilities of all 3 parties; the remainder show the probabilities for each party separately.

Interpretation

If we think of polls and the probability of voting for a party, then we would generally not expect the probability to vary widely over a short period of time. We, therefore, expect the time series plot to be quite smooth over a shorter period of time, but possibly to change a bit, though still quite modestly, over a longer period of time. We must use this criterion when studying the types of simulations we get per model.

It is clear from the ARCH plots that the time series is highly variable, effectively implying that over a short period of time the probabilities vary quite noticeably. This is unrealistic, but we should remember that this is only a stationary model. However, the effect of varying the lags is unclear from the graphs; we will need to look at alternative ways to assess relative goodness of fit. The same is in fact true for all our models.

By contrast, the SV and GARCH plots are much smoother as required, in that they vary slightly but in general do not steer off too widely in the given time period. It is not easy to differentiate between the SV and GARCH plots. This is because when we change our weight u the impact is not obvious, which suggests possibly dropping one of these types, and perhaps the GARCH models since the SV models are more parsimonious.

These figures also stimulate our intuition in terms of goodness of fit when in the context of election poll data (see Section 7.4.3).

Choice of values

- n: we set this at 35 days, which is a reasonable length of time during which in practice a set of polls is conducted.

- r_t: we fix this to be 1500 for ease of comparison. This value is roughly the sample size taken for each poll.

- α and β: we consider $\alpha > \beta$ as well as $\alpha < \beta$ for typical values which we actually obtained from maximising using our method.

- ϕ: we use 0.01 which to two decimal places is the value which we tend to get each time when maximising using our method.

- δ_t: we fix this to be 1 merely in order to compare with the ARCH models. Note that this gives $c_t \approx 99.5$.

- u: we vary this to study the dynamics of whether the previous probabilities or observed data dominate.
- w_1 and w_2: we vary these to study the effect of letting different lags dominate.

6.6.6 Simulations: volatility models

Recall 6.5.3 earlier. The estimation method for the stochastic volatility and GARCH models differed slightly from that presented in Lee and Nelder (1996), in terms of how we maximise the latent data. We have no theory justifying the asymptotic efficiency of the parameters and latent data in these cases. Therefore, it is necessary to check that the MHLE parameters are reliable for use. One way to do this is to simulate data from chosen values and subsequently to find the MHLE of these data, and so these are given in Tables 6.2 for stochastic volatility models and 6.3 for GARCH models. We repeated this fifty times for each model and obtained the mean of $\tilde{\theta}$. Ideally, we want the mean MHLEs to be as close to the chosen values as possible. Note that we chose $x_1 = 0.45$ for the two-party models and $x_{1(1)} = 0.40$, $x_{2(1)} = 0.35$ for the three-party models. From looking at these tables, we can see that the mean MHLEs obtained are reasonably close to the original choices of parameter values. This empirical evidence suggests that the slight adaptions to the estimation methods outlined in 6.5.3 are reliable enough to use.

6.7 Goodness of Fit

We are interested in two aspects concerning the goodness of fit of our models to the data, that is, both *relative* and *absolute*. Of these, the former is probably the easier to discuss and carry out, in which we compare all our models for a given number of parties in order to determine which fits best to the given dataset. Therefore, what we end up comparing are the broad type of model (stationary or stochastic volatility) as well as the number of lags per model. We would perhaps expect the stochastic volatility models to outperform the stationary ones, as they involve more information than the stationary models, namely the difference in time between polls. Recall how for all our models we obtained both fixed effect and random effect estimates. Our method of assessing relative goodness of fit is via the deviance method.

Model	$\theta^{(0)}$	mean MHLE (to 3 decimal places)
SV(2,1)	ϕ=0.01	ϕ=0.009
SV(2,2)	ϕ=0.01 w=0.5	ϕ=0.009 w=0.416
SV(2,2)	ϕ=0.01 w=0.9	ϕ=0.008 w=0.756
SV(2,2)	ϕ=0.01 w=0.1	ϕ=0.007 w=0.091
SV(2,3)	ϕ=0.01 w_1=0.6 w_2=0.3	ϕ=0.009 w_1=0.523 w_2=0.325
SV(2,3)	ϕ=0.01 w_1=0.2 w_2=0.3	ϕ=0.008 w_1=0.195 w_2=0.326
SV(3,1)	ϕ=0.01	ϕ=0.008
SV(3,2)	ϕ=0.01 w=0.75	ϕ=0.009 w=0.752
SV(3,3)	ϕ=0.01 w_1=0.7 w_2=0.2	ϕ=0.009 w_1=0.685 w_2=0.100

Table 6.2: H-likelihood method with stochastic volatility models when fit to simulated data.

6.7.1 Relative goodness of fit

The Scaled Deviance Test

In Chapter 3, we mentioned model criteria which may be applied to time series models, whether nested or not, namely the Aikaike Information Criterion (AIC):

$$AIC = 2d - 2l,$$

where d is the number of parameters and Bayesian Information Criterion (BIC):

$$BIC = \ln(n)d - 2l,$$

where n is the number of observations (in our case the number of polls recorded). These measure the amount of information lost when the model is used to describe reality, and attempt to find the model which best explains the data with a minimum number of free parameters. Both reward goodness of fit while including a penality which is a linear increasing function of the number of estimated parameters, thus discouraging overfitting. They effectively judge by how close the fitted values of the model tend to be to the true values, in terms of expected values. The basic idea is to choose the model with the lowest value therefore, where the BIC imposes a greater penality for less parsimony via the $\ln(n)$ (typically greater than 2 in the AIC) term. As with estimation, due to the presence of latent variables we cannot use these criteria. To relieve this problem, Lee and Nelder (1996) introduced a similar statistic to the above, known as the scaled deviance, to select the best hierarchical generalised linear model (HGLM). This is based on writing

$$S = S(y,\, \hat{\mu}') = 2\{l(y;\, y|\, x) - l(\hat{\mu}';\, y|\, x)\}, \tag{6.14}$$

in which $\mu' = \mathbb{E}(y|\, x)$, $l(\hat{\mu}';\, y|\, x) = \ln(f(y|\, x; \tilde{\beta}))$ and $\tilde{\beta}$ are the estimated fixed effect parameters, all of whose terminology we introduced in Chapter 5 when we described HGLMs. Then, for models with random parameters we may use the statistic

$$D = 2(n - \mathbb{E}(S) - \ln f_\theta(y|\, x)).$$

However, in our binomial/multinomial models for $y|\, x$ there are no fixed effect parameters present; our parameters only appear in the $x|\, y$ part. This simplifies matters for us somewhat, since $S = 0$ and we are left with

$$D = 2(n - \ln f_\theta(y|\, x)),$$

which we call the **deviance**.

Note how the x are now treated as if they were known values and effectively left out of further analysis, as their task of probability generation in order to find the MHLEs is now over.

The idea is that we choose the model with *smallest* deviance value.

6.7.2 Absolute goodness of fit

Using the deviance criterion, we can compare *within* models (ARCH(2,1) and ARCH(2,2) for instance) but we cannot compare *between* models (i.e. ARCH and volatility). Hong and Preston (2008, [58]), for instance, consider weak conditions under which consistent model selection is possible, regardless of whether the models are nested or not. However, it is not known whether we may apply this to our scenario, in which there are *latent variables*. We therefore rely on simulation plots using MHLEs obtained in order to determine which of ARCH and SV is better for poll modelling.

Assessing relative goodness of fit does not immediately mean that the models actually suit the data well. One useful and simple way to assess whether this is so is to look at typical time series simulations of our chosen models. We would expect the plots generally to take the shape of the actual proportions obtained from the polls in the corresponding pre-elections.

Hypothesis Tests

Lee and Nelder (1996) present a test very similar to the Likelihood Ratio test, which is suited to cover HGLMs like ours. For a two-tailed test we have the hypotheses

$$H_0 : \theta = \boldsymbol{\theta_0}$$
$$H_1 : \theta \neq \boldsymbol{\theta_0}$$

and the test statistic

$$\lambda = 2(\tilde{h} - h_0),$$

in which $\tilde{h}$ is the maximum hierarchical likelihood value obtained by using our MHLEs and h_0 is the hierarchical likelihood value using $\boldsymbol{\theta_0}$.

Also,

$$\lambda \sim \chi^2,$$

which holds due to the asymptotic equivalence of the MLE and MHLE vectors:

$$|\,\hat{\theta} - \tilde{\theta}\,| = O_p\left(\frac{1}{r_t}\right),$$

seen earlier in the chapter and also more generally in Chapter 5; that is, because $\hat{\theta}$ is asymptotically normal then so is $\tilde{\theta}$.

We proceed by rejecting the null at $2\alpha\%$ if

$$\lambda > {\chi^2}_d(\alpha).$$

6.8 Review

This chapter has detailed the theory behind our modelling of the opinion polls. We looked at various aspects, beginning with the general model specification. We fit a Dirichlet-multinomial model to the data, in which the votes (observations) are multinomially distributed and probabilities of voting (latent variables) are Dirichlet distributed. Having explained this carefully, we then defined each specific model which we are interested in fitting. There are three types: ARCH, which is a stationary model, and stochastic volatility and GARCH, both of which are nonstationary. For each broad type, we have a two-party and three-party case, the latter of which may be easily extended to any number of parties; for each we can consider a different number of time lags, which we limit to three in this thesis. We gave their basic properties and then expressed the h-likelihood function for each model. Next, we discussed the estimation by h-likelihood, specifying how the theory from the previous chapter is applied in this time series context. Importantly, we verified that the properties of estimators proved in Chapter 5 follow through for us, for the ARCH models, which boils down to requiring that the sample size of polls must increase with the number of polls conducted. Further theory was given, outlining how we can perform inference, having estimated the h-likelihood parameters. This included an approximate covariance matrix and thus approximate confidence intervals. We discussed the model selection criterion available in the presence of latent variables, which is known as the deviance method. Having chosen the best model, we then need to check that it suits the data well, for which we have hypothesis tests as well as simulation plots given the estimated parameters. We introduced how the EM algorithm may be applied for the simplest ARCH model, for the integrals conveniently end up quite simple; this made it possible to compare the performance of the h-likelihood with the EM algorithm, which in turn showed that the h-likelihood method is reliable enough to use. Finally, since we have no comparative theory for our volatility models, we studied simulations and found the MHLEs of them, which proved promising as the values were similar to the starting values which we chose to simulate using initially.

Model	$\theta^{(0)}$	mean MHLE (to 3 decimal places)
GARCH(2,1)	ϕ=0.01 u=0.5	ϕ=0.007 u=0.512
GARCH(2,1)	ϕ=0.01 u=0.9	ϕ=0.008 u=0.856
GARCH(2,1)	ϕ=0.01 u=0.5	ϕ=0.009 u=0.500
GARCH(2,2)	ϕ=0.01 w=0.5 u=0.5	ϕ=0.008 w=0.501 u=0.455
GARCH(2,2)	ϕ=0.01 w=0.5 u=0.5	ϕ=0.009 w=0.476 u=0.602
GARCH(2,2)	ϕ=0.01 w=0.5 u=0.1	ϕ=0.009 w=0.515 u=0.096
GARCH(2,2)	ϕ=0.01 w=0.75 u=0.75	ϕ=0.009 w=0.751 u=0.682
GARCH(2,2)	ϕ=0.01 w=0.25 u=0.75	ϕ=0.008 w=0.242 u=0.672
GARCH(2,3)	ϕ=0.01 w_1=0.6 w_2=0.3 u=0.75	ϕ=0.007 w_1=0.588 w_2=0.273 u=0.512
GARCH(2,3)	ϕ=0.01 w_1=0.6 w_2=0.3 u=0.25	ϕ=0.008 w_1=0.575 w_2=0.225 u=0.200
GARCH(2,3)	ϕ=0.01 w_1=0.2 w_2=0.3 u=0.75	ϕ=0.010 w_1=0.202 w_2=0.228 u=0.787
GARCH(2,3)	ϕ=0.01 w_1=0.2 w_2=0.3 u=0.25	ϕ=0.008 w_1=0.225 w_2=0.276 u=0.453
GARCH(3,1)	ϕ=0.01 u=0.6	ϕ=0.009 u=0.666
GARCH(3,2)	ϕ=0.01 w=0.6 u=0.75	ϕ=0.008 w=0.587 u=0.728
GARCH(3,3)	ϕ=0.01 w_1=0.6 w_2=0.3 u=0.20	ϕ=0.009 w_1=0.589 w_2=0.221 u=0.256

Table 6.3: H-likelihood method with GARCH models when fit to simulated data.

Chapter 7

Illustrations

7.1 Introduction

Now that we have specified a range of potential models, we will demonstrate how they work and assess how useful they are. Our aim is to select the best model out of those considered, and then to see how useful it is in reflecting the behaviour of the time series. Subsequently, we want to apply our methodology to the actual well-established forecasting method adopted by the British Broadcasting Corporation (BBC) on election night.

We consider two datasets, which we introduce in Section 7.2. Section 7.3 is devoted to the first of these and Section 7.4 to the second. In each section, first we fit our models to the data to find our parameters. After a discussion of their meaning we provide approximate correlation matrices for them, which are used to give approximate confidence intervals. We then need to select the best model, via the scaled deviance test developed in Lee and Nelder (1996, [69]). Following this, we concentrate on our models of choice, examining their specific properties. Finally, Section 7.5 provides a comparison of statistics from our method with that from existing methods.

7.2 Real Data

7.2.1 Components

To work in conjunction with the fact that our model is time series, we require data recordings as time progresses. We choose a starting time, which is likely to be the point at which fieldwork begins for the first opinion poll recorded for an upcoming election. We set this to zero, such that the fieldwork finishes for that poll at time δ_1. We then work in days where, for instance, $3\frac{1}{2}$ days after the first set of recordings will for us translate into $\delta = 3.5$. We then include all known recording time periods as far as we can; we define the latest poll as 'current time', n. Note that we will look at stationary models, which disregard the differences in time between poll collections, in addition to the nonstationary ones which do account for time variation.

At each collection of recordings, we will have the distribution of total votes for each contending election party. Of course, we choose only to focus on specific parties, and so we can either pool the remainder or ignore them altogether. The actual recordings themselves will be the total number of votes for that party at that particular poll. Hence, for example, 425 for Labour at 16.5 means that $16\frac{1}{2}$ days after the first set of recordings, a new set of poll recordings was taken, and then 425 people voted Labour.

It is clearly a good idea to convert the nominal data into proportions in order to gauge relative strengths at those separate times. This would be useful as it is ultimately relative strength - given the data as a whole - with which we are concerned. Also, obviously the total number of votes differs amongst polls, so the data are of less use in absolute terms.

We will be dealing with three parties only (Labour, Conservative and Liberal/Alliance), and so will simply disregard information on the remainder, redefine the total number of votes, and rescale the proportions. As we have said, we can of course include however many number of parties (and lags) we like, though this unnecessarily affects the clarity and simplicity of our computation and subsequent analysis.

7.2.2 Elections

We will focus on two historical general elections: 1983 and 1987. The poll data for these two are shown at Appendix C and plots of proportions of votes for each party are shown in Figures 7.1 and 7.2 for two-parties and Figures 7.3 and 7.4 for three parties. The reasons for choosing these elections are both that more poll recordings are available here and they are both three-party contests, similar to what we generally expect in the UK nowadays. Clearly from the plots, we see that the Conservative party has a striking lead in both pre-elections, Labour is almost always second and Liberal/Alliance mostly third.

As we have stressed, our main objective is to use these data in order to find maximum hierarchical likelihood estimates (MHLEs).

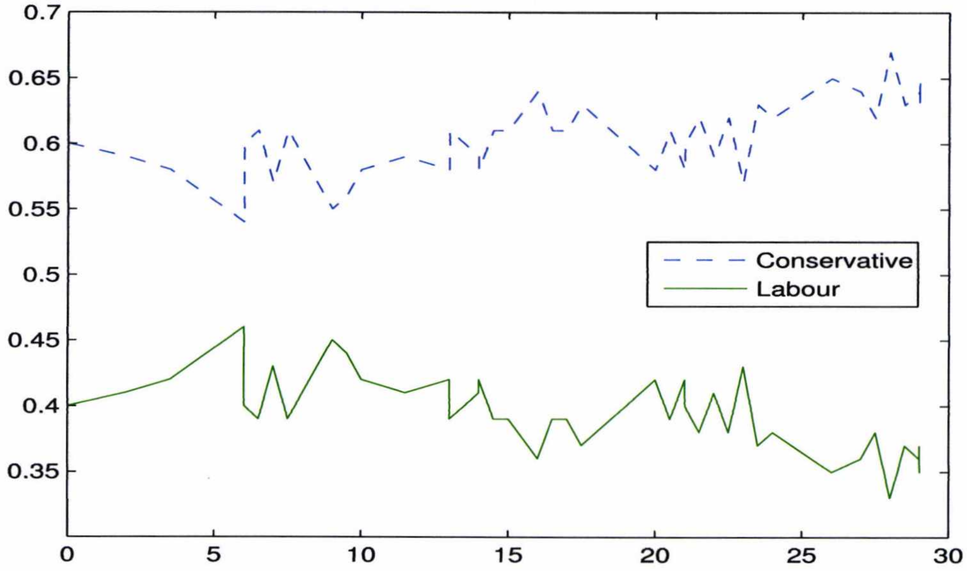


Figure 7.1: Poll outcomes for the main two parties in the run-up to the 1983 general election (time in days on x-axis, proportions of votes on y-axis).

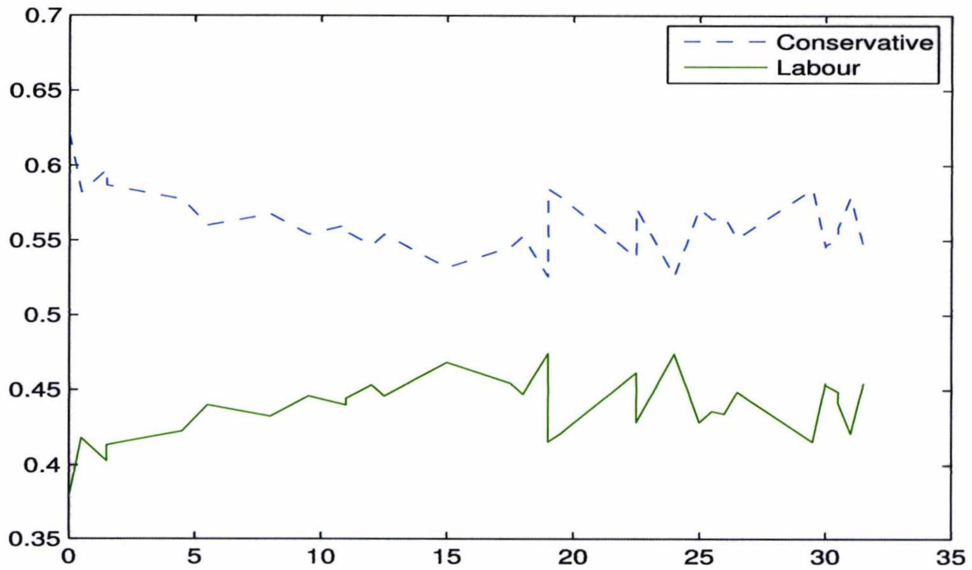


Figure 7.2: Poll outcomes for the main two parties in the run-up to the 1987 general election (time in days on x-axis, proportions of votes on y-axis).

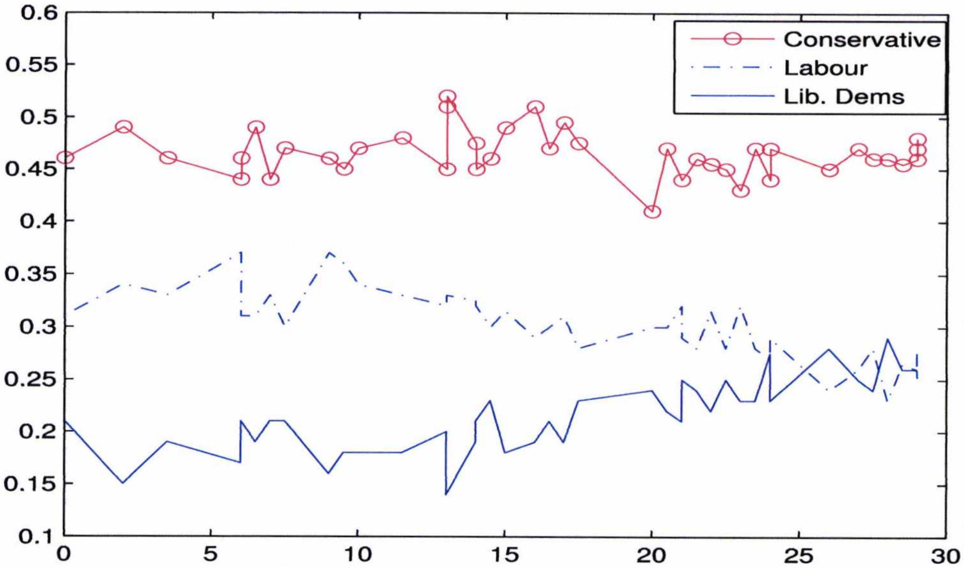


Figure 7.3: Poll outcomes for the main (three) parties in the run-up to the 1983 general election (time in days on x-axis, proportions of votes on y-axis); 'Lib. Dems'=Liberal.

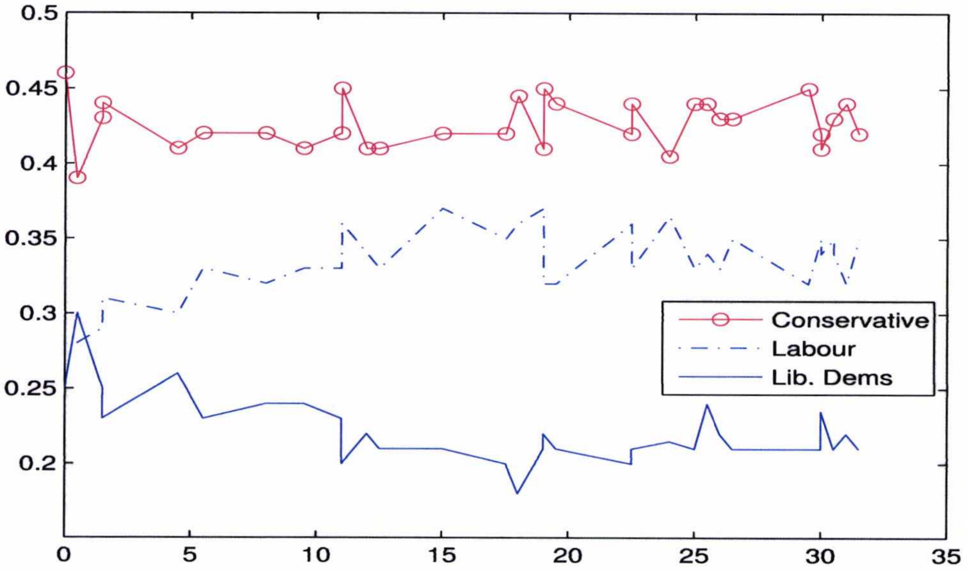


Figure 7.4: Poll outcomes for the main (three) parties in the run-up to the 1987 general election (time in days on x-axis, proportions of votes on y-axis); 'Lib. Dems'=Alliance.

7.3 Illustration 1: 1983 Data

7.3.1 MHLEs

Table 7.1 shows the MHLE values obtained for all models when fit to the 1983 data. Also shown are the values for $\tilde{h}$, which we may interpret in the same way that we would with log-likelihood values, that is, the MHLEs found give us the maximum absolute value of h.

Model	$\tilde{\theta}'$	$\tilde{h}$
ARCH(2,1)	$(\tilde{\alpha} = 1271.7,\ \tilde{\beta} = 826.8)$	-2160.8
ARCH(2,2)	$(\tilde{\alpha} = 590.4,\ \tilde{\beta} = 383.8,\ \tilde{w} = 0.6250)$	-1947.2
ARCH(2,3)	$(\tilde{\alpha} = 530.5,\ \tilde{\beta} = 345.0,\ \tilde{w}_1 = 0.5250,\ \tilde{w}_2 = 0.1250)$	-1758.6
SV(2,1)	$(\tilde{\phi} = 0.4030)$	-161.6
SV(2,2)	$(\tilde{\phi} = 0.2797,\ \tilde{w} = 0.4626)$	-155.3
SV(2,3)	$(\tilde{\phi} = 0.1176,\ \tilde{w}_1 = 0.1863,\ \tilde{w}_2 = 0.1051)$	-143.4
ARCH(3,1)	$(\tilde{\alpha_1} = 1882.7,\ \tilde{\alpha_2} = 1228.1,\ \tilde{\alpha_3} = 941.2)$	-3338.7
ARCH(3,2)	$(\tilde{\alpha_1} = 879.4,\ \tilde{\alpha_2} = 573.6,\ \tilde{\alpha_3} = 439.6,\ \tilde{w} = 0.5750)$	-3043.6
ARCH(3,3)	$(\tilde{\alpha_1} = 634.6,\ \tilde{\alpha_2} = 413.9,\ \tilde{\alpha_3} = 317.3,\ \tilde{w}_1 = 0.4250,\ \tilde{w}_2 = 0.1250)$	-2640.0
SV(3,1)	$(\tilde{\phi} = 0.1978)$	-281.8
SV(3,2)	$(\tilde{\phi} = 0.1303,\ \tilde{w} = 0.4402)$	-272.4
SV(3,3)	$(\tilde{\phi}=0.0548,\ \tilde{w}_1 = 0.1738,\ \tilde{w}_2 = 0.0408)$	-252.3

Table 7.1: MHLEs for candidate models when fit to the 1983 poll data.

Inference

ARCH models: We see that the party strength parameters α are large in magnitude, which is due to the large figures for the number of votes in the polls. This does not matter since we are effectively interested in the relative strength of the parties and so we work with the proportions $\alpha_g / \sum_{k=1}^{p} \alpha_k$ for party g, where p is the number of parties. Then, the outputs are similar regardless of the model. For the 1983 data and in a two-party contest, all three models suggest a distribution of about 60 per cent in favour of Conservative and the remaining 40 per cent for Labour. In a three-party contest, all three suggest about 45 per cent in favour of Conservative, 30 per cent for Labour and the remaining 25 per cent for Liberal. These figures reflect what we saw in Figures 7.1 and 7.3, showing the poll outcomes by proportions, to suggest that MHLEs α are very reliable for further use; rather than relying on graphs we now have determined values to use.

The lag parameters w vary, depending on the model. First consider two parties. ARCH(2,2) attributes 63 per cent of the information to the previous lag and the remainder to the second lag. ARCH(2,3) also attributes more to lag one, but only 53 per cent now; it attributes 13 per cent to lag two, but the remaining 44 per cent to lag three. With three parties, ARCH(3,2) like ARCH(2,2) attributes more to lag one (58 per cent), whereas ARCH(3,3) like ARCH(3,2) reduces this (this time to 43 per cent and less than half). This time, ARCH(3,3) actually attributes most (44 per cent) to lag three. Obviously it would be useful to extend our models to more lags in order to assess this further. It seems for the above as if the models never put much weight on the second lag.

Stochastic volatility models: Recall that

$$c_{s(t)} = \frac{\exp(-\phi\delta_{s(t)})}{1 - \exp(-\phi\delta_{s(t)})},$$

where s is the s^{th} lag, is a value which the proportionate strengths of the parties divide amongst their corresponding shape parameters inside the beta/Dirichlet density. Strictly speaking, it would be the latent probabilities which distribute the $c_{s(t)}$, but we decided to replace them with the proportion of votes per party. Therefore, the larger $c_{s(t)}$ is the more there is to share out. $c_{s(t)}$ is governed in size by both ϕ and $\delta_{s(t)}$: the smaller their product the larger $c_{s(t)}$ is. For all of our models here we find that our value of ϕ is very small, roughly somewhere between 0.01 and 0.1. Typically our values for $\delta_{s(t)}$ will also be small; considering both datasets the maximum number of days between polls was 3. This implies that our value for $c_{s(t)}$ is reasonably large, say 100, and so we have very different probability densities at each poll. (When we actually go on to consider the election night, seats often declare consecutively in very small intervals.) Note also that recent lags are more influential than older lags since $\delta_{s(t)} < \delta_{s(t+1)}$ which leads to $c_{1(t)} > c_{2(t)} > c_{3(t)}$ for some ϕ and poll t.

For the 1983 data and with two parties SV(2,2) attributes only 46 per cent to the first lag yet 54 to the second. This contrasts with the comparative ARCH models, which put more emphasis on the first lag than the second. Similar is true in the three-party case. As with the ARCH models, when we go on to consider three parties for the 1983 data all stochastic volatility models attribute a majority (about 80 per cent) of weight on the third lag, a reasonable amount on the first, and least on the second.

7.3.2 Dispersion and interval estimates

Now we present the approximate correlation matrices for the MHLEs, as well as their approximate 99% confidence intervals.

ARCH(2,1)

$$\begin{pmatrix} 43.2994 & 0.9778 \\ 0.9778 & 37.5054 \end{pmatrix}$$

$$\alpha \in (1160.2, 1383.2)$$

$$\beta \in (730.2, 923.4)$$

ARCH(2,2)

$$\begin{pmatrix} 23.9971 & 0.9549 & 0 \\ 0.9549 & 20.7131 & 0 \\ 0 & 0 & 0.3266 \end{pmatrix}$$

$$\alpha \in (447.6, 571.2)$$

$$\beta \in (330.4, 437.2)$$

$$w \in (0, 1)$$

ARCH(2,3)

$$\begin{pmatrix} 19.3702 & 0.9349 & 0 & 0 \\ 0.9349 & 16.6401 & 0 & 0 \\ 0 & 0 & 0.0790 & -0.3974 \\ 0 & 0 & -0.3974 & 0.0523 \end{pmatrix}$$

$$\alpha \in (480.6, 580.4)$$

$$\beta \in (302.1, 387.9)$$

$$w_1 \in (0.3215, 0.7285)$$

$$w_2 \in (0, 0.2597)$$

SV(2,1)

$$(0.0641)$$

$$\phi \in (0.2380, 0.5681)$$

SV(2,2)

$$\begin{pmatrix} 0.0507 & 0.4157 \\ 0.4157 & 0.0173 \end{pmatrix}$$

$$\phi \in (0.1492, 0.4102)$$

$$w \in (0.1037, 0.7814)$$

SV(2,3)

$$\begin{pmatrix} 0.0241 & 0.3732 & 0.0300 \\ 0.3732 & 0.0664 & -0.5519 \\ 0.0300 & -0.5519 & 0.1221 \end{pmatrix}$$

$$\phi \in (0.0554, 0.1797)$$

$$w_1 \in (0.0153, 0.3573)$$

$$w_2 \in (0, 0.0658)$$

ARCH(3,1)

$$\begin{pmatrix} 60.1633 & 0.9660 & 0.9776 \\ 0.9660 & 39.6119 & 0.9705 \\ 0.9776 & 0.9705 & 25.9207 \end{pmatrix}$$

$$\alpha_1 \in (1727.7, 2037.7)$$

$$\alpha_2 \in (1126.1, 1330.1)$$

$$\alpha_3 \in (874.4, 1008.0)$$

ARCH(3,2)

$$\begin{pmatrix} 32.7110 & 0.9312 & 0.9540 & 0 \\ 0.9312 & 21.6909 & 0.9401 & 0 \\ 0.9540 & 0.9401 & 14.2935 & 0 \\ 0 & 0 & 0 & 0.3198 \end{pmatrix}$$

$\alpha_1 \in (795.1, 963.7)$

$\alpha_2 \in (517.7, 629.5)$

$\alpha_3 \in (402.7, 476.3)$

$w \in (0, 1)$

ARCH(3,3)

$$\begin{pmatrix} 21.8766 & 0.8819 & 0.9186 & 0 & 0 \\ 0.8819 & 15.2163 & 0.8955 & 0 & 0 \\ 0.9186 & 0.8955 & 10.0830 & 0 & 0 \\ 0 & 0 & 0 & 0.0782 & -0.3249 \\ 0 & 0 & 0 & -0.3249 & 0.0523 \end{pmatrix}$$

$\alpha_1 \in (578.3, 690.9)$

$\alpha_2 \in (374.7, 453.1)$

$\alpha_3 \in (291.3, 343.3)$

$w_1 \in (0.2236, 0.6264)$

$w_2 \in (0, 0.2597)$

SV(3,1)

$$(0.0254)$$

$\phi \in (0.1326, 0.2631)$

SV(3,2)

$$\begin{pmatrix} 0.0183 & 0.4180 \\ 0.4180 & 0.3959 \end{pmatrix}$$

$$\phi \in (0.0831, 0.1774)$$

$$w \in (0.1889, 0.6914)$$

SV(3,3)

$$\begin{pmatrix} 0.0087 & 0.3954 & 0.0185 \\ 0.3954 & 0.0489 & -0.4872 \\ 0.0185 & -0.4872 & 0.0881 \end{pmatrix}$$

$$\phi \in (0.0325, 0.0772)$$

$$w_1 \in (0.0478, 0.2997)$$

$$w_2 \in (0, 0.1862)$$

Inference

ARCH models: The standard errors for the party parameters are quite large, but not relative to the parameter sizes themselves. There is strong positive correlation between them, but this is to be expected, because if the poll sample size increases, then we would expect not only more people to vote one party but all parties. Due to the way in which the mixture transition density (MTD) models are constructed, these parameters are independent from the lag parameters. The lag parameters are negatively correlated, which also makes sense since they all sum to one, and so increasing one would mean decreasing the others. In all instances above, the confidence intervals tell us (with 99 per cent certainty) that overall during the poll period, Conservative was more powerful than Labour which in turn was more powerful than Liberal/Alliance. When the model only has one weight parameter the confidence intervals are not very useful, at least at the 99 per cent level. They are more useful when there are several weight parameters; for instance, for the 1983 data ARCH(2,3) says with 99 per cent certainty that lag one accounts for more of the modelling than lag two.

Stochastic volatility models: The standard errors are suitably low for all parameters, which will give us better confidence intervals. Again, the weight parameters are negatively correlated as would be expected.

7.3.3 Relative goodness of fit

Table 7.2 shows the deviance values for each model as well as the number of parameters, d.

Model	d	D
ARCH(2,1)	2	3791.5
ARCH(2,2)	3	3085.7
ARCH(2,3)	4	2761.3
SV(2,1)	1	389.8
SV(2,2)	2	387.1
SV(2,3)	3	387.1
ARCH(3,1)	3	6080.9
ARCH(3,2)	4	5018.9
ARCH(3,3)	5	4231.6
SV(3,1)	1	711.2
SV(3,2)	2	710.5
SV(3,3)	3	710.6

Table 7.2: Deviances for candidate models when fit to the 1983 poll data.

Inference

For the 1983 data, the deviance method selects SV(2,3) when there are two parties and SV(3,2) when there are three parties.

7.3.4 Absolute goodness of fit

We have chosen the best models out of two different sets (ARCH and SV) of nested models, that is, of a similar form to one another but with slightly different parameterisations; this is relative goodness of fit. However, this does not immediately mean that the models actually suit the data well. One useful and simple way to assess whether this is so is to look at typical time series simulations of our chosen models. We would expect the plots generally to take the shape of the actual proportions obtained from the polls in the corresponding pre-elections.

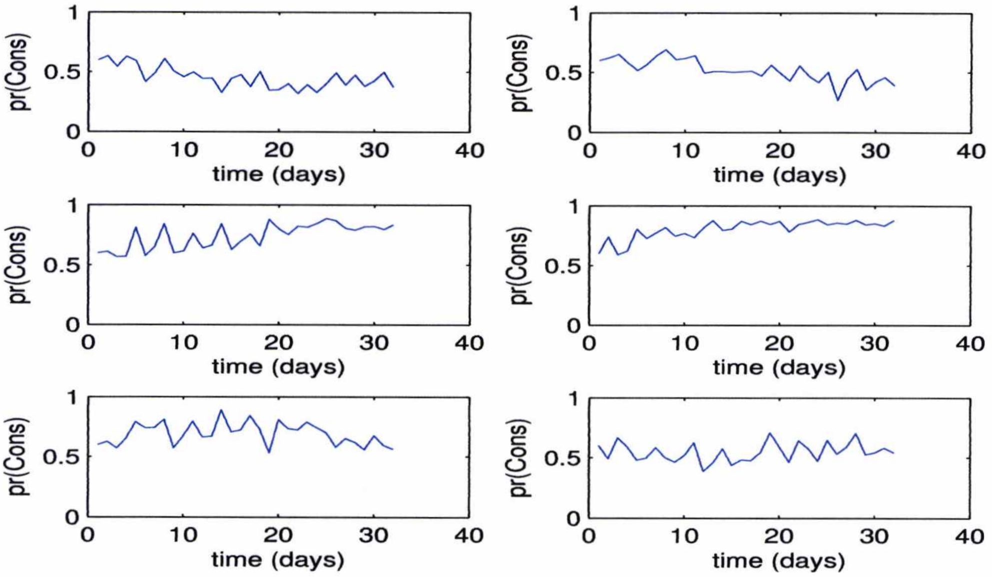


Figure 7.5: Simulations of probabilities of voting from SV(2,3) using $\tilde{\phi} = 0.12$, $\tilde{w}_1 = 0.19$ and $\tilde{w}_2 = 0.11$, obtained from the 1983 data.

Simulations

Recall that we saw typical simulations of our models in the previous chapter, but now we update them to contain our MHLE parameters. Figure 7.5 shows typical simulations obtained using the MHLEs found from fitting the 1983 poll data to the best model (SV(2,3)), in a two-party contest of Conservative versus Labour. Figure 7.6 shows the best model in a three-party contest, which was SV(3,2), with their unique MHLEs when fit to the 1983 poll data.

We firstly note that all plots are clearly smoother than those obtained from the spiky ARCH models. The spikes here are by comparison much smaller, and are similar to those seen in Figure 7.3, showing the actual poll data proportions. This suggests that the models seem to be able to replicate quite adequately the type of development expected with polls before a general election. The (latent) ever-changing probabilities offer natural progressions of the votes as we get closer to the exit poll, and our simulations show how the probabilities of voting can reflect changes in power of the parties as time goes on. The simulations for 1983 with three parties show change of strength of Labour versus Liberal/Alliance (but with Conservative clearly in lead). Such changes were apparent in the actual poll votes, as we saw at the end of the 1983 polls and the start of the 1984 polls. Another advantage over ARCH models is that the stochastic volatility models

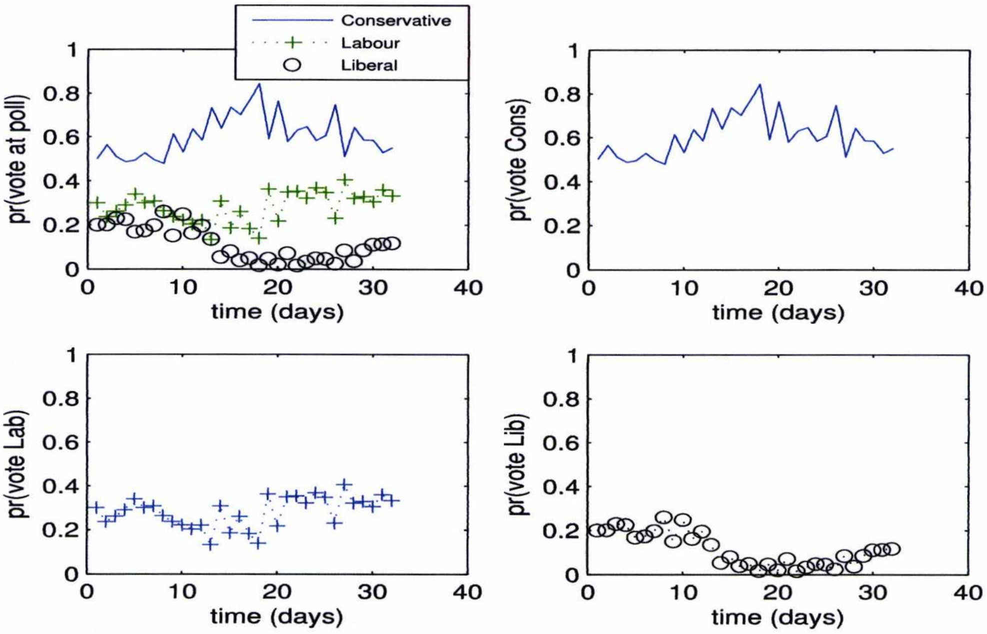


Figure 7.6: Simulations of probabilities of voting from SV(3,2) using $\tilde{\phi} = 0.13$ and $\tilde{w} = 0.44$, obtained from the 1983 data.

consider time differences between polls.

An objective of the modelling is for the actual data to look as if a realisation of our best models. It is clear by comparing these simulations with the proportions of votes seen earlier in the chapter (Figures 7.1 and 7.3) that the actual proportions could indeed be realisations from our best models.

Expected values

Figures 7.7 and 7.8 show the proportion of votes for each party compared with the expected values given the best two-party model, SV(2,3), and the best three-party model, SV(3,2), respectively. We can see that the model gives generally quite accurate expected values, which suggests that the models are appropriate. However, our choice at the first poll, $t = 1$, could have been better.

Hypothesis tests

At this point of the inference, this test would be useful if we wished to assess the importance of particular weights representing the lags, especially if we suspected that one, say, was almost negligible in size, say only 0.05. For our examples though the weights are large enough, so there is little to be gained in performing this test.

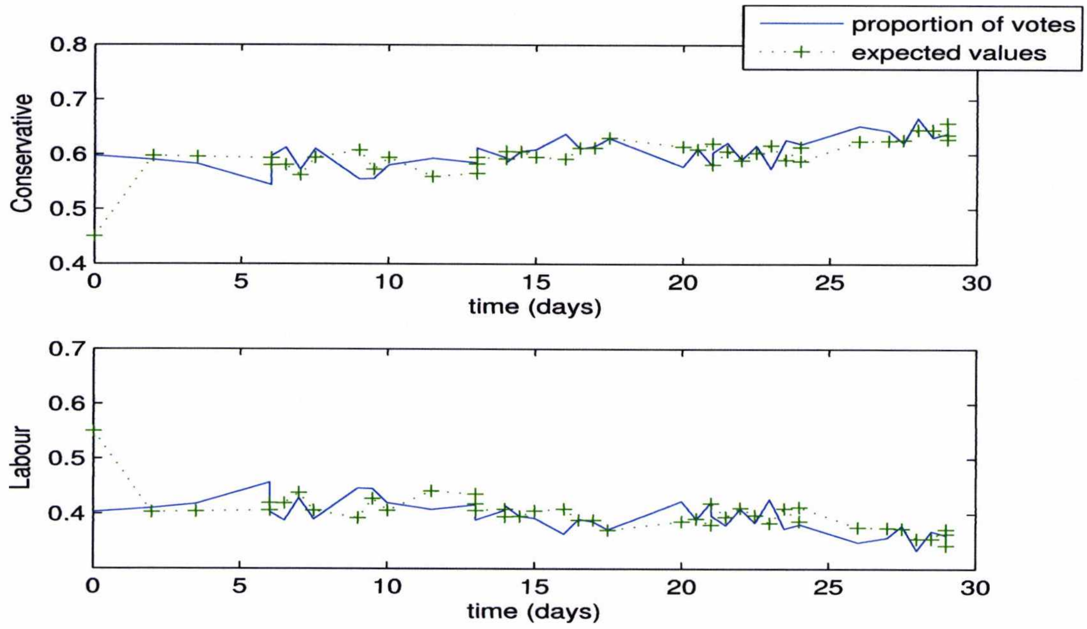


Figure 7.7: Proportion of votes for Conservative and Labour versus SV(2,3) expected values.

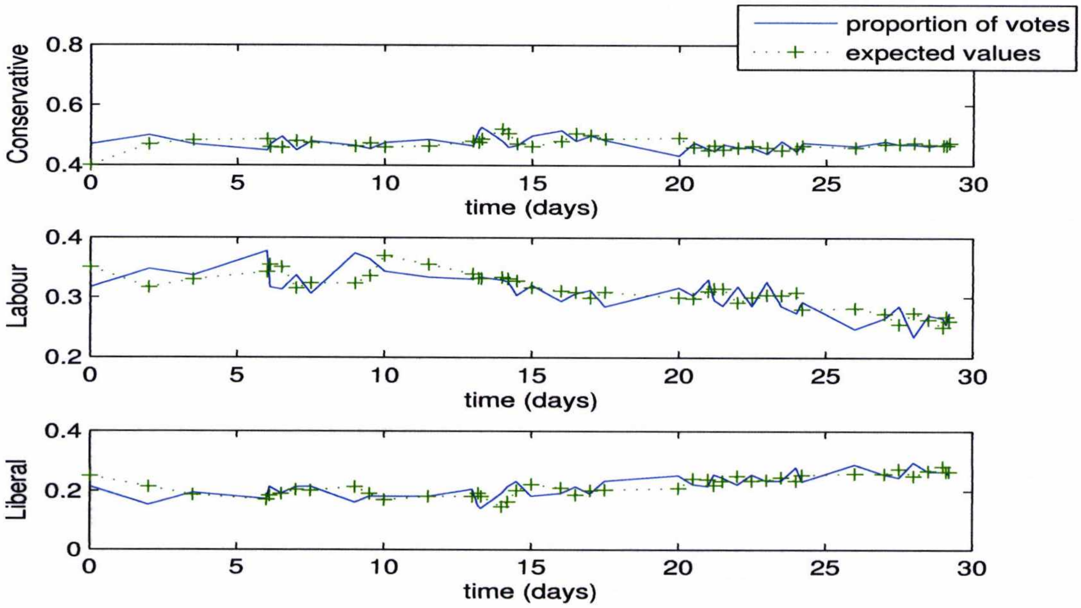


Figure 7.8: Proportion of votes for Conservative, Labour and Liberal versus SV(3,2) expected values.

7.4 Illustration 2: 1987 Data

7.4.1 MHLEs

Table 7.3 shows the MHLE values obtained for all models when fit to the 1987 data. Also shown are the values for $\tilde{h}$, which we may interpret in the same way that we would with log-likelihood values, that is, the MHLEs found give us the maximum absolute value of h.

Model	$\tilde{\theta}'$	$\tilde{h}$
ARCH(2,1)	$(\tilde{\alpha} = 2879.3, \tilde{\beta} = 2245.1)$	-2451.7
ARCH(2,2)	$(\tilde{\alpha} = 2066.4, \tilde{\beta} = 1611.2, \tilde{w} = 0.5667)$	-2418.1
ARCH(2,3)	$(\tilde{\alpha} = 1081.0, \tilde{\beta} = 842.9, \tilde{w}_1 = 0.3333, \tilde{w}_2 = 0.2000)$	-2091.4
SV(2,1)	$(\tilde{\phi} = 0.3409)$	-128.2
SV(2,2)	$(\tilde{\phi} = 0.1716, \tilde{w} = 0.2555)$	-118.6
SV(2,3)	$(\tilde{\phi} = 0.0736, \tilde{w}_1 = 0.0563, \tilde{w}_2 = 0.1924)$	-109.6
ARCH(3,1)	$(\tilde{\alpha_1} = 3584.3, \tilde{\alpha_2} = 2798.6, \tilde{\alpha_3} = 2012.9)$	-4775.3
ARCH(3,2)	$(\tilde{\alpha_1} = 2127.6, \tilde{\alpha_2} = 1661.2, \tilde{\alpha_3} = 1194.8, \tilde{w} = 0.5333)$	-4597.4
ARCH(3,3)	$(\tilde{\alpha_1} = 1287.0, \tilde{\alpha_2} = 1004.9, \tilde{\alpha_3} = 722.8, \tilde{w}_1 = 0.3667, \tilde{w}_2 = 0.2000)$	-3960.9
SV(3,1)	$(\tilde{\phi} = 0.2193)$	-236.7
SV(3,2)	$(\tilde{\phi} = 0.1033, \tilde{w} = 0.2420)$	-218.1
SV(3,3)	$(\tilde{\phi}{=}0.0303, \tilde{w}_1 = 0.0255, \tilde{w}_2 = 0.1940)$	-201.7

Table 7.3: MHLEs for candidate models when fit to the 1987 poll data.

Inference

ARCH models: As before, the party strength parameters α are large in magnitude, due to the large figures for the number of votes in the polls. Then, the outputs are similar regardless of the model. For the 1987 data the models suggest a split of about 55 per cent and 45 per cent for Conservative versus Labour respectively, and about 45 per cent, 30 per cent and 25 per cent for Conservative, Labour and Alliance respectively. These figures reflect what we saw in Figures 7.1-7.4, showing the poll outcomes by proportions, to suggest that MHLEs α are very reliable for further use, rather than relying on graphs we now have determined values to use.

The lag parameters w vary, depending on the model. With two parties only, ARCH(2,2) favours the previous lag (57 per cent), but ARCH(2,3) reduces this to

33 per cent and allocates almost half to lag three. With three parties ARCH(3,2) attributes 53 per cent to lag one and ARCH(3,3) attributes 37 per cent to it, but 43 per cent to lag three. It seems for the above as if the models never put much weight on the second lag, similarly to with the 1983 data.

Stochastic volatility models: Again, we decided to replace the latent probabilities with the proportion of votes per party. For all of our models here our value of ϕ is very small, roughly somewhere between 0.01 and 0.1. Typically our values for $\delta_{s(t)}$ will also be small; considering both datasets the maximum number of days between polls was 3. This implies that our value for $c_{s(t)}$ is reasonably large, say 100, and so we have very different probability densities at each poll.

For the 1987 data SV(3,2) attributes only 26 per cent to the first lag yet 74 to the second. This contrasts with the comparative ARCH models, which put more emphasis on the first lag than the second. Similar is true in the three-party case. As with the ARCH models, when we go on to consider three parties, with the 1987 data again most weight (about 75 per cent) was placed on the third lag, but more on the second than on the first.

7.4.2 Dispersion and interval estimates

Now we present the approximate correlation matrices for the MHLEs, as well as their approximate 99% confidence intervals.

ARCH(2,1)

$$\begin{pmatrix} 110.9241 & 0.9910 \\ 0.9910 & 97.2728 \end{pmatrix}$$

$$\alpha \in (2593.6, 3165.0)$$

$$\beta \in (1994.5, 2495.7)$$

ARCH(2,2)

$$\begin{pmatrix} 75.5363 & 0.9850 & 0 \\ 0.9850 & 66.1551 & 0 \\ 0 & 0 & 0.3684 \end{pmatrix}$$

$$\alpha \in (1871.8, 2261.0)$$

$$\beta \in (1440.8, 1781.6)$$

$$w \in (0, 1)$$

ARCH(2,3)

$$\begin{pmatrix} 32.8767 & 0.9504 & 0 & 0 \\ 0.9504 & 28.6181 & 0 & 0 \\ 0 & 0 & 0.0861 & -0.3536 \\ 0 & 0 & -0.3536 & 0.0730 \end{pmatrix}$$

$$\alpha \in (996.3, 1165.7)$$

$$\beta \in (769.2, 916.6)$$

$$w_1 \in (0.1115, 0.5551)$$

$$w_2 \in (0.0120, 0.3880)$$

SV(2,1)

$$(0.0593)$$

$$\phi \in (0.1881, 0.4936)$$

SV(2,2)

$$\begin{pmatrix} 0.0357 & 0.3898 \\ 0.3898 & 0.0096 \end{pmatrix}$$

$$\phi \in (0.0796, 0.2637)$$

$$w \in (0.0035, 0.5074)$$

SV(2,3)

$$\begin{pmatrix} 0.0171 & 0.3714 & 0.0789 \\ 0.3714 & 0.0465 & -0.3256 \\ 0.0789 & -0.3256 & 0.1017 \end{pmatrix}$$

$\phi \in (0.0296, 0.1176)$

$w_1 \in (-0.0636, 0.1762)$

$w_2 \in (0.0725, 0.3123)$

ARCH(3,1)

$$\begin{pmatrix} 111.5159 & 0.9798 & 0.9851 \\ 0.9798 & 74.4733 & 0.9827 \\ 0.9851 & 0.9827 & 58.1956 \end{pmatrix}$$

$\alpha_1 \in (3296.8, 3871.2)$

$\alpha_2 \in (2606.8, 2990.4)$

$\alpha_3 \in (1863.0, 2162.8)$

ARCH(3,2)

$$\begin{pmatrix} 62.7116 & 0.9579 & 0.9689 & 0 \\ 0.9579 & 42.2192 & 0.9638 & 0 \\ 0.9689 & 0.9638 & 32.9577 & 0 \\ 0 & 0 & 0 & 0.3660 \end{pmatrix}$$

$\alpha_1 \in (1966.0, 2289.1)$

$\alpha_2 \in (1552.5, 1769.9)$

$\alpha_3 \in (1109.9, 1279.7)$

$w \in (0, 1)$

ARCH(3,3)

$$\begin{pmatrix} 32.0489 & 0.8938 & 0.9194 & 0 & 0 \\ 0.8938 & 22.2579 & 0.9074 & 0 & 0 \\ 0.9194 & 0.9074 & 17.5752 & 0 & 0 \\ 0 & 0 & 0 & 0.0880 & -0.3804 \\ 0 & 0 & 0 & -0.3804 & 0.0730 \end{pmatrix}$$

$\alpha_1 \in (1204.4, 1369.6)$

$\alpha_2 \in (947.6, 1062.2)$

$\alpha_3 \in (677.5, 768.1)$

$w_1 \in (0.1400, 0.5934)$

$w_2 \in (0.0120, 0.3880)$

SV(3,1)

$$(0.0296)$$

$\phi \in (0.1431, 0.2956)$

SV(3,2)

$$\begin{pmatrix} 0.0158 & 0.3842 \\ 0.3842 & 0.0707 \end{pmatrix}$$

$\phi \in (0.0627, 0.1440)$

$w \in (0.0599, 0.4240)$

SV(3,3)

$$\begin{pmatrix} 0.0033 & 0.2023 & 0.0265 \\ 0.2023 & 0.0249 & -0.4020 \\ 0.0265 & -0.4020 & 0.0599 \end{pmatrix}$$

$\phi \in (0.0219, 0.0387)$

$w_1 \in (-0.0387, 0.0897)$

$w_2 \in (-0.0396, 0.3484)$

Inference

ARCH models: As we saw with first dataset, standard errors for party parameters are quite large, but not relative to corresponding parameter sizes. There is strong positive correlation between them. Again, MTD model parameters are independent from the lag parameters. The lag parameters are negatively correlated, which also makes sense since they all sum to one, and so increasing one would mean decreasing the others. In all instances above, the confidence intervals tell us (with 99 per cent certainty) that overall during the poll period, Conservative was more powerful than Labour which in turn was more powerful than Liberal/Alliance. Confidence intervals are more useful when there are several weight parameters.

Stochastic volatility models: As with the 1983 data, standard errors are suitably low for all parameters, enabling more accurate confidence intervals. Again, the weight parameters are negatively correlated.

7.4.3 Relative goodness of fit

Table 7.4 shows the deviance values for each model as well as the number of parameters, d.

Model	d	D
ARCH(2,1)	2	4507.9
ARCH(2,2)	3	4236.4
ARCH(2,3)	4	3360.0
SV(2,1)	1	524.6
SV(2,2)	2	433.1
SV(2,3)	3	433.3
ARCH(3,1)	3	8765.6
ARCH(3,2)	4	7885.2
ARCH(3,3)	5	6391.7
SV(3,1)	1	554.5
SV(3,2)	2	554.0
SV(3,3)	3	554.7

Table 7.4: Deviances for candidate models when fit to the 1987 poll data.

Inference

For the 1987 data, the deviance method selects SV(2,2) when there are two parties and SV(3,2) when there are three parties.

Remarks: For both pre-election datasets, the best model in a three-party contest is SV(3,2). If we reconsider the types of simulations in Chapter 6 which the ARCH models gave us, the changes with time are too extreme compared with those seen in Figures 7.3 and 7.4. The stochastic volatility models by contrast take into account time difference between observations, and we ended up with much smoother time series plots. Also these models generally have fewer parameters. We therefore now drop the ARCH models and focus on the stochastic volatility models, firstly, to assess absolute goodness of fit and then in the modelling on election night.

One interesting observation is that the best ARCH model for both polls was ARCH(2,3), that is, with three lags. For both the two-party and three-party contests lag three was allocated a large proportional value. This may mean that if we considered even more lags then the deviance values for the corresponding models would be even lower, though would still not compete with the stochastic volatility models, whose deviance values are far lower.

For our stochastic volatility models, it should be noted that the deviance values are close to one another. For example, for the 1983 data the deviance of the best model SV(2,3) in a two-party contest is 387.07, whereas the second best is SV(2,2) with deviance 387.08. For simplicity we might then choose the latter as it is simpler in terms of computation, yet still is not much worse than the best model.

7.4.4 Absolute goodness of fit

Simulations

Recall that we saw typical simulations of our models in the previous chapter, but now we update them to contain our MHLE parameters. Figure 7.9 shows simulations from the best two-party model for the 1987 data (SV(2,2)). Figure 7.10 shows the best model, which was SV(3,2), with their unique MHLEs when fit to the 1987 poll data.

As before, it is clear that all plots are smoother than those obtained from the spiky ARCH models. The spikes here are by comparison much smaller, and are similar to those seen in Figure 7.4, showing the actual poll data proportions

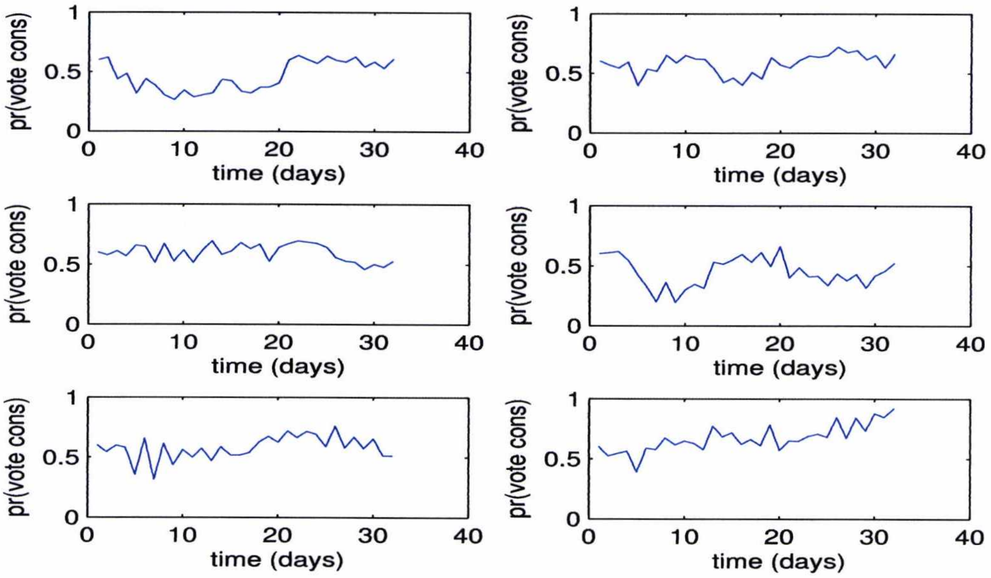


Figure 7.9: Simulations of probabilities of voting from SV(2,2) using $\tilde{\phi} = 0.17$ and $\tilde{w} = 0.26$, obtained from the 1987 data.

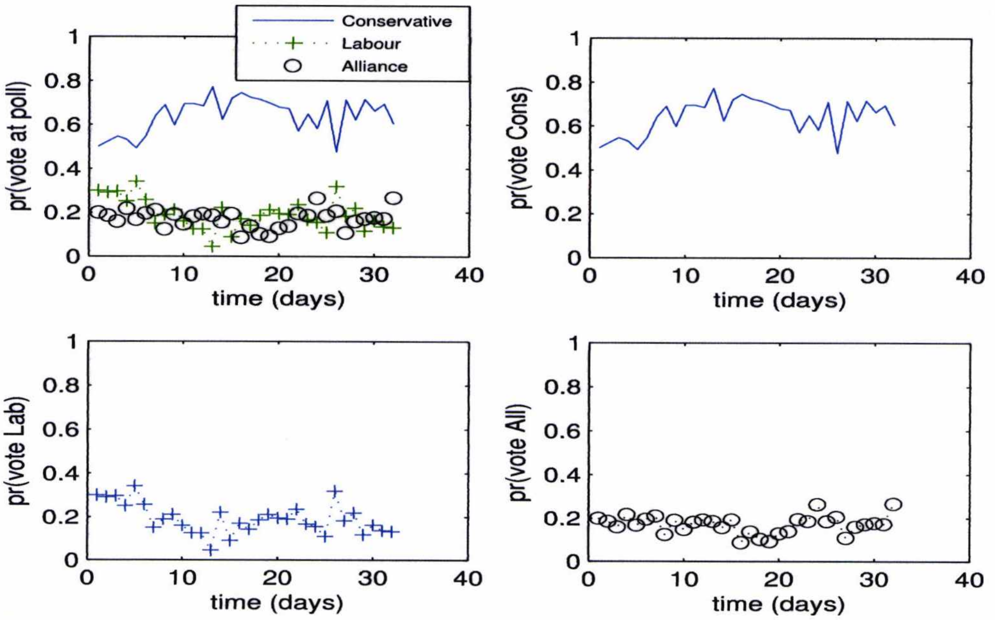


Figure 7.10: Simulations of probabilities of voting from SV(3,2) using $\tilde{\phi} = 0.10$ and $\tilde{w} = 0.24$, obtained from the 1987 data.

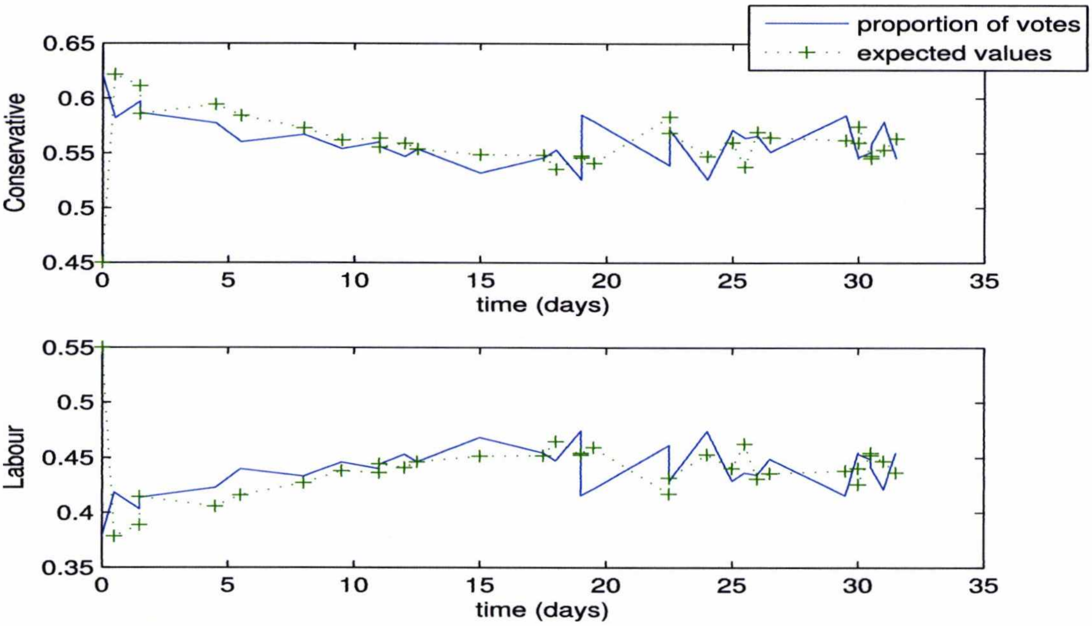


Figure 7.11: Proportion of votes for Conservative and Labour versus SV(2,2) expected values.

themselves. This suggests that the models seem to be able to replicate quite adequately the type of development expected with polls before a general election. The (latent) ever-changing probabilities offer natural progressions of the votes as we get closer to the exit poll, and our simulations show how the probabilities of voting can reflect changes in power of the parties as time goes on. The simulations for 1987 with three parties show change of strength of Labour versus Liberal/Alliance (but with Conservative clearly in lead).

Also, if we compare these simulations with the proportions of votes seen earlier in the chapter (Figures 7.2 and 7.4) that the actual proportions could indeed be realisations from our best models.

Expected values

Figures 7.11 and 7.12 show the proportion of votes for each party compared with the expected values given the best two-party model, SV(2,2), and the best three-party model, SV(3,2), respectively. We can see that the model gives generally quite accurate expected values, which suggests that the models are appropriate. However, as with the 1983 data, our choice at the first poll, $t = 1$, could have been better.

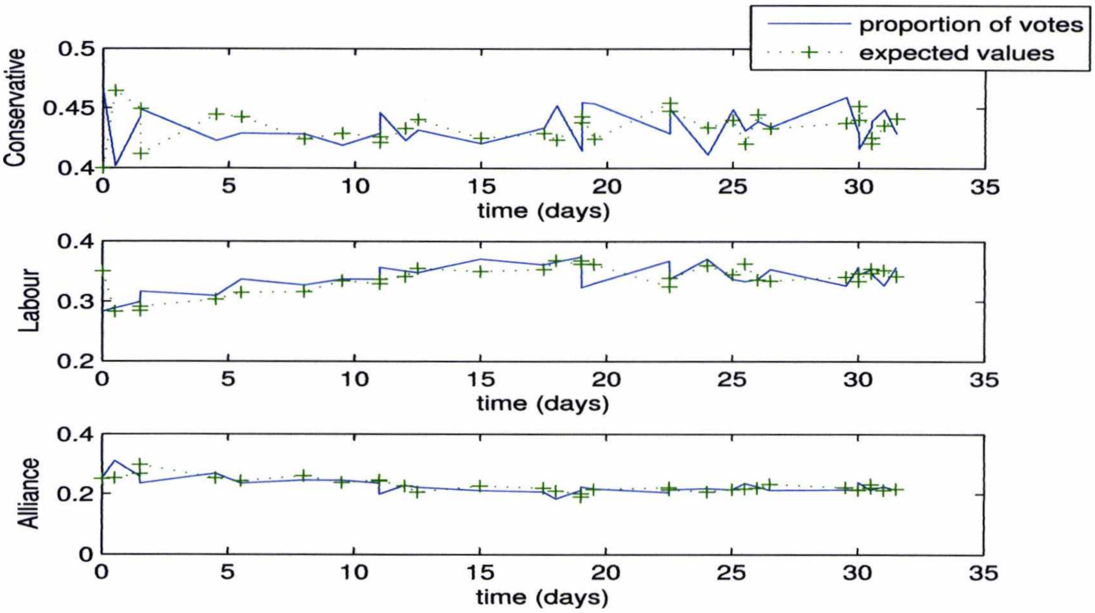


Figure 7.12: Proportion of votes for Conservative, Labour and Liberal versus SV(3,2) expected values.

Final models:

- SV(2,3) (two parties) and SV(3,2) (three parties) for 1983; and
- SV(2,2) (two parties) and SV(3,2) (three parties) for 1987.

7.5 Comparisons with Other Models

7.5.1 Background

It is worth comparing the performance of our method with other methods. In Chapter 4 we mentioned both the poll of polls and the multivariate structural time series approach to model polls, each of which provides a prior forecast *before* the exit poll results are known. Although we have just discussed how our method may be easily incorporated into the regression-based forecast throughout election night, we may use parts of our method to derive our own comparative prior forecast. To do this, however, we would need to employ an ARCH model rather than the even better stochastic volatility model. This is because the former is a stationary model, effectively meaning that in the three party case we have the stationary case that

$$\left(\frac{\alpha_1}{\sum_{k=1}^{3}\alpha_k}, \frac{\alpha_2}{\sum_{k=1}^{3}\alpha_k}, \frac{\alpha_3}{\sum_{k=1}^{3}\alpha_k}\right),$$

in which α_1 represents the strength of Conservative, α_2 Labour and α_3 Liberal/Alliance. Therefore we have fixed parameters to work with, whereas with SV(3,2) the probability densities evolve with each poll and without any common parameters regarding party strength. We found that our best ARCH model given the 1983 data was ARCH(3,3). The same was true given the 1987 data. In Harvey and Shephard (1990, [55]) the results for the local level model and the poll of polls were given; we now contribute our results - these are displayed in Tables 7.5 and 7.6. Also shown is the actual outcome at the end of election night. It is important to realise that these figures involve comparing party strength in terms of votes and not seats, when in fact it is the latter which determines which party becomes Government.

Model	Conservative	Labour	Liberal
Outcome	**44.5**	**28.9**	**26.6**
ARCH(3,3)	46.5 (-2.0)	30.3 (-1.4)	23.2 (+3.4)
Shephard	46.8 (-2.3)	26.7 (+2.1)	26.5 (+0.1)
Poll of Polls	47.3 (-2.8)	26.9 (+2.0)	25.8 (+0.8)

Table 7.5: Comparison of our best (ARCH) model with other models when fit to the 1983 poll data.

Model	Conservative	Labour	Alliance
Outcome	**44.2**	**32.2**	**23.6**
ARCH(3,3)	42.7 (+1.5)	33.3 (-1.1)	24.0 (-0.4)
Shephard	43.5 (+0.7)	34.6 (-2.4)	21.9 (+1.7)
Poll of Polls	43.1 (+1.1)	34.7 (-2.5)	22.3 (+1.2)

Table 7.6: Comparison of our best (ARCH) model with other models when fit to the 1987 poll data.

7.5.2 Analysis

We see that our forecasts using MHLEs are similar overall to those of the other two methods. In the 1983 election, our result for Conservative was the best of the three methods with the least overestimation. Unlike the other methods, we overestimated Labour, but again the absolute error was the smallest. Our forecast for Liberal was somewhat larger than the other two methods, which were both quite close in their predictions. With the other two methods the overestimation of Conservative was offset by the underestimation of Labour, which led to their predictions for Liberal being accurate. In our case, we overestimated both Conservative and Labour, which meant that the forecast for Liberal had to be more underestimated than the other methods.

In the 1987 election, our forecasts were quite good throughout. We underestimated Conservative only by 1.5 per cent and overestimated Labour only by 1.1 per cent. This offset meant that we only overestimated Alliance by 0.4 per cent. Our forecast for Labour was the most accurate, as was that for Alliance, and our forecast for Conservative although the least accurate was not much worse (less than 1 per cent compared to both methods).

Regardless of the comparative accuracy of these prior forecasts, both our method and that of Harvey and Shephard (1990) provide more information than the poll of polls, by operating within time series. This information is often of importance to psephologists who carry out trend analysis, for instance, and politicians who might use such analysis to adapt their campaigns accordingly.

7.6 Review

We fitted our ARCH and stochastic volatility models to the 1983 and 1987 preelection poll data; note that we would expect the GARCH models to behave very similarly to the latter, just as they did in the simulations seen in the previous chapter. Through the h-likelihood approach we obtained estimates in replacement of MLEs. Generally, the party parameters for all models had similar implications in terms of summarising party strength. What we found varied from model to model, however, is the distribution of weight given to the lags. Subsequently, we made use of theory given in Lee and Nelder (1996) and Lee, Nelder and Pawitan (2001, [72]) to perform inference on our estimates. This included deriving approximate covariance matrices and confidence intervals, which were not surprising in what they showed. It also included the scaled deviance test used to choose the best of a set based on comparising a numerical value. This statistic chose the stochastic volatility models in all cases, and preferred the multi-lag models. Simulations of the chosen models confirmed that the models were reliable in the modelling of poll data treated as time series. Finally, we provided a point estimate prior forecast based on the number of votes, in order to compare part of our method with other established methods, for which our results are encouraging. We have seen that our general models involve gauging an idea of the distribution of popularity for parties, based on observed data. These data come from polls and cover a reasonable period of time in the run up to election night. Clearly, this is merely the first of two stages. The subsequent stage is to incorporate our modelling in the forecasting during election night.

Chapter 8

Election Night

8.1 Introduction

In this chapter, we firstly provide an overview of how we go about modelling each seat, in terms of the number of votes and the probability of voting. In Section 8.2 we devote much time reviewing the BBC's regression-based method of forecasting on election night, which has seen much development and revision in the experience of recent general elections. Section 8.3 then discusses in detail our forecasting method, starting with the simplest two-party contest and then developing to a more general three-party contest. Finally, in Section 8.4 we illustrate forecasting of a three-party contest in a simplified version of the UK election for the 1983 and 1987 data.

8.1.1 Overview

Similarly to opinion polls, once we observe for any seat the number of votes for Conservative (without loss of generality) among the total number of votes for that seat, we assume the former to be binomially distributed, that is,

$$y_j \sim Bin(r_j, x_j),$$

where y_j is the number of Conservative votes in seat j, r_j the total number of votes in that seat, and x_j the probability of voting Conservative within that seat (which has a separate distribution). This extends straightforwardly to the multinomial distribution for more parties than two, so clearly each constituency has a unique distribution both for y_j and the latent x_j.

Our general formulation thus has two purposes: firstly to model the opinion polls *before* the election night and therefore to produce an initial forecast, and secondly to model the updates of the predictive distributions *throughout* the election night to lead to a final forecast, that is, where we recognise a convergence of predicted number of seats won for each party, via a series of revised forecasts. In the former, the data involve numbers who *say* that they will vote for a party and the latter numbers who *did* vote for a party. Further, they involve in the former *national* probabilities of voting for a party at that time, yet the latter involves the probabilities of voting for a party in a *constituency*. In both cases, however, these probabilities are latent, that is to say, they work in the background to generate the numbers of votes per party. The link between the two cases is that modelling behind the opinion polls reappears in the election night forecasting, via the *probability* models. Put another way, the modelling of the polls is not confined merely to providing a prior forecast before election night. Our general forecasting approach is based heavily on that taken by the BBC, which is outlined in Brown and Payne (1975, [23], 1984, [24]) and Brown, Firth and Payne (1999, [22]).

Exit polls

Until exit polls are carried out, the results (from opinion poll analysis covered in previous chapters) we assume to hold nationally across the UK. (Our theory developed thus far may be as broad or as narrow as required, and so it is possible to apply this to constituency level using constituency polls, exit polls and previous election results.) Once exit polls have been obtained, we focus on revising to local level and constituency-specific models. First, suppose naively that it is possible to carry out an exit poll for each constituency. Thus each seat has its unique (and prior) distribution for the probability of voting in the exit poll. For example, suppose that in a two-party contest and with x_j denoting the probability of voting Conservative in seat j in the exit poll

$$\pi(x_j) \sim Be\left(q_j N_j,\, (1-q_j)N_j\right), \tag{8.1}$$

in which

- N_j is the number of respondents in the exit poll conducted at seat j; and
- q_j is a proportion in order to apportion the N_j among the shape parameters of the beta distribution, with the first representing Conservative and the second Labour. Therefore,

  $$q_j = \frac{C_j}{N_j},$$

  in which C_j is the number of votes for Conservative in constituency j.

From (8.1) we have that $\mathbb{E}(x_j) = q_j$, and so we end up *expecting* a local probability of voting Conservative (and thus Labour), before any seats have declared; at the same time though we see that each seat has its *unique* beta distribution.

Ideally, we would want exit poll data for each constituency, but this is not collected due to time and expense. Therefore, we can use what *is* collected regarding exit polls *by pooling* for the UK and apply this to each constituency as relevant. This means that lots of constituencies have the same prior; for example, the seats in the UK where Conservative is strong will be allocated the same prior as those in the exit poll sample where Conservative is strong. See Curtice and Firth (2008, [32]) for how to go about pooling exit poll data.

Election night

Next, we concentrate on modelling the election night probabilities of voting given the seats currently declared. This clearly means that we must sequentially update these probabilities as each seat declares, until either all seats have declared or ideally when we are able to give a final prediction. In our revised probability models we continue to use the opinion poll methodology as well as the option to use the exit poll information above, combined in a weight function for instance. We may choose to give the latter less and less weight as the number of declared seats increases, since we can then rely more on what the actual declarations say rather than what the now somewhat historic exit polls say. Alternatively, we could just keep the weight fixed and argue that what an exit poll taken at seat j says will be independent from all other seats; therefore, the exit poll at seat 659 would have as important a role as that at seat 1.

8.2 Brown and Payne Model

8.2.1 Objective

Brown and Payne (1975) discussed a regression-based approach to election forecasting. As the night goes on and seats are declared, a **multiple linear regression** modelling of seats so far declared is performed on explanatory variables of possible influence. These variables collectively cover a range of factors, including previous election votes, regional dummy variables, interaction dummy variables and socioeconomic variables. For example, in the 1997 model there were 36 such variables. The method is **ridge regression** in a slightly-modified form, and has the advantage of enabling any number of variables to be included even without any data (in which case the corresponding coefficients would be shrunk to zero).

8.2.2 Data

A prior forecast is made at the start of election night; that is, before any seats have been declared. The BBC method actually makes forecasts of the *change in share* of vote from previous elections per party per seat. The sources of information here are chiefly exit polls and also opinion polls. Three seats are chosen and used as if they were dummy seats, and their results were turned into artificial votes, assuming an 80 per cent turnout of the electorate. Then, all the declared results per constituency on the night are used in the regression-based forecasting.

8.2.3 The model

Initial grouping

Brown and Payne (1975) split the seats into three categories: **special seats**, two-party contests (Conservative versus Labour) and three-party contests (Conservative, Labour and Liberal). For the special seats, no modelling is done as such seats will be unreliable if used to forecast other seats and vice versa. Instead, they are assigned fixed a priori probabilities of winning (until declared; these probabilities are modified manually if needed). Assumptions with the two-party seats is that the nationalist parties 'remain local' and that Liberal has no chance of winning, due to the fact that it received less than 8 per cent of votes in the last election.

Seats declared in the two-party contest are used with those in the three-party contest to predict the three-party forecast and vice versa for the two-party contest. This is achieved by defining two *overlapping* groups:

1. One group comprising all seats in which Liberal is currently standing; and
2. One group comprising all seats in which Liberal received no more than 30 per cent of the votes in the last election.

This implies the need for a **dummy variable** for Liberal intervention in both groups.

Probability matrix

Overall, assume that there are 635 constituencies and 5 parties: Conservative, Labour, Liberal, Nationalist and Other. $\hat{\mathbf{P}}$ is a 635 × 5 probability matrix

$$\hat{\mathbf{P}} = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{15} \\ P_{21} & P_{22} & \dots & P_{25} \\ \dots & \dots & \dots & \dots \\ P_{6351} & P_{6352} & \dots & P_{6355} \end{pmatrix}$$

in which the element P_{jg} is the probability that the j^{th} seat goes to the g^{th} party, and where $\sum_{g=1}^{5} P_{jg} = 1$ (until that seat is declared). Then, the expected number of seats that party g wins is simply the sum of the g^{th} column.

Two-party contest (Conservative versus Labour)

- The multiple regression model is

$$y_j = \beta_0 + \sum_{g=1}^{r} \beta_g x_{jg} + \epsilon_j,$$

  in which y_j is the change in the Conservative share of the Conservative plus Labour vote (known as 'two-party swing') in constituency j, x_{jg} are explanatory variables for constituency j, β_g are the regression coefficients to be estimated, with β_0 a constant, and $\epsilon_j \sim N(0, \sigma^2)$ are errors. (Brown and Payne mention that the normality assumption is well justified empirically.)

- Suppose that there are only five variables of interest ($r = 5$), in priority order:

  1. Conservative in the last election (x_{j1});
  2. Liberal intervention (x_{j2});
  3. Liberal in the last election (x_{j3});
  4. Scottish nationalists in the last election (x_{j4}); and
  5. South East of England constituencies (x_{j5}).

  The method has a **staged inclusion** for these variables, meaning that the number of explanatory variables increases as more seats are declared: when $5 + 3(d - 1)$ seats have declared we may include d explanatory variables.

- 'Optimum' estimation for the model with identical design matrices is by way of univariate least squares, but this is equivalent to maximum likelihood under normality. With different design matrices, the degree of optimality depends on the generalised canonical correlations of the sets of regressor variates. In the three-party contest (see later) there will be a small degree of optimality, especially because two variables are highly correlated.

- We would follow an equation-by-equation approach to estimate the β vector, $\hat{\beta}$, and so may drop the subscript defining the specific party. Hence, we must solve
$$\hat{\beta} = (\mathbf{X'X}) + k \cdot Diag(0,\, 1,\, \ldots,\, 1))^{-1}\mathbf{X'Y},$$
in which k is a ridge constant. Brown and Payne (1975) set this equal to 4.0, based on the behaviour of the program in the previous election at the time of writing. Obviously, we regress using only the declared results and corresponding explanatory variables, the latter of which form the design matrix $\mathbf{X}$. Hence, letting n denote the number of constituencies within the two-party category,
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \ldots \\ \epsilon_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_2 \\ \ldots \\ \beta_5 \end{pmatrix},$$
$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{15} \\ 1 & x_{21} & \ldots & x_{25} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{n5} \end{pmatrix}, \quad Diag(0,\, 1,\, \ldots,\, 1) = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

- The explanatory variables are introduced in priority order (except when $\mathbf{X}$ is nearly singular, when a lower priority variable is fitted).

- Also, we estimate the variance of the errors, σ^2, by
$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})}{n_0 - r},$$
where r is the number of parameters in the model and n_0 is the number of constituencies which have declared.

- The predicted swing of the j^{th} undeclared seat then is
$$\hat{y}_j = \mathbf{x'_j}\hat{\beta},$$

where $\mathbf{x_j}$ is the $(r+1)\times 1$ explanatory variable for that constituency, and if we let

$$\mathbf{H}^{-1} = (\mathbf{X}'\mathbf{X}) + k \cdot Diag(0,\, 1,\, \ldots,\, 1)$$

then the predicted variance ${a_j}^2$ for constituency j is

$$ {a_j}^2 = \hat{\sigma}^2(1 + \mathbf{x'_j}\mathbf{H}^{-1}\mathbf{x_j}). \tag{8.2}$$

- To estimate the probabilities, for Conservative say, we use

$$\hat{P}_{j1} = \Phi\left(\frac{c_j + \hat{y}_j - 0.5}{a_j}\right),$$

in which c_j is the Conservative share at the two-party vote in the last election, $c_j + \hat{y}_j$ is the Conservative predicted share, and 0.5 is the tie value. Also, Φ is the cumulative distribution function of the normal distribution and, clearly, $\hat{P}_{j2} = 1 - \hat{P}_{j1}$.

The three-party scenario is more complicated to generate the probabilities. The regression model extends to three equations now, and we have $\{y_{j1}\}$, $\{y_{j2}\}$ and $\{y_{j3}\}$.

Three-party contest

- The multiple regression model is

$$\begin{aligned} \mathbf{y_1} &= \mathbf{X_1}\beta_\mathbf{1} + \epsilon_\mathbf{1} \\ \mathbf{y_2} &= \mathbf{X_2}\beta_\mathbf{2} + \epsilon_\mathbf{2} \\ \mathbf{y_3} &= \mathbf{X_3}\beta_\mathbf{3} + \epsilon_\mathbf{3}, \end{aligned} \tag{8.3}$$

where $\mathbf{y}'_m = (y_{1m},\, y_{2m},\, \ldots,\, y_{nm})$, $\epsilon'_m = (\epsilon_{1m},\, \epsilon_{2m},\, \ldots,\, \epsilon_{nm})$, $\beta'_m = (\beta_{0m},\, \beta_{1m},\, \ldots,\, \beta_{rm})$ and $\mathbf{X}_m = (\mathbf{x}_{0m},\, \mathbf{x}_{1m},\, \ldots,\, \mathbf{x}_{rm})$, an $n \times (r+1)$ matrix of explanatory variables, with $\mathbf{x}_{0m}$ a vector of ones, $m = 1,\, 2,\, 3$. Now let n denote the number of constituencies within the three-party category.

Also, $\mathbb{E}(\epsilon_m) = \mathbf{0}$ (a column of zeros) and $\mathbb{E}(\epsilon_m \epsilon'_q) = \sigma_{mq}\mathbf{I}$, where $\mathbf{I}$ is an $n \times n$ identity matrix and $m,\, q = 1,\, 2,\, 3$ and we assume multivariate normality of the error structure. We define the matrix (σ_{mq}) by $\mathbf{\Sigma}$.

Here y_{j1} is the change in Conservative share, y_{j2} Labour and y_{j3} Liberal for the j^{th} seat.

- Equation-by-equation ridge estimation is applied to give estimates $\hat{\beta}_m$, $m = 1, 2, 3$. This does not correspond in the Bayesian sense to the components $\beta_\mathbf{1}, \beta_\mathbf{2}, \beta_\mathbf{3}$ being a random sample from $N(0, {\sigma^2}_\beta)$. Brown and Payne (1975) show that equation-by-equation multivariate ridge regression has similar mean-squared error benefits to univariate ridge regression, but is not a Bayesian procedure. Therefore, it is deficient. However, Bayes specification of a realistic prior variance-covariance structure for $\beta' = (\beta_\mathbf{1}', \beta_\mathbf{2}', \beta_\mathbf{3}')$ is not a great problem given its likely dependence on $\mathbf{\Sigma}$.

- For the j^{th} undeclared seat we have a trivariate vector of the predicted change in share of the electorate $\hat{\mathbf{y}}'_m = (\hat{y}_{j1}, \hat{y}_{j2}, \hat{y}_{j3})$, where

$$\hat{y}_{jm} = \mathbf{x}'_{jm}\beta_m,$$

with $\mathbf{x}'_{jm}$ the $(r+1) \times 1$ vector of explanatory variables for the m^{th} equation, $m = 1, 2, 3$.

- The covariance matrix $\mathbf{\Sigma}$ is estimated by $\mathbf{V} = (v_{ml})$ where

$$v_{ml} = \frac{(\mathbf{y}_m - \mathbf{X}_m\hat{\beta}_m)'(\mathbf{y}_l - \mathbf{X}_l\hat{\beta}_l)}{n_0 - r},$$

where n_0 is the number of relevant declared seats and $m,\, l = 1, 2, 3$.

- If we now let $\mathbf{Z}_j$ be a random vector denoting the share predicted to go to each party for the j^{th} seat, then we estimate that

$$\mathbf{Z}_j \sim N_3(\mathbf{c_j} + \hat{\mathbf{y}}_\mathbf{j}, \mathbf{V}),$$

where $\mathbf{c_j}$ is the vector of shares at the last election for seat j.

- The variability due to uncertainty of estimation of the parameter (as in (8.2)) is neglected; this second-rate contribution would mostly cancel when estimating the total number of seats per party.

- We have that

$$\begin{aligned}\hat{P}_{jm} &= Pr(Z_m > Z_{m'},\, Z_{m''}),\ m \neq m' \neq m'' = 1, 2, 3,\\ &= Pr\{(Z_m - Z_{m'} > 0) \cap (Z_m - Z_{m''} > 0)\}\\ &= Pr\{(W_{m'} > 0) \cap (W_{m''} > 0)\},\end{aligned}$$

where $(W_{m'},\, W_{m''})$ is bivariate normal with mean and covariance matrix obtained directly from $\mathbf{Z}_j$ and $\mathbf{V}$.

This means that the estimation of the probability that the j^{th} undeclared seat will go to Conservative, Labour or Liberal involves a bivariate normal numerical integration.

Other issues

Explanatory variables: Multicolinearity between explanatory variables is dealt with automatically; when few seats have declared some of the explanatory variables are so correlated that it is difficult with ordinary least squares to get reasonably precise estimates of the relative effects of the variables. The ridge modification shrinks parameter estimates towards zero in ill-conditioned directions. Brown and Payne (1984, [24]) state that ridge regression has also been shown to be very useful for prediction. The explanatory variables are scaled to have mean 0 and variance 1. The obvious but important point is made that we must be selective in choosing explanatory variables, or else if we try to include all the possible information then we will end up with a variety of nuisance parameters scattered throughout the model.

Probabilities: The BBC method assumes a multivariate normal distribution for the predictive probability that a seat is won by a party. This probability is then fed into the BBC's battleground computer graphic system to get the forecasted number of seats in the final outcome. Regarding the prior forecast, a covariance matrix is derived for the change in share, using past elections, which is used to obtain the probabilities. Therefore, the forecasted number of seats is a *weighted function* of the two types of probability:

$$\hat{p}_{jm} = w_n p_{jm} + (1 - w_n) q_{jm},$$

where p_{jm} = the predicted probability that the seat will go to party m and q_{jm} = the prior probability of the same event.

The more value believed to be in the prior information the larger the weight it is assigned (which is usually 0 to 3 seats). The prior probabilities are usually given larger weight for those constituencies in which declarations are not expected early. For instance, in the 1983 election the weight assigned to the predicted probability w_n, where n is the number of seats which have been declared, was $n/(n + w)$, and w was assigned the value 2 for England and Wales and 3 in Scotland. This means that the prior information is essentially equivalent to two declared seats in England and Wales and three in Scotland. Thus, for example, when two seats have actually been declared in England and Wales the prior and actual results get equal measure, but when twenty have been declared the former $(1 - w_{20})$ has far less influence. This prior forecast is discussed in full detail in Brown and Payne (1984). Brown and Payne (1984) point out that these prior probabilities are very important in the forecasting, not only due to their entertainment value for viewers, but also since they provide a partial solution to a key problem: declaration order is unrepresentative of all results, and so bias is present in the early results-based probabilities.

Special seats: Non-special seats may, if necessary, be transferred to the special type and therefore omitted from future regression. This happens where declared results in which at least one major party has a change in share of the vote which differs noticeably from the overall pattern of change for that party. Such a device traps gross errors in data capture and deals with outliers (Brown and Payne, 1984).

8.2.4 Inference

Confidence intervals

Brown and Payne (1975) suggest how one might go about getting confidence intervals for the forecasted number of seats, although they stress how the method is ad-hoc. One assumes that the forecasted number won by party g, N_g, follows a normal distribution. Then, all we need are the standard errors of N_g. If we define an indicator variable I_{jg}, to take value 1 if party g wins seat j, or 0 otherwise, (a Bernoulli model) then we end up with

$$\widehat{\mathbb{V}ar}(I_{jg}) = \hat{p}_{jg}(1-\hat{p}_{jg}).$$

Clearly

$$\mathbb{E}(N_g) = \sum_{j=1}^{635} \mathbb{E}(I_{jg}) = \sum_{j=1}^{635} P_{jg}.$$

The covariance between two indicator variables for a party is chosen according to previous election behaviour, but restricted to being monotonic linear decreasing using a simple piecewise function (linear decreasing for the first 100 seats, constant between 100 and 500 and linear decreasing to zero for the remainder). Together, we can then estimate

$$\mathbb{V}ar(N_g) = \sum_i \mathbb{V}ar(I_{jg}) + 2\sum_{i<j}\sum \mathbb{C}ov(I_{jg}, I_{kg})$$

and end up with the standard form of

$$\mathbb{E}(N_g) \pm 1.96 \times \sqrt{(\widehat{\mathbb{V}ar}(N_g))}$$

for a 95% confidence interval for the forecasted number of seats won by party g.

Prediction curves

When used for the October 1974 election, prediction curves seemed to behave very well; they were unstable early on, but soon settled, and were never far from the final result. For the 1983 election, for which there were several changes including the number of constituencies (see Section (2.4)), the updated model was adaptable enough to cope and work well. The revised 1997 model worked well, and was one of the best since regression models were introduced at the BBC.

Critical values

It is also possible to calculate probabilities of exceeding critical values of interest, such as the probability of gaining a majority.

8.2.5 Discussion

Revisions

In Brown, Firth and Payne (1999, [22]), the model was improved such that exit polls were designed to allow for differential refusal and ensured that the model was properly responsive to patterns emerging in the actual results. Basically the former was relieved by the guessing of experts. For the latter, there was effort on a particular exit poll design as well as changes to the design matrix, denote this by $\mathbf{X}$ for any party. The main change was to use the BBC exit poll to prespecify some tactical and regional regression coefficients.

Exit poll design: The first s columns of $\mathbf{X}$ are now 0-1 dummy variables corresponding to an s-category prior classification in order to identify likely cleavages in the pattern of change in share of vote. Let $\tilde{\beta} = (\beta_1, \beta_2, \ldots, \beta_s)'$, $s < r$ in Equation (8.3), which denote the mean changes in share of the vote for the prior categories. Then, the prior model is

$$\tilde{\beta} = \mathbf{b} - \alpha\mathbf{I} - \eta,$$

in which $\mathbf{b}$, obtained using the exit poll, estimates $\tilde{\beta} + \alpha$, with α a scalar representing systematic bias due to exit poll methodology (assumed common to each of the s components) and with $\eta_{s\times 1}$ a vector of exit poll sampling errors. For the 1997 election, there were four groups, that is, $s = 4$: 'Scotland' and then, for England and Wales, 'Conservative versus Labour', 'Conservatives versus Liberal Democrats' and 'Remainder'.

In Appendix A of Brown, Firth and Payne (1999) it is shown how to incorporate this prior design into the overall methodology by data augmentation. The conclusion is that the β coefficients may still be estimated by ordinary least squares and the estimates then used to predict the changes of share in vote in the undeclared seats.

Other updates: As well as exit polls, there were various other changes implemented in order to improve on the model and also to reflect the many changes (see Chapter 2, 2.4) which occurred. Some examples include minimising the number of special seats in order to predict winners in this category more accurately and greater care with the inclusion of explanatory variables which adjusted for differences, in Scotland particularly, since it took time to separate general election papers from the local election papers which occurred around the same time.

For further details of the methodology and performance, see Brown and Payne (1975 and 1984) and Brown, Firth and Payne (1999).

8.3 Forecasting Method

8.3.1 Two-party contest

We will now illustrate our forecasting method, initially in the context of a two-party contest and in the simplest-case scenario.

1. First we find q_j as described earlier. We employ these parameters in a vector of proportions

   $$\mathbf{q_j} = [q_j, 1 - q_j]',$$

   corresponding with the parties of interest, say Conservative and Labour, respectively, for each constituency j. Intuitively, at the time of the exit poll,

   $$Pr(\text{voting Conservative at constituency}j) = q_j,$$

   and the remainder vote Labour. Until now, we have assumed stationarity across the nation, but want to make the probability of voting Conservative constituency-specific, in order to correspond with how the UK is divided up in election terms.

2. We combine information contained within $\mathbf{q_j}$ with the *exit poll* sample size, to define prior probability densities for each constituency, for instance,

$$\pi(x_j) \sim Be(q_j N_j,\, (1 - q_j)N_j), \tag{8.4}$$

for $j = 1,\, 2,\, \ldots,\, 659$ (i.e. for 659 constituencies), and where N_j denotes the total sample size of the exit poll for constituency j. Here, we may infer that the proportion q_j distributes the total sample size for the exit poll amongst the two shape parameters, the former of which represents Conservative and the latter Labour. The fact that these sample sizes will differ among constituencies provides variation of probability densities as required. The next task is to update these densities when seats are actually declared throughout election night, taking into account all the known seat outcomes all the while. We now make use of the regression techniques discussed in Section 8.2. This requires us to observe a sufficient number of declarations, which depends upon the number of explanatory variables which we wish to consider. As mentioned, this is to ensure that the regression is stable enough to provide reliable estimates. In the easiest case, suppose that we only wish to consider one explanatory variable, the number who voted Conservative in the last election, say. Then we must observe at least five declarations. Subsequently, we forecast the remainder *in sequence*, starting with the sixth, before the sixth is actually declared. When the sixth is declared we simply repeat this whole method but now obviously using the sixth declared result rather than what we predicted it to be. We aim to state an accurate prediction and as early as possible in the election night.

   (a) We regress using the currently declared seats as our response and those corresponding from the previous election as our regressor:

   $$y_j = \beta_0 + y_{j(old)}\beta_1 + \epsilon_j,$$

   for $j = 1, 2, 3, 4, 5$, in order to find the least-square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, and also $\hat{\sigma}^2$.

   (b) From this we compute the predicted total number of Conservative votes for seat 6, $\tilde{y}_6$. As with usual linear regression, we assume normality among the errors so that here

   $$\tilde{y}_6 \sim N(\hat{\beta}_0 + y_{6(old)}\hat{\beta}_1,\, \hat{\sigma}^2), \tag{8.5}$$

   thereby implying that we would *expect* $\tilde{y}_6$ to be somewhere around the point $\hat{\beta}_0 + y_{6(old)}\hat{\beta}_1$.

(c) The fundamental part of the forecasting involves obtaining the expected - or predicted - probability of voting given the declared seats. For each constituency undeclared, we must update the prior to take account of the known results. Our model makes use of the relationship

$$\pi(x_j|\,\text{history}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi(x_j|\,\tilde{y}_j,\,\tilde{r}_j)\pi(\tilde{y}_j|\,\text{history}) \times \pi(\tilde{r}_j|\,\text{history})\,\mathrm{d}\tilde{y}_j\mathrm{d}\tilde{r}_j,$$

which translates into

$$\pi(x_6|\,y_1,\ldots,y_5) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Be(a,b;\,x_6,\tilde{y}_6,\tilde{r}_6)N(\mu_1,\,\sigma_1^2;\,\tilde{y}_6,y_1,\ldots,y_5) \times N(\mu_2,\,\sigma_2^2;\,\tilde{r}_6,r_1,\ldots,r_5)\,\mathrm{d}\tilde{y}_6\mathrm{d}\tilde{r}_6.$$

This shows that we must integrate out the forecasted data in order essentially to get a density which only considers the seats declared. Note that here we are doing a separate regression forecast for the next total number of votes $\tilde{r}_6$. This takes a similar form to that for $\tilde{y}_6$, i.e.

$$\tilde{r}_6 \sim N(\hat{\beta}_{(r)0} + r_{6(old)}\hat{\beta}_{(r)1},\,\hat{\sigma}^2_{(r)}). \tag{8.6}$$

We do not need to state explicitly a regression model for the Labour votes, since we obtain this by subtraction of (8.6) and (8.5); dropping one of the groups is standard practice in systems of such regression equations. Clearly, we may then obtain proportions, y/r, which are more useful for comparison between constituencies than looking at just y alone (as r varies between constituencies).

First of all assuming that the double integral were solvable, we would then calculate our predicted probability via

$$\mathbb{E}(x_6|\,\text{history}) = \int_0^1 x_6\pi(x_6|\,\text{history})\,\mathrm{d}x_6,$$

whose limits are $[0,1]$ since $\pi(x_j|\,\text{history})$ is a probability.

We have yet to explain the beta density above. We want to incorporate the type of modelling which we implemented on the opinion poll data, sequentially throughout election night. Of course, this is in addition to including the necessary $\tilde{y}_6$ and $\tilde{r}_6$. Therefore, we choose a form of $\pi(x_6|\,\tilde{y}_6,\,\tilde{r}_6)$ such that we satisfy

$$\pi(x_6|\,\tilde{y}_6,\,\tilde{r}_6) \propto \pi(\tilde{y}_6|\,x_6)\pi(x_6|\,y_5),$$

the right-hand side of which is derived from our specification of the general model for the opinion poll data; this is covered in Chapter 6. If we have that

$$\begin{aligned}\pi(y_6|\, x_6) &\sim Bin(\tilde{r}_6|\, x_6)\\ \pi(x_6|\, y_5) &\sim Be(q_6 N_6 + y_5,\, (1-q_6)N_6 + r_5 - y_5),\end{aligned}$$

in which $\tilde{r}_6$ is the total number of votes in the next constituency, then a possible choice would be

$$\pi(x_6|\, \tilde{y}_6,\, \tilde{r}_6) \sim Be(q_6 N_6 + \tilde{y}_6,\, (1-q_6)N_6 + \tilde{r}_6 - \tilde{y}_6), \tag{8.7}$$

since

$$\frac{x_6^{q_6 N_6 + \tilde{y}_6 - 1}(1-x_6)^{(1-q_6)N_6+\tilde{r}_6-\tilde{y}_6-1}}{B(q_6 N_6 + \tilde{y}_6,\, (1-q_6)N_6 + \tilde{r}_6 - \tilde{y}_6)} \propto \frac{\binom{\tilde{r}_6}{\tilde{y}_6} x_6^{q_6 N_6 + y_5 + \tilde{y}_6 - 1}(1-x_6)^{(1-q_6)N_6 + r_5 - y_5 + \tilde{r}_6 - \tilde{y}_6 - 1}}{B(q_6 N_6 + y_5,\, (1-q_6)N_6 + r_5 - y_5)}$$

with proportionality constant

$$\frac{\binom{\tilde{r}_6}{\tilde{y}_6} x_6{}^{y_5}(1-x_6)^{r_5-y_5} B(q_6 N_6 + \tilde{y}_6,\, (1-q_6)N_6 + \tilde{r}_6 - \tilde{y}_6)}{B(q_6 N_6 + y_5,\, (1-q_6)N_6 + r_5 - y_5)}.$$

Recall also that

$$\mathbb{E}(x_6|\, \tilde{y}_6,\, \tilde{r}_6) = \frac{q_6 N_6 + \tilde{y}_6}{N_6 + \tilde{r}_6},$$

and so, more generally, $\mathbb{E}(x_j|\,\text{history})$ may be expressed as

$$\begin{aligned}\mathbb{E}(x_j|\,\text{history}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} &\mathbb{E}(x_j|\, \tilde{y}_j,\, \tilde{r}_j)\pi(\tilde{y}_j|\,\text{history})\\ &\times \pi(\tilde{r}_j|\,\text{history})\, \mathrm{d}\tilde{y}_j \mathrm{d}\tilde{r}_j,\end{aligned}$$

as the latter two functions do not contain x_j.

The problem is that the double integral above proves difficult to compute on packages such as Maple and MATLAB. An alternative is to sample from all the necessary distributions and then to obtain the mean of the sample realisations. Therefore, here we would simulate from models (8.5) and (8.6) to obtain $\tilde{y}_6$ and $\tilde{r}_6$ respectively, and subsequently use them in $Be(q_6,\, N_6;\, \tilde{y}_6, \tilde{r}_6)$, from which we simulate. We would then repeat this process many times, say, $K = 10000$ and instead obtain

$$\mathbb{E}(x_6|\,\text{history}) = \frac{1}{K}\sum_{k=1}^{K} \tilde{x}_6^{(k)}.$$

Compared with the prior probabilities earlier (8.4), we might *alternatively* introduce a weight $0 \leq u \leq 1$, in order to retain use of the exit poll information. Then we would end up with

$$\pi(x_j | \, \tilde{y}_6, \, \tilde{r}_6) \sim Be(uq_6N_6 + (1-u)\tilde{y}_6, \, u(1-q_6)N_6 + (1-u)(\tilde{r}_6 - \tilde{y}_6)). \tag{8.8}$$

The motivation behind this would be if the exit poll data are hard to obtain, so that we could still continue our overall approach (setting $u = 0$).

Here, *we* would choose the value of u; we could even let it vary, u_j, for constituency $j = 1, 2, \ldots, 659$, such that it has decreasing impact as $j \rightarrow 659$: for instance, we know that u_0 (before the first seat is declared) will equal 1; by the time of, say, the 600^{th} declaration u_{601} may have been reduced to 0.40. A possible model in this case might be

$$u_j = \begin{cases} \frac{a}{a+j} & j = 0, 1, 2, \ldots, 659 \\ 0 & \text{otherwise,} \end{cases}$$

for some large positive constant a such as 450, such that when $j = 0$ (no seats have declared) the exit poll has the only influence on the probability of voting, whereas $\lim_{j \to 659} u_j = 0.4$ say, so that by the time that all seats have declared the exit poll only has 40% of the influence. Several possibilities for u_j are possible, a simpler one of which would be $u_j = 1 - (3/3295)j$ for $j = 0, 1, 2, \ldots, 659$, which is linear and decreases much more slowly than the previous example. Of course, our control of the weight is governed by how influential we regard the exit polls to be relative to the actual declarations, which may change through the night.

Assuming that all exit polls are conducted satisfactorily, we might expect the information contained within them to be correlated positively with that within the actual voting patterns. Therefore, an arrangement like (8.8) would not 'hide' away information. For instance, if for a given constituency the exit polls suggest that Labour will win in the end, and since the exit polls are carried out close in time to election night, then we would expect the actual votes to reflect this.

(d) The product of $\tilde{r}_6$ and $\mathbb{E}(x_6 | \, \text{history})$ determines $\hat{y}_6$, the number of predicted votes for Conservative (and thus the remainder $\tilde{r}_6 - \hat{y}_6$ are predicted to go to Labour).

(e) Finally, the seat is won by the party with the greatest number of votes. The program to do such forecasting is outlined in Appendix A.

This is the simplest case for ease of explanation. We are actually interested in three parties and a somewhat more complex model for the opinion poll data. Furthermore, we would like to incorporate more explanatory variables into our regression modelling.

Remark: Whiteley (2005, [109]) points out how the forecasting of elections has in the past tended to focus on votes won rather than seats, whereas it is the latter which determines which party wins overall. The advantage of our approach to forecasting is that we actually take both the voting numbers *and* the number of seats, with the underlying interest being in the latter, just as is done on election night.

8.3.2 Three-party contest

Brown and Payne (1975) discuss the splitting of constituencies into those in which there tends to be a two-party contest and those in which there tends to be a three-party contest. However, in Chapter 2 we explained how - since that publication - the Liberal/Alliance (now evolved into the Liberal Democrat) party has led to a three-party contest almost nationally (forgetting nationalist parties). The Liberal Democrats now stands for election in almost all constituencies, and so we may focus just on three-party modelling. Where nationalist parties are popular also, it is straightforward to extend our modelling to have four parties, or indeed any number of parties.

The model

Exit polls: We now have that

$$\pi\left(x_{1(j)}, x_{2(j)}\right) \sim Dir\left(q_{1j}N_j,\ q_{2j}N_j,\ q_{3j}N_j\right), \tag{8.9}$$

in which for $m = 1,\ 2,\ 3$ we set

$$q_{mj} = \frac{y_{mj}}{N_j},$$

where j is the constituency number. N_j is the sample size for the exit poll conducted at constituency j and y_{mj} the number of votes in the exit poll at constituency j for party m.

Then, the probabilities given all known history make use of the following:

$$\pi\left(x_{1(j)}, x_{2(j)} \middle| \tilde{y}_{1(j)}, \tilde{y}_{2(j)} \tilde{r}_{1(j)}, \tilde{r}_{2(j)}, \tilde{\delta}_{1(j)}, \tilde{\delta}_{2(j)}\right) \sim Dir\left(b_1, b_2, b_3\right), \qquad (8.10)$$

where for $m = 1, 2, 3$

$$b_m = \tilde{y}_{m(j)} + q_{mj} N_j. \qquad (8.11)$$

Earlier, we mentioned that we may incorporate the exit poll information via a weight function, u, say. Hence, we may choose to extend b_m in (8.11) to

$$b_m = u \cdot \tilde{y}_{m(j)} + (1 - u) \cdot \left(q_{mj} N_j\right). \qquad (8.12)$$

By this, we are able to let u increase as more seats declare, thereby effectively shrinking the influence of the exit polls.

Because the elections which we have used for examples (Chapter 7) are historical, we have access to all the number of votes broken down by constituency. This permits us to perform the regression-based forecast as if we did not know the outcome and then at the same time compare how close we are to the final outcome. What we, however, do not have are exit poll data to carry out the forecasting using b_m as in (8.11). We would hope that such information would become available around the time of the election. Nevertheless, it is suffice to say that having an arrangement such as (8.9) makes sense and fits in neatly with our modelling both of opinion polls and through election night. Therefore, we can progress using the special case of (8.11) which is (8.12), obviously setting $u = 1$.

Explanatory variables: Following the BBC approach, we increase the number of explanatory variables in our regressions as we have more data, according to requiring $5 + 3(v - 1)$ observations in order to regress with v variables. This means that the model may only get more realistic when we have more observations. Hence we must observe the first five seat declarations and then forecast the remainder, each in sequence, starting with the sixth, $\mathbf{y_6} = (y_{1(6)}\ \ y_{2(6)}\ \ y_{3(6)})'$. In Brown and Payne (1975) the first explanatory variable is that party's vote count (in that seat only) in the previous election.

The regression: Again, we let y_1 denote Conservative, y_2 Labour and y_3 Liberal/Alliance. Similar to before, we have

$$\tilde{r}_6 \sim N(\hat{\beta}_{0r} + \hat{\beta}_{1r} r_{6(old)}, \hat{\sigma}_r^2),$$

but now also require an estimated time when the next seat will declare and so we regress such that

$$\tilde{\delta}_{16} \sim N(\hat{\beta}_{0\delta} + \hat{\beta}_{1\delta} \delta_{16(old)}, \hat{\sigma}_\delta^2).$$

Finally, we have the multivariate normal distribution

$$\mathbf{\tilde{y}_6} = \begin{pmatrix} \tilde{y}_{1(6)} \\ \tilde{y}_{2(6)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \hat{\beta}_{0c} + \hat{\beta}_{1c} y_{1(6(old))} \\ \hat{\beta}_{0l} + \hat{\beta}_{1l} y_{2(6(old))} \end{pmatrix}, \Sigma \right),$$

where

$$\Sigma = \begin{pmatrix} \mathbb{V}ar(y_{16)}) & \mathbb{C}ov(y_{16)}, y_{26)}) \\ \mathbb{C}ov(y_{16)}, y_{26)}) & \mathbb{V}ar(y_{26)}) \end{pmatrix}.$$

Predictive probabilities: Using the above, we find our predictive probability density via[1]

$$\pi(x_{16}, x_{26} | \, \text{history}) = \int_{\infty}^{\infty} \int_{\infty}^{\infty} \int_{\infty}^{\infty} \pi \left(x_{1(6)}, x_{2(6)} | \, \tilde{y}_{1(6)}, \, \tilde{y}_{2(6)} \, \tilde{r}_{1(6)}, \, \tilde{r}_{2(6)}, \, \tilde{\delta}_{1(6)}, \, \tilde{\delta}_{2(6)} \right)$$
$$\times \, \pi(\mathbf{\tilde{y}_6} | \, \text{history}) \pi(\tilde{r}_6 | \, \text{history}) \pi(\tilde{\delta}_6 | \, \text{history}) \, \mathrm{d}\mathbf{\tilde{y}}_6 \, \mathrm{d}\tilde{r}_6 \, \mathrm{d}\tilde{\delta}_6,$$

whose expectation we may approximate by sampling from $\mathbf{\tilde{y}_6}$, $\tilde{r}_6$ and $\tilde{\boldsymbol{\delta}}_\mathbf{6} = (\delta_{1(6)} \;\; \delta_{2(6)})'$ and then using them in $\pi \left(x_{1(6)}, x_{2(6)} | \, \tilde{y}_{1(6)}, \, \tilde{y}_{2(6)} \, \tilde{r}_{1(6)}, \, \tilde{r}_{2(6)} \, \tilde{\delta}_{1(6)}, \, \tilde{\delta}_{2(6)} \right)$ from which we also sample. We repeat this 10000 times, say, and calculate the average of the samples of both $x_{1(6)}$ and $x_{2(6)}$.

Predicted votes: The predicted number of votes, thus is found using

$$\mathbf{\hat{y}_6} = \tilde{r}_6 \cdot \mathbb{E}(\mathbf{x_6} | \, \text{history}),$$

for $\mathbf{x_6} = (x_{1(6)} \;\; x_{2(6)} \;\; x_{3(6)})'$ and the winner of seat 6 is the party with the most votes. Recall that we do not need to regress the Liberal/Alliance voting, since we obtain this by subtracting the Conservative and Labour votes from the total votes.

Process: We then treat $\mathbf{\hat{y}_6}$ as if it were the declared result and next forecast the remaining undeclared seat outcomes in sequence. From this we have our first final forecasted outcome. When $\mathbf{y_6}$ actually is declared we revise $\mathbf{\hat{y}_7}$ onwards and follow this routine as each seat is declared. When we have seen eight declarations we may include our second explanatory variable, and so we have a new regression parameter for the r and $\mathbf{y}$ models:

$$\tilde{r}_j \sim N(\hat{\beta}_{0r} + \hat{\beta}_{1r} r_{j(old)} + \hat{\beta}_{2r} r_{j(old)}, \hat{\sigma}_r^2),$$

$$\mathbf{\tilde{y}_j} = \begin{pmatrix} \tilde{y}_{1(j)} \\ \tilde{y}_{2(j)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \hat{\beta}_{0c} + \hat{\beta}_{1c} y_{1(j(old))} + \hat{\beta}_{2c} y_{1(j(old2))} \\ \hat{\beta}_{0l} + \hat{\beta}_{1l} y_{2(j(old))} + \hat{\beta}_{2l} y_{2(j(old2))} \end{pmatrix}, \Sigma \right);$$

[1] When we say 'history' we mean any relevant known information used in the specific modelling, $r_1, r_2, \ldots, r_5$ for $\tilde{r}_6$, for example.

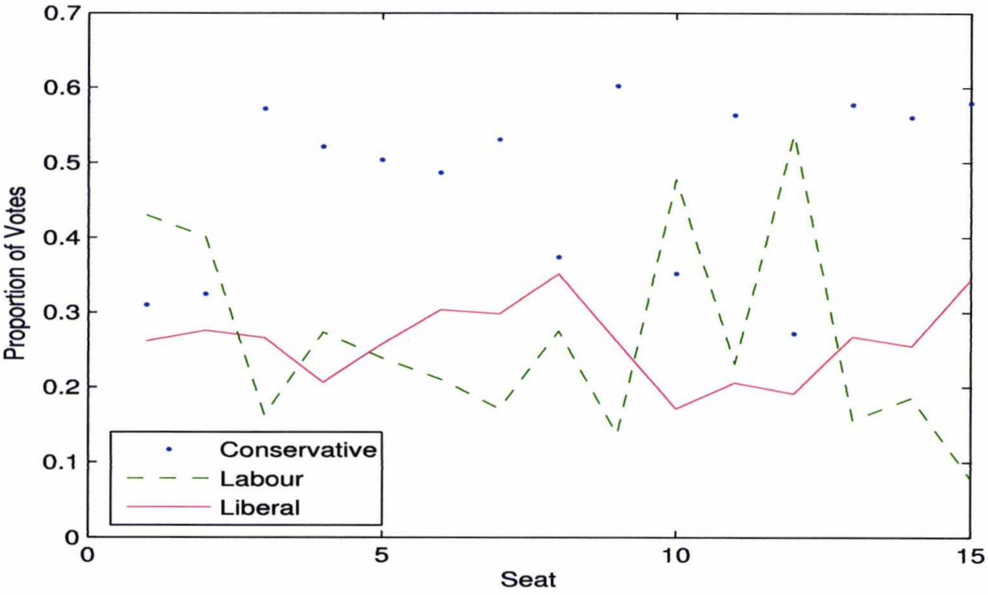


Figure 8.1: Proportion of votes for each main party per seat in the truncated 1983 election.

these variables have a particular order of inclusion. In Brown and Payne (1975) the second variable included was the number of votes (in that seat) in the previous election of that party's main rival party. In the $\tilde{r}_j$ model above we would have here that $\hat{\beta}_{1r} = \hat{\beta}_{2r}$ since $r_{j(old)}$ appears twice, only due to the fact that the second explanatory variable introduced here requires the same total as the first, that is, because it also concerns the *last* election. If, otherwise, the second explanatory variable were, say, a dummy variable such as a nationalist party's intervention, then we would not have $r_{j(old)}$ again because it does not concern the last election, and thus $\hat{\beta}_{1r} \neq \hat{\beta}_{2r}$.

We want our forecast of the final outcome at each stage of 'waiting' to get closer and closer to the actual outcome, which we know to be (397, 209, 23) for 1983 and (376, 229, 22) for 1987.

8.4 Illustration

We now demonstrate how our approach ties in with the regression just described, by forecasting the first fifteen seat declarations. We imagine that there are only fifteen seats in the UK, and suppose for convenience that these come from the London Boroughs. Obviously we know the results already and Figures 8.1 and 8.2 show the proportions of votes for each party per seat. We know that in 1987

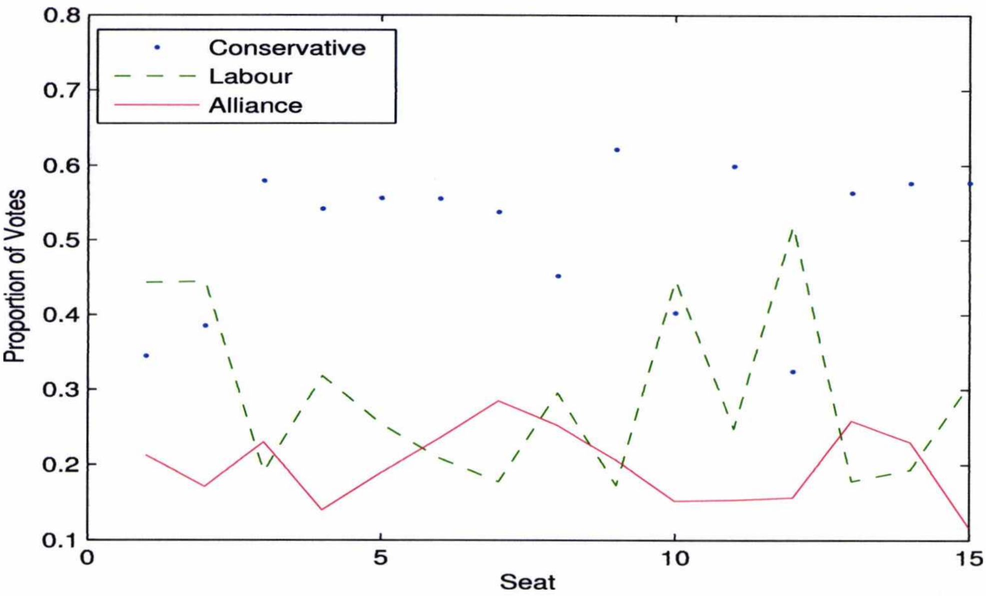


Figure 8.2: Proportion of votes for each main party per seat in the truncated 1987 election.

of these seats, Conservative won 11, Labour 4 and Alliance 0. (Proportionately this is different from the final result for the UK.) Table 8.2 shows the number of votes won in 1987 while Table 8.1 shows that in 1983 as they are involved in the explanatory variables.

8.4.1 Analysis

We must observe the first five declarations:

1. Barking and Dagenham, Barking: Winner LABOUR (0, 1, 0)
2. Barking and Dagenham, Dagenham: Winner LABOUR (0, 2, 0)
3. Barnet, Chipping Barnet: Winner CONSERVATIVE (1, 2, 0)
4. Barnet, Finchley: Winner CONSERVATIVE (2, 2, 0)
5. Barnet, Hendon North: Winner CONSERVATIVE (3, 2, 0)

While waiting for both seats six and seven, we make forecasts of the nine/eight undeclared seats using only one explanatory variable: that party's number of votes in the 1983 election:

- Barnet, Hendon South: Prediction (3, 2, 10) ALLIANCE to win **overall**

Seat No.	Name	Conservative	Labour	Liberal	Total
1	Barking and Dagenham, Barking	10389	14415	8770	33574
2	Barking and Dagenham, Dagenham	12668	15665	10769	39102
3	Barnet, Chipping Barnet	23164	6599	10771	40534
4	Barnet, Finchley	19616	10302	7763	37681
5	Barnet, Hendon North	18499	8786	9474	36759
6	Barnet, Hendon South	17115	7415	10682	35212
7	Barnet, Bexleyheath	23411	7560	13153	44124
8	Barnet, Erith and Crayford	15289	11260	14369	40918
9	Bexley, Old Bexley and Sidcup	22422	5116	9704	37242
10	Brent, Brent East	13529	18363	6598	38490
11	Brent, Brent North	24842	10191	9082	44115
12	Brent, Brent South	10740	21259	7557	39556
13	Bromley, Beckenham	23603	6386	10936	40925
14	Bromley, Chislehurst	22108	7320	10047	39475
15	Bromley, Orpington	25569	3439	15148	44156

Table 8.1: Vote share in the truncated 1983 election.

- Barnet, Bexleyheath: Prediction (4, 2, 9) ALLIANCE to win overall

Then, we introduce the second explanatory variable, Labour (the main rival for Conservative) votes in the 1983 election, Conservative (the main rival for Labour) votes in the 1983 election and Labour (the main rival for Alliance) votes in the 1983 election:

- Barnet, Erith and Crayford: Prediction (5, 2, 8) ALLIANCE to win overall
- Bexley, Old Bexley and Sidcup: Prediction (6, 2, 7) ALLIANCE to win overall
- Brent, Brent East: Prediction (7, 2, 6) CONSERVATIVE to win overall
- Brent, Brent North: Prediction (7, 3, 5) CONSERVATIVE to win overall
- Brent, Brent South: Prediction (8, 3, 4) CONSERVATIVE to win overall
- Bromley, Beckenham: Prediction (8, 4, 3) CONSERVATIVE to win overall
- Bromley, Chislehurst: Prediction (9, 4, 2) CONSERVATIVE to win overall

Seat No.	Name	Conservative	Labour	Liberal	Total
1	Barking and Dagenham, Barking	11898	15307	7336	34541
2	Barking and Dagenham, Dagenham	15985	18454	7088	41527
3	Barnet, Chipping Barnet	24686	8115	9815	42616
4	Barnet, Finchley	21603	12690	5580	39873
5	Barnet, Hendon North	20155	9223	6859	36237
6	Barnet, Hendon South	19341	7261	8217	34819
7	Barnet, Bexleyheath	24866	8218	13179	46263
8	Barnet, Erith and Crayford	20203	13209	11300	44712
9	Bexley, Old Bexley and Sidcup	24350	6762	8076	39188
10	Brent, Brent East	15119	16772	5710	37601
11	Brent, Brent North	26823	11103	6868	44794
12	Brent, Brent South	13209	21140	6375	40724
13	Bromley, Beckenham	24903	7888	11439	44230
14	Bromley, Chislehurst	24165	8115	9658	41938
15	Bromley, Orpington	27261	14486	5512	47259

Table 8.2: Vote share in the truncated 1987 election.

- Bromley, Orpington (final seat): Prediction (10, 4, 1) CONSERVATIVE to win overall

8.4.2 Inference

We see how the forecast greatly overestimates Alliance early on, with Conservative in second place and Labour third. This swings to Conservative taking the lead by the time the ninth seat has declared, and putting (and retaining) Labour correctly in second place. Plots of the forecasting accuracy for each party are shown in Figure 8.3. Clearly we only have a very small number of seats and so it is hard to see how accurate how method is. Nevertheless, we can see that in all cases the accuracy improves gradually as more seats declare. Ideally, when extended to the UK as a whole we expect that the forecast will tend towards the final outcome early on in order that we may state a final forecast equally early.

Table 8.3 shows for each party the predictive probabilities of voting at each stage of waiting for the next seat to declare. From this we can see that the predictions get better as more seats declare, and correctly predict certain important features, for instance, the domination of Labour in Brent South. Because

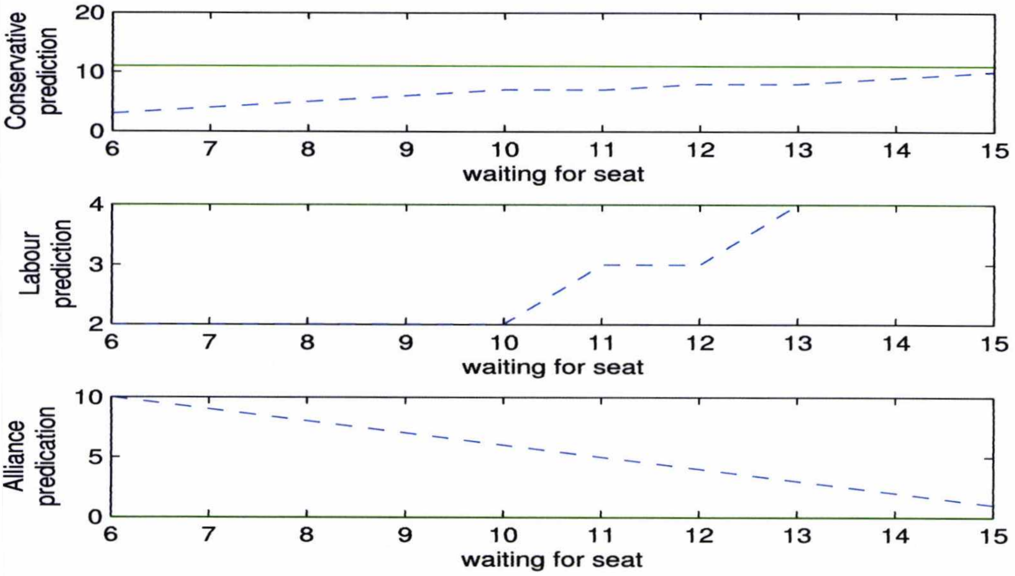


Figure 8.3: Revised forecasts per party in the truncated 1987 election compared with the final outcome.

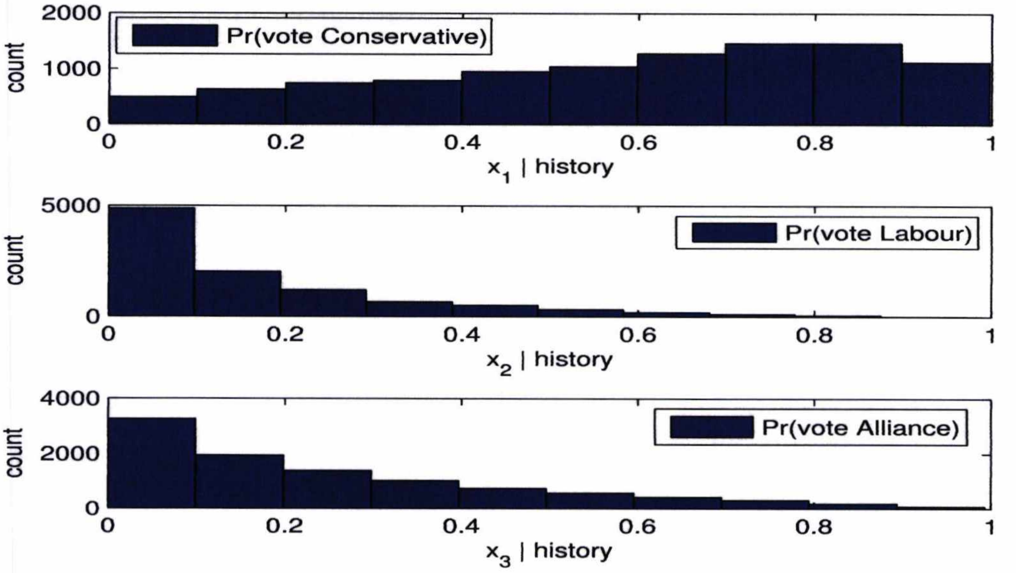


Figure 8.4: Histograms of sample predictive probabilities of voting each party in the truncated 1987 election given 14 declared seats.

of the fact that we sampled from various distributions in order to arrive at these probabilities, due to the difficulty in evaluating it exactly, it is important to check whether the realisations of the inevitable probabilities take the shape of the Dirichlet distribution. Figure 8.4 shows histograms of the 10000 samples for each party while waiting for the final seat to declare; these are typical spreads at any stage of waiting. We see that these form the shapes expected from Dirichlet densities. Here, the negative skewness of Conservative realisations is to be expected due to the party's popularity in the 1987 election. Consequently we have positive skewness for the realisations of Labour and Alliance, whose shapes look similar to each other. For information, also shown in Table 8.4 are regression-based estimates of how long we must wait until the next declarations. This gives us generally the kinds of waiting times which we expect.

Table 8.5 shows the regression parameters including the estimated variances for each of the parties and also for the time difference between declarations.

An obvious limitation is that in practice some constituencies are known typically to declare later than others, for instance, those in which Alliance won in 1987. Our truncated version has not allowed for this, which explains why Alliance has won no seats. Recall also that while waiting for seat 14 to declare we should ideally introduce the third variable, whereas we still only have two. Because of its simplification to only a fraction of the seats, there is no scope or flexibility for our forecast to smoothen itself out over time. When used to forecast in a more realistic context we expect that the forecasting error like that shown in Figure 8.3 will be large as early on as what our illustration above shows but would become more and more modest later on, such that we may soon get similar results each time a new seat declares and we update the forecast.

8.5 Review

To summarise, we have introduced our own new method, using a simple scenario for illustration. Our methodology from Chapter 6 is used inside the BBC's regression framework, of which we provided a detailed review. Overall, our approach attractively has several of the benefits of the other approaches considered, and often improves on them. We focused on the three-party case and with the 1987 data, the interest now being in the forecasting on election night itself. We detailed the approach, which is based on the BBC's adopted method but with our new contribution of probability modelling at each stage. Next we illustrated the approach by looking at the first fifteen declarations of seats. Our findings indicate that, after somewhat poor forecasts initially, the fluctuations stretch and get smaller as more seats declare and we have more data in our model. Subsequently, the method would be accurate and thus useful if implemented on the night. Furthermore, if we had the exit poll data per constituency then we would expect our method to be more accurate.

Waiting for seat	6	7	8	9	10	11	12	13	14	15
6	(0.46, 0.29, 0.25)	(0.46, 0.26, 0.28)	(0.42, 0.55, 0.03)	(0.48, 0.26, 0.26)	(0.39, 0.48, 0.13)	(0.44, 0.32, 0.23)	(0.39, 0.55, 0.06)	(0.47, 0.27, 0.27)	(0.44, 0.33, 0.22)	(0.45, 0.20, 0.35)
7	- -	(0.47, 0.25, 0.28)	(0.41, 0.30, 0.29)	(0.49, 0.23, 0.28)	(0.39, 0.43, 0.19)	(0.46, 0.30, 0.24)	(0.34, 0.50, 0.16)	(0.46, 0.25, 0.29)	(0.46, 0.28, 0.27)	(0.47, 0.17, 0.37)
8	- -	- -	(0.48, 0.33, 0.19)	(0.45,0.24, 0.31)	(0.40, 0.39, 0.21)	(0.41, 0.28, 0.31)	(0.31, 0.41, 0.28)	(0.50, 0.26, 0.25)	(0.47, 0.27, 0.25)	(0.55, 0.24, 0.20)
9	- -	- -	- -	(0.40, 0.24, 0.36)	(0.39, 0.39, 0.22)	(0.35, 0.28, 0.38)	(0.35, 0.40 0.24)	(0.45, 0.26, 0.30)	(0.43, 0.27, 0.30)	(0.55, 0.26, 0.19)
10	- -	- -	- -	- -	(0.45, 0.40, 0.15)	(0.42, 0.27, 0.31)	(0.37, 0.41, 0.22)	(0.49, 0.25, 0.26)	(0.47, 0.26, 0.26)	(0.52, 0.25, 0.23
11	- -	- -	- -	- -	- -	(0.41, 0.28, 0.31)	(0.31, 0.42, 0.27)	(0.49, 0.25, 0.27)	(0.47, 0.27, 0.25)	(0.59, 0.22, 0.26)
12	- -	- -	- -	- -	- -	- -	(0.29, 0.41, 0.30)	(0.55, 0.25, 0.20)	(0.52, 0.27, 0.21)	(0.58, 0.22, 0.20)
13	- -	- -	- -	- -	- -	- -	- -	(0.56, 0.25, 0.20)	(0.52, 0.28, 0.20)	(0.61, 0.20, 0.20)
14	- -	- -	- -	- -	- -	- -	- -	- -	(0.52, 0.27, 0.21)	(0.58, 0.18, 0.24)
15	- -	- -	- -	- -	- -	- -	- -	- -	- -	(0.57, 0.17, 0.26)

Table 8.3: Predictive probabilities in the truncated 1987 election.

Waiting for seat	6	7	8	9	10	11	12	13	14	15
6	0.12	0.04	0.01	0.03	0.04	0.01	0.05	0.02	0.02	0.02
7	-	0.25	0.05	0.02	0.04	0.05	0.02	0.04	0.03	0.04
8	-	-	0.02	0.05	0.05	0.02	0.06	0.04	0.03	0.04
9	-	-	-	0.05	0.05	0.02	0.06	0.04	0.03	0.04
10	-	-	-	-	0.05	0.03	0.07	0.04	0.04	0.05
11	-	-	-	-	-	0.02	0.06	0.04	0.04	0.04
12	-	-	-	-	-	-	0.06	0.04	0.03	0.04
13	-	-	-	-	-	-	-	0.04	0.03	0.03
14	-	-	-	-	-	-	-	-	0.03	0.03
15	-	-	-	-	-	-	-	-	-	0.03

Table 8.4: Predictive waiting time (hours) for declarations in the truncated 1987 election.

Waiting for seat	Conservative	Labour	Total	δ
6	$\hat{\beta}_0$=1911.8, $\hat{\beta}_1$=0.4, $\hat{\sigma}^2$=5461800	$\hat{\beta}_0$=626.7, $\hat{\beta}_1$=0.7, $\hat{\sigma}^2$=1396600	$\hat{\beta}_0$=1419.1, $\hat{\beta}_1$=0.5, $\hat{\sigma}^2$=17883000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$ = 0.1, $\hat{\sigma}^2$=0.02
7	$\hat{\beta}_0$=2279.1, $\hat{\beta}_1$=0.5, $\hat{\sigma}^2$=5278600	$\hat{\beta}_0$=896.8, $\hat{\beta}_1$=0.7, $\hat{\sigma}^2$=1397000	$\hat{\beta}_0$=1934.8, $\hat{\beta}_1$=0.6, $\hat{\sigma}^2$=17117000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$=0.1, $\hat{\sigma}^2$=0.02
8	$\hat{\beta}_0$=1562.1, $\hat{\beta}_1$=0.5, $\hat{\beta}_2$=0.1, $\hat{\sigma}^2$=5899000	$\hat{\beta}_0$=525.1, $\hat{\beta}_1$=0.6, $\hat{\beta}_2$=0.1, $\hat{\sigma}^2$=1516000	$\hat{\beta}_0$=779.7, $\hat{\beta}_1$=1.1, $\hat{\beta}_2$=-0.4, $\hat{\sigma}^2$=17020000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$ = 0.1, $\hat{\sigma}^2$=0.02
9	$\hat{\beta}_0$=1336.7, $\hat{\beta}_1$=0.2, $\hat{\beta}_2$=0.8, $\hat{\sigma}^2$=6039700	$\hat{\beta}_0$=19.4, $\hat{\beta}_1$=0.6, $\hat{\beta}_2$=0.3, $\hat{\sigma}^2$=1420500	$\hat{\beta}_0$=563.2, $\hat{\beta}_1$=1.0, $\hat{\beta}_2$=-0.2, $\hat{\sigma}^2$=16874000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$ = 0.1, $\hat{\sigma}^2$=0.02
10	$\hat{\beta}_0$=1354.8, $\hat{\beta}_1$=0.6, $\hat{\beta}_2$=0.4, $\hat{\sigma}^2$=5627200	$\hat{\beta}_0$=120.4, $\hat{\beta}_1$=0.6, $\hat{\beta}_2$=0.3, $\hat{\sigma}^2$=1417500	$\hat{\beta}_0$=844.2, $\hat{\beta}_1$=0.9, $\hat{\beta}_2$=-0.1, $\hat{\sigma}^2$=16958000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$ = 0.1, $\hat{\sigma}^2$=0.02
11	$\hat{\beta}_0$=1985.9, $\hat{\beta}_1$=0.6, $\hat{\beta}_2$=0.3, $\hat{\sigma}^2$=5414200	$\hat{\beta}_0$=208.3, $\hat{\beta}_1$=0.8, $\hat{\beta}_2$=0.1, $\hat{\sigma}^2$=725090	$\hat{\beta}_0$=641.9, $\hat{\beta}_1$=1.5, $\hat{\beta}_2$=-0.6, $\hat{\sigma}^2$=12584000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$ = 0.1, $\hat{\sigma}^2$=0.02
12	$\hat{\beta}_0$=1886.3, $\hat{\beta}_1$=1.0, $\hat{\beta}_2$=-0.0, $\hat{\sigma}^2$=4654500	$\hat{\beta}_0$=294.6, $\hat{\beta}_1$=0.9, $\hat{\beta}_2$=0.1, $\hat{\sigma}^2$=675740	$\hat{\beta}_0$=150.4, $\hat{\beta}_1$=1.6, $\hat{\beta}_2$=-0.7, $\hat{\sigma}^2$=11444000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$ = 0.1, $\hat{\sigma}^2$=0.02
13	$\hat{\beta}_0$=2140.4, $\hat{\beta}_1$=1.0, $\hat{\beta}_2$=-0.0, $\hat{\sigma}^2$=4640400	$\hat{\beta}_0$=136.9, $\hat{\beta}_1$=1.0, $\hat{\beta}_2$=0.01, $\hat{\sigma}^2$=262380	$\hat{\beta}_0$=67.6, $\hat{\beta}_1$=1.8, $\hat{\beta}_2$=-0.8, $\hat{\sigma}^2$=10364000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$ = 0.1, $\hat{\sigma}^2$=0.02
14	$\hat{\beta}_0$=1942.5, $\hat{\beta}_1$=1.1, $\hat{\beta}_2$=-0.0, $\hat{\sigma}^2$=4325500	$\hat{\beta}_0$=158.9, $\hat{\beta}_1$=1.0, $\hat{\beta}_2$=0.02, $\hat{\sigma}^2$=278650	$\hat{\beta}_0$=-21.1, $\hat{\beta}_1$=1.7, $\hat{\beta}_2$=-0.7, $\hat{\sigma}^2$=10670000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$ = 0.1, $\hat{\sigma}^2$=0.02
15	$\hat{\beta}_0$=1918.4, $\hat{\beta}_1$=1.1, $\hat{\beta}_2$=-0.4, $\hat{\sigma}^2$=4153300	$\hat{\beta}_0$=135.9, $\hat{\beta}_1$=1.0, $\hat{\beta}_2$=0.01, $\hat{\sigma}^2$=270130	$\hat{\beta}_0$=-4.4, $\hat{\beta}_1$=1.7, $\hat{\beta}_2$=-0.7, $\hat{\sigma}^2$=10881000	$\hat{\beta}_0$=0.1, $\hat{\beta}_1$=0.1, $\hat{\sigma}^2$=0.02

Table 8.5: Regression parameters at each stage in the truncated 1987 election.

Chapter 9

Discussion

9.1 Summary and Conclusions

Chapter 2 began with us familiarising ourselves with the process of the general election, including the night itself, over which constituencies declare results. We covered how the UK is broken down into constituencies and clarified the main parties in history as well as key events per recent election. This includes the evolution from a two-party contest (Conservative and Labour) into a three-party contest (Conservative, Labour and Liberal Democrats) in recent times. Fundamental here is forecasting, mainly of the final result, but also of the result per undeclared seat. We then focused on the sources of information used to gear this, namely opinion polls, exit polls, previous elections and declared seats on the night itself.

In Chapter 3 we presented an account of time series modelling in general, beginning with its numerous applications in real life. We explained how very generally one may decompose the time series into three parts: the trend, the seasonal component and the irregular part. The data may be stationary, and the autocorrelation function plays a vital role in assessing which lags most influence the current value. We then went on to look at non-stationarity of the mean and variance, the latter of which involves the modelling of volatility. It is of course important to choose the best model for the time series of interest and there are model criteria for doing this; we may then assess specific goodness of fit using such tools as F-tests, for instance. We looked at some elementary approaches for forecasting upcoming values given what is known. We then presented models which are strictly stationary, distinguishing clearly between three types of model: stationary, stochastic volatility and generalised autoregressive conditional heteroskedacity (GARCH) models. These each use different historical information in the modelling of probabilities of voting. Furthermore, we reviewed the mixture transition distribution for stationary models, which enable more than the previous lag to be considered in modelling. After looking at strict stationarity, we showed how to adapt these so as to model differing time intervals.

Chapter 4 was devoted to the existing forecasting methods used for elections. The simplest of these is the poll of polls (the average of poll results collected), which provides little scope for analysis. A more sophisticated approach, set in the area of time series, involved a multivariate local level model, which puts more emphasis on recent data. The cube rule was a very early attempt at predicting the final outcome, but became somewhat redundant in the UK with the concept of the third party in the contest. We also briefly reviewed some other approaches taken.

Chapter 5 then reviewed the method of maximisation in the presence of unobserved variables, known as the h-likelihood. Most of the work is due to the research of Lee and Nelder, who since the mid-1990s have developed the theory. A key property is that the estimates are asymptotically equivalent to the unavailable maximum likelihood estimates, the latter of which we would have sought in the absence of latent data.

Next, in Chapter 6, we moved onto the main focus of the thesis, which is defining our approach to modelling the election poll data. We have, in the simplest case, a beta-binomial model, in which the latent variables take beta (continuous) densities and subsequently the number of votes take binomial (discrete) densities, which we showed was an appealing yet straightforward arrangement. Using the results of Chapter 5, we specified a form for the likelihood and put a lot of detail into tailoring this theory for our time series scenario, including showing that the estimators are still asymptotically equivalent to maximum likelihood estimates. We reintroduced the ARCH, stochastic volatility and GARCH models from Chapter 3. Subsequently, we defined each specific model of interest. We showed how the stationary model provides parameters via the h-likelihood which are very similar and asymptotically equivalent to those provided by the popular EM algorithm, asymptotically in this case actually requiring that the sample size of a poll is equal to the number of polls taken before the election. We acknowledged that this is of course unrealistic in practice. The next section looked at simulations from each main model, given realistic parameter values. Importantly, the simulations suggested that the stationary models were perhaps unsuitable in context, due to the simulations fluctuating too much. By contrast the simulations of the stochastic volatility and GARCH models looked more realistic, although too similar to one another; this was due to the proportions of votes behaving too similarly to the probabilities to be able to distinguish clearly one from the other. This led to us dropping the latter type, since they are more cumbersome than the former, especially when thinking of computation and programming. Regarding the volatility models, the h-likelihood parameter estimates obtained from simulated data were close to the initial choices, to confirm our slight adaption to the methodology.

Chapter 7 was a chapter of illustration. First, we focused on fitting the stationary and stochastic volatility models to real data, with the inevitable goal of setting up a framework for forecasting on election night. We used opinion poll data from the 1983 and 1987 consecutive elections. H-likelihood parameters were obtained as well as approximations of their correlation matrices and confidence intervals. We applied the scaled deviance test, used for selecting the best model and in applying this found that the stochastic volatility models easily outperformed the stationary models. This was not surprising when we reconsidered the simulations shown earlier.

Chapter 8 focused on election night. We looked in detail at the British Broadcasting Corporation regression approach, which has seen some refinement based on lessons learned from previous elections. Next, we described in detail our method, detailing how our methodology appears at each stage of waiting for the seats to declare their results. Having set up the theory we initiated the approach using the 1987 data, although in a much simplified context. Our probabilities consider recent history, but also update exit poll models and consider regression-based forecasts of votes and fit in nicely with the regression approach adopted. Our method is sensible and convenient in that we obtain underlying probabilities of voting at each stage, similar to with the opinion polls, and these are likely to be of interest to psephologists and perhaps politicians.

Griliches (1974, [50]) states that some unobservable variables act as 'carriers of some of the content of our theories.' The focus on this thesis has been on using the probabilities of voting as instruments in the modelling of election data. These probabilities are determined by unique historical information, thus putting us in a time series framework. The history could be probabilities of voting (at previous stages), numbers of votes (at previous stages) or both. The observations and probabilities each have their own distributions. Clearly the probabilities are latent data; in their absence we would have used usual maximum likelihood estimation to get model parameters, but employed a newer approach to deal with such unobserved information. Much of the subsequent inference is similar to that, provided we had the typical maximum likelihood parameter estimates. The election data could be pre-polls or seats on election night, but we focused more so on the former yet outlined how to proceed with the latter, and also incorporating exit poll information. We conclude that this appears to be a sensible and sophisticated yet conceptually simple contribution to the modelling of election data, in which the probabilities evolve naturally alongside the actual voting.

Our modelling and overall approach improves on the current models for elections in various ways. The simplest statistic is the poll of polls. We provide scope for more analysis of our results, which are in comparison more than just a point estimate. The infamous cube law is limited to two parties whereas fight for UK Government is now at least a three-party contest; our model can cater for as many parties as is needed. The multivariate structural time series by Harvey and Shephard (1990, [55]) simply assumes the normality of votes whereas we give them a distribution (binomial/multinomial) which makes sense - voters choose one party out of a fixed set of parties. None of these methods discussed thus far provides probabilities of voting; in Brown, Firth and Payne (1999, [22]) the authors comment that such probabilities should have a prominent role in the broadcasted presentation of the forecast. Not only do probabilities help to convey the inevitable uncertainty in any forecast, but they would also make election night more interesting for the audience. The BBC method presented in Brown and Payne (1975, [23]) provides probabilities but assumes their normality. We have modelled them as beta/Dirichlet data which seems a natural choice. Note, however, that Lee and Nelder (1996, [69]) analysed various hierarchical generalised linear models (HGLMs) when fit to real datasets, focusing on conjugate HGLMs including the beta-binomial, Poisson-gamma and gamma-inverse gamma models. They commented that the distributions for the unobserved variables in conjugate HGLMs behave similarly to the normal distribution, and often tend to normality rapidly as the maximum hierarchical likelihood estimate values increase.

Available is a range of models within our general specification. We can consider different history; we saw with our data that the stochastic volatility models were best, which consider time differences between outcomes. However, to simplify matters or if we suspect stationarity then we may use our autoregressive conditional heteroskedacity models. If we suspect that the historical outcomes have different influence from the probabilities then we may employ our GARCH models; in our cases and with our particular specifications we found that this difference was not obvious. Furthermore, with all models we may assess historical information (lags) as far back as we wish. We should not necessarily reject a forecasting model altogether, on the basis of only one election trial.

9.2 Extensions

Below we list some possible avenues of future work to extend what we have done.

Multi-latent processes Pitt and Walker (2006, [96]) consider multi-latent processes. We said that having continuous probability density functions in between the election votes gives us flexibility in the modelling. We could introduce more flexibility if required using the theory in this paper, which would basically involve a new latent variable modelling the latent probabilities.

Additional weights In our GARCH model the shape parameters were already quite elaborate. However, we could have even more weights, u, letting them differ according to lags rather than assuming them to be fixed no matter what lag. Obviously though, the idea of parsimony will determine whether this is necessary when it comes to choosing the best model for the data of interest.

Use of prior information In the forecasting on election night we discussed including the exit poll information in a weight function with the regression modelling. However, in Brown, Firth and Payne (1999) prior information is introduced into the regression design matrix and so is done differently. We could employ this approach with a few minor adjustments. Consequently, this would slightly change the regression modelling which we presented in Chapters 4 and 8.

Next election It will be interesting to perform our analysis when the next general election takes place, which will occur no later than June 2010. We will also be able to examine precisely how influential our predictive probabilities are on election night, since we do not know the final outcome yet (before-the-fact model), as opposed to having worked with historical elections for which we already know the results (after-the-fact model). It will be exciting to do a complete analysis of the whole night. The first task would be to collate opinion poll and exit poll data.

Other polls/elections There is no reason why we cannot apply our overall approach to the modelling of *any* polls or elections conducted, under the necessary conditions. The competitors would still be modelled binomially/multinomially and the probability of voting still beta/Dirichlet.

Bayesian/nonparameteric approach It would be possible to switch to a Bayesian and/or nonparametric perspective, to compare with our classical and parametric approach. Mena and Walker (2005, [84]) cover the necessary theory to enable this with our stationary models.

Continuous time Our time series analysis has been in discrete time. A much more complicated approach would be to investigate whether it is possible to move into continuous time, whereby we have the *exact* times when the polls take place, and when the seats declare. A useful theoretical starting point would be the famous book 'The Theory of Stochastic Processes' by Cox and Miller (1977, [30]), which, for instance, has a chapter on Markox processes with discrete states in continuous time, a section on stationary processes in continuous time and a chapter on non-Markovian processes.

Appendix A

Programs

We made use of the useful mathematical software 'Maple' (Versions 10 and 11) to help solve some of the more difficult mathematics such as the integrals within using the EM algorithm. Our main software package, however, was 'MATLAB' (Version 7.2.0.232 (R2006a)), in which we wrote and ran all the programs. This includes the running of simulations and production of most plots. We also plotted using 'Minitab' (Version 15.1.0.0).

A.1 H-Likelihood Evaluation

Essentially, with each model the program had two chief components: maximisation of the latent variables followed by maximisation of the parameters via the Newton-Raphson method. These were combined into an iterative procedure in order to produce the desired (converged) result.

Below is some example code, both from one of our stationary models (see Figures A.1-A.5, ARCH(3,2)) and one of our stochastic volatility models (see Figures A.6-A.11, SV(3,2)). Note that generally there is a typical structure for evaluation of parameters in all models, although they differ in level of complexity. Note also that in the code text after the symbol '%' are brief explanations of the meaning at that point in the program; they do not interrupt the programs.

Methods

The following summarises the steps required to evaluate the maximum hierarchical likelihood estimate (MHLE) for the mixture transition distribution models, given the way in which we structured the likelihood. Recall that these are stationary models. A fuller explanation of the components was given in Chapter 6.

For $k = 1 : m$, $m =$ a large number of iterations:

1. Intuitively, choose $\theta^{(k)}$ and $\mathbf{Z}^{(k)}$
2. $\mathbf{W}^{(k)} | \mathbf{Z}^{(k)}$
3. $\mathbf{X}^{(k)} | \theta^{(k)}$, $\mathbf{Z}^{(k)}$ (i.e. k=1 is now complete)
4. $\theta^{(k+1)} | \mathbf{X}^{(k)}$, $\mathbf{Z}^{(k)}$
5. $\mathbf{Z}^{(k+1)} | \mathbf{X}^{(k)}, \mathbf{W}^{(k)}, \theta^{(k+1)}$
6. Loop from 2. to 5. until convergence of all

In the algorithm above:

- the main components of interest are θ (party strengths) and $\mathbf{w}$ (lag strengths); and
- $\mathbf{z}$ and $\mathbf{x}$ are latent but still of some interest ($\mathbf{z}$ identifies the dominating lag at each poll t and $\mathbf{x}$ states $Pr(\mathbf{Y} = \mathbf{y})$ at each poll).

The following summarises the steps required to evaluate the MHLE for the stochastic volatility models. A fuller explanation of the components was given in Chapter 6. Note that this is the more standard form of algorithm for finding the MHLEs of the h-likelihood for most of our models, and so the program for the generalised autoregressive conditional heterskedacity models (as well as the simplest stationary models) would be very similar.

For $k = 1 : m$, $m =$ a large number of iterations:

1. Intuitively, choose $\theta^{(k)}$
2. $\mathbf{X}^{(k)} | \theta^{(k)}$ (i.e. k=1 is now complete)
3. $\theta^{(k+1)} | \mathbf{X}^{(k)}$
4. Loop from 2. to 3. until convergence of all

In the algorithm above:

- the main components of interest are θ (party strengths and lag strengths); and
- $\mathbf{x}$ are latent but still of some interest ($\mathbf{x}$ states $Pr(\mathbf{Y} = \mathbf{y})$ at each poll).

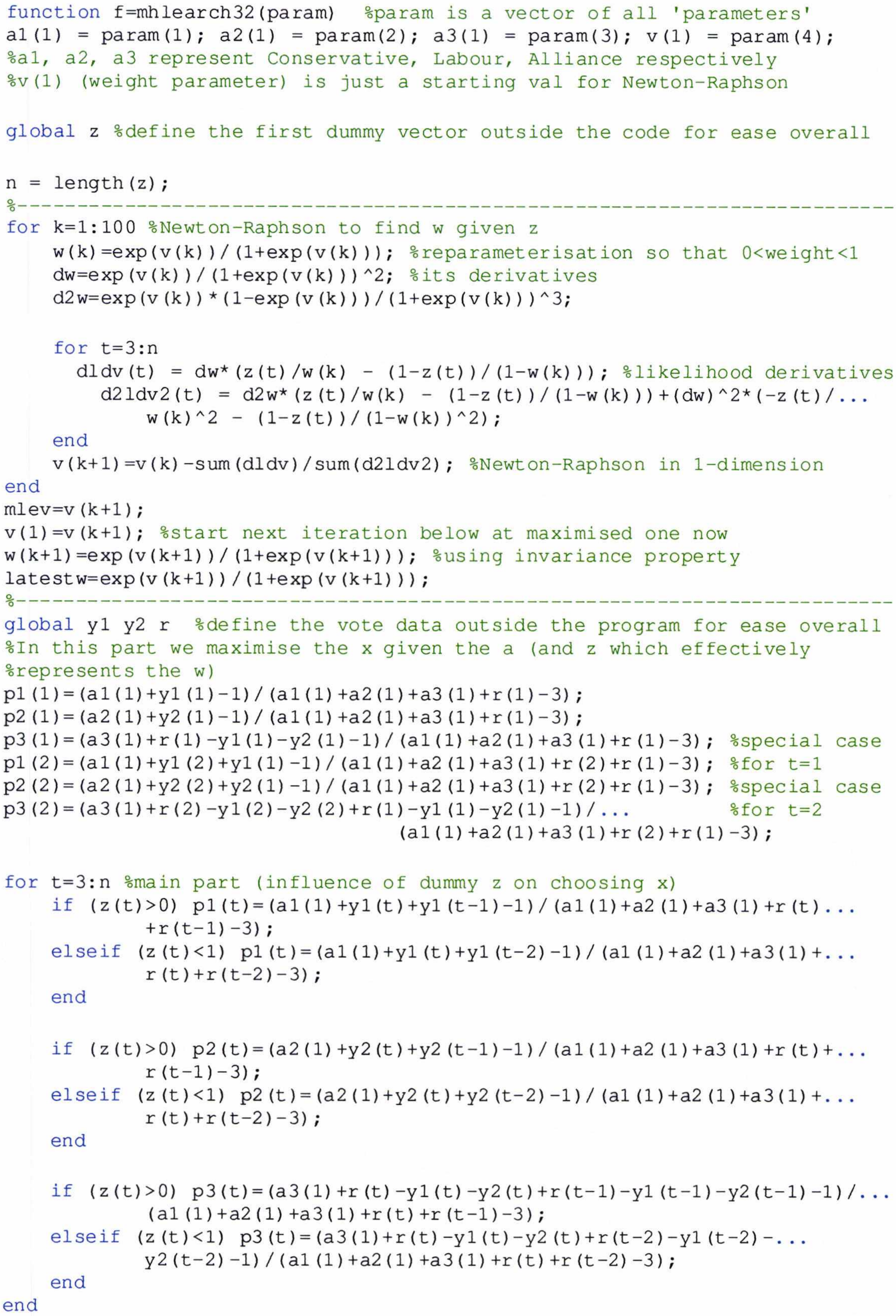
A.2 Simulations

In order to study model behaviour we ran several simulations, an example of which is shown in Figures A.12-A.13 (for SV(3,2)). Again, all models follow a similar procedure.

A.3 Forecasting

Following the poll parameter estimates, we have programs to forecast the seat distribution among the main three parties. Our programs give the continually-revised predictions, that is, imagining that we observed each declaration and then accordingly updated our forecast of the remaining undeclared seat outcomes. An example program is shown in Figures A.14-A.15.

Note that in all three cases above it is straightforward to adapt any program as and when necessary.



```
function f=mhlearch32(param)  %param is a vector of all 'parameters'
a1(1) = param(1); a2(1) = param(2); a3(1) = param(3); v(1) = param(4);
%a1, a2, a3 represent Conservative, Labour, Alliance respectively
%v(1) (weight parameter) is just a starting val for Newton-Raphson

global z %define the first dummy vector outside the code for ease overall

n = length(z);
%-----------------------------------------------------------------------
for k=1:100 %Newton-Raphson to find w given z
    w(k)=exp(v(k))/(1+exp(v(k))); %reparameterisation so that 0<weight<1
    dw=exp(v(k))/(1+exp(v(k)))^2; %its derivatives
    d2w=exp(v(k))*(1-exp(v(k)))/(1+exp(v(k)))^3;

    for t=3:n
      dldv(t) = dw*(z(t)/w(k) - (1-z(t))/(1-w(k))); %likelihood derivatives
        d2ldv2(t) = d2w*(z(t)/w(k) - (1-z(t))/(1-w(k)))+(dw)^2*(-z(t)/...
            w(k)^2 - (1-z(t))/(1-w(k))^2);
    end
    v(k+1)=v(k)-sum(dldv)/sum(d2ldv2); %Newton-Raphson in 1-dimension
end
mlev=v(k+1);
v(1)=v(k+1); %start next iteration below at maximised one now
w(k+1)=exp(v(k+1))/(1+exp(v(k+1))); %using invariance property
latestw=exp(v(k+1))/(1+exp(v(k+1)));
%-----------------------------------------------------------------------
global y1 y2 r  %define the vote data outside the program for ease overall
%In this part we maximise the x given the a (and z which effectively
%represents the w)
p1(1)=(a1(1)+y1(1)-1)/(a1(1)+a2(1)+a3(1)+r(1)-3);
p2(1)=(a2(1)+y2(1)-1)/(a1(1)+a2(1)+a3(1)+r(1)-3);
p3(1)=(a3(1)+r(1)-y1(1)-y2(1)-1)/(a1(1)+a2(1)+a3(1)+r(1)-3); %special case
p1(2)=(a1(1)+y1(2)+y1(1)-1)/(a1(1)+a2(1)+a3(1)+r(2)+r(1)-3); %for t=1
p2(2)=(a2(1)+y2(2)+y2(1)-1)/(a1(1)+a2(1)+a3(1)+r(2)+r(1)-3); %special case
p3(2)=(a3(1)+r(2)-y1(2)-y2(2)+r(1)-y1(1)-y2(1)-1)/...        %for t=2
                                  (a1(1)+a2(1)+a3(1)+r(2)+r(1)-3);

for t=3:n %main part (influence of dummy z on choosing x)
    if (z(t)>0) p1(t)=(a1(1)+y1(t)+y1(t-1)-1)/(a1(1)+a2(1)+a3(1)+r(t)...
            +r(t-1)-3);
    elseif (z(t)<1) p1(t)=(a1(1)+y1(t)+y1(t-2)-1)/(a1(1)+a2(1)+a3(1)+...
            r(t)+r(t-2)-3);
    end

    if (z(t)>0) p2(t)=(a2(1)+y2(t)+y2(t-1)-1)/(a1(1)+a2(1)+a3(1)+r(t)+...
            r(t-1)-3);
    elseif (z(t)<1) p2(t)=(a2(1)+y2(t)+y2(t-2)-1)/(a1(1)+a2(1)+a3(1)+...
            r(t)+r(t-2)-3);
    end

    if (z(t)>0) p3(t)=(a3(1)+r(t)-y1(t)-y2(t)+r(t-1)-y1(t-1)-y2(t-1)-1)/...
            (a1(1)+a2(1)+a3(1)+r(t)+r(t-1)-3);
    elseif (z(t)<1) p3(t)=(a3(1)+r(t)-y1(t)-y2(t)+r(t-2)-y1(t-2)-...
            y2(t-2)-1)/(a1(1)+a2(1)+a3(1)+r(t)+r(t-2)-3);
    end
end
```

Figure A.1: H-likelihood code for ARCH(3,2).

```
for t=1:n
x1(t)=p1(t); x2(t)=p2(t); x3(t)=p3(t);
end
maxx1=x1; maxx2=x2; maxx3=x3;   %so all variables for h=0 found now; onto...
                                %next, i.e. h=1
%---------------------------------------------------------------------------
for h=1:50 %this is the h-likelihood; first get a|x then x|a til ...
           %convergence of both

for j=2:50 %Newton-Raphson to find the a1, a2, a3 given current z and x

  for t=3:n %relevant components to the partial derivatives
    di0(t)=psi(0,a1(j-1)+a2(j-1)+a3(j-1)+r(t-1)); di1(t)=psi(0,a1(j-1)+...
        y1(t-1)); di2(t)=psi(0,a2(j-1)+y2(t-1));
    di3(t)=psi(0,a3(j-1)+r(t-1)-y1(t-1)-y2(t-1));
    tr0(t)=psi(1,a1(j-1)+a2(j-1)+a3(j-1)+r(t-1)); tr1(t)=psi(1,a1(j-1)+...
        y1(t-1)); tr2(t)=psi(1,a2(j-1)+y2(t-1));
    tr3(t)=psi(1,a3(j-1)+r(t-1)-y1(t-1)-y2(t-1));
    di00(t)=psi(0,a1(j-1)+a2(j-1)+a3(j-1)+r(t-2)); di11(t)=psi(0,...
        a1(j-1)+y1(t-2)); di22(t)=psi(0,a2(j-1)+y2(t-2));
    di33(t)=psi(0,a3(j-1)+r(t-2)-y1(t-2)-y2(t-2));
    tr00(t)=psi(1,a1(j-1)+a2(j-1)+a3(j-1)+r(t-2)); tr11(t)=psi(1,...
        a1(j-1)+y1(t-2)); tr22(t)=psi(1,a2(j-1)+y2(t-2));
    tr33(t)=psi(1,a3(j-1)+r(t-2)-y1(t-2)-y2(t-2));
    l1(t) =log(x1(t)); l2(t) =log(x2(t)); l3(t) =log(1-x1(t)-x2(t));
  end
    di0(2)=psi(0,a1(j-1)+a2(j-1)+a3(j-1)+r(1));
    di1(2)=psi(0,a1(j-1)+y1(1)); di2(2)=psi(0,a2(j-1)+y2(1)); di3(2)=...
        psi(0,a3(j-1)+r(1)-y1(1)-y2(1));
    tr0(2)=psi(1,a1(j-1)+a2(j-1)+a3(j-1)+r(1));
    tr1(2)=psi(1,a1(j-1)+y1(1)); tr2(2)=psi(1,a2(j-1)+y2(1)); tr3(2)=...
        psi(1,a3(j-1)+r(1)-y1(1)-y2(1));
    di0(1)=psi(0,a1(j-1)+a2(j-1)+a3(j-1));
    di1(1)=psi(0,a1(j-1)); di2(1)=psi(0,a2(j-1)); di3(1)=psi(0,a3(j-1));
    tr0(1)=psi(1,a1(j-1)+a2(j-1)+a3(j-1));
    tr1(1)=psi(1,a1(j-1)); tr2(1)=psi(1,a2(j-1)); tr3(1)=psi(1,a3(j-1));
    l1(2) =log(x1(2)); l2(2) =log(x2(2)); l3(2) =log(1-x1(2)-x2(2));
    l1(1) =log(x1(1)); l2(1) =log(x2(1)); l3(1) =log(1-x1(1)-x2(1));

    for t=3:n
   d2pda12(t) = z(t)*(tr0(t)-tr1(t)) + (1-z(t))*(tr00(t)-tr11(t));
   d2pda22(t) = z(t)*(tr0(t)-tr2(t)) + (1-z(t))*(tr00(t)-tr22(t));
   d2pda32(t) = z(t)*(tr0(t)-tr3(t)) + (1-z(t))*(tr00(t)-tr33(t));
   d2pda1da2(t) = z(t)*(tr0(t)) + (1-z(t))*(tr00(t));
  dpda1(t) = z(t)*(di0(t)-di1(t)+l1(t)) + (1-z(t))*(di00(t)-di11(t)+l1(t));
  dpda2(t) = z(t)*(di0(t)-di2(t)+l2(t)) + (1-z(t))*(di00(t)-di22(t)+l2(t));
  dpda3(t) = z(t)*(di0(t)-di3(t)+l3(t)) + (1-z(t))*(di00(t)-di33(t)+l3(t));
    end

  A = tr0(1)-tr1(1)+tr0(2)-tr1(2)+sum(d2pda12); %relevant partial
  B = tr0(1)-tr2(1)+tr0(2)-tr2(2)+sum(d2pda22); %derivatives
  C = tr0(1)-tr3(1)+tr0(2)-tr3(2)+sum(d2pda32);
  D = tr0(1)+tr0(2)+sum(d2pda1da2);
  E = di0(1)-di1(1)+di0(2)-di1(2)+l1(1)+l1(2)+sum(dpda1);
  F = di0(1)-di2(1)+di0(2)-di2(2)+l2(1)+l2(2)+sum(dpda2);
  G = di0(1)-di3(1)+di0(2)-di3(2)+l3(1)+l3(2)+sum(dpda3);
```

Figure A.2: H-likelihood code for ARCH(3,2) (*continued*).

```
%-------------------------------------------------------------------------
  M = [a1(j-1);a2(j-1);a3(j-1)]; %last parameter vector
  M2 = M - [(B*C-D*D)*E + (D*D-C*D)*F + (D*D-B*D)*G; ...
            (D*D-C*D)*E + (A*C-D*D)*F + (D*D-A*D)*G; ...
            (D*D-B*D)*E + (D*D-A*D)*F + (A*B-D*D)*G]/(A*(B*C-D*D)-D*...
            (C*D-D*D)+D*(D*D-B*D));
  a1(j)=M2(1); a2(j)=M2(2); a3(j)=M2(3); %latest parameter vector
end %get the converged a|x
%-------------------------------------------------------------------------
mlea1=a1(j); mlea2=a2(j); mlea3=a3(j);
a1(1)=a1(j); a2(1)=a2(j); a3(1)=a3(j); %starts new iteration at last mhles
latestmle = [a1(j); a2(j); a3(j)];
%-------------------------------------------------------------------------
for t=3:n      %first re-evaluate z now given new a1,a2,a3
f1(t)=log(w(k+1))+(y1(t)+a1(j)+y1(t-1)-1)*log(x1(t))+(y2(t)+a2(j)+...
    y2(t-1)-1)*log(x2(t))+(r(t)-y1(t)-y2(t)+a3(j)+r(t-1)-y1(t-1)-...
    y2(t-1)-1)*log(1-x1(t)-x2(t)) - ...
(gammaln(a1(j)+y1(t-1))+gammaln(a2(j)+y2(t-1))+gammaln(a3(j)+r(t-1)-...
y1(t-1)-y2(t-1))-...
gammaln(a1(j)+a2(j)+a3(j)+r(t-1)));
f2(t)=log(1-w(k+1))+(y1(t)+a1(j)+y1(t-2)-1)*log(x1(t))+(y2(t)+a2(j)+...
    y2(t-2)-1)*log(x2(t))+(r(t)-y1(t)-y2(t)+a3(j)+r(t-2)-y1(t-2)-...
    y2(t-2)-1)*log(1-x1(t)-x2(t)) - (gammaln(a1(j)+y1(t-2))+...
    gammaln(a2(j)+y2(t-2))+gammaln(a3(j)+r(t-2)-y1(t-2)-y2(t-2))-...
gammaln(a1(j)+a2(j)+a3(j)+r(t-2)));
end
for t=3:n %use this principle to choose whether 1st or 2nd lag dominates:
    if (f1(t) > f2(t)) z(t)=1;
    elseif (f1(t) < f2(t)) z(t)=0;
    end
end
mlez=z;
%-------------------------------------------------------------------------
for k=1:100 %need to re-evaluate the w given the new zs (same way as above)
w(k)=exp(v(k))/(1+exp(v(k)));
dw=exp(v(k))/(1+exp(v(k)))^2;
d2w=exp(v(k))*(1-exp(v(k)))/(1+exp(v(k)))^3;
for t=3:n
dldv(t) = dw*(z(t)/w(k) - (1-z(t))/(1-w(k)));
d2ldv2(t) = d2w*(z(t)/w(k) - (1-z(t))/(1-w(k))) + (dw)^2*(-z(t)/w(k)^2...
- (1-z(t))/(1-w(k))^2);
end
v(k+1)=v(k)-sum(dldv)/sum(d2ldv2);
end
mlev=v(k+1);
v(1)=v(k+1);
mlev2=v(k+1);
w(k+1)=exp(v(k+1))/(1+exp(v(k+1)));
mlew=exp(v(k+1))/(1+exp(v(k+1)));
%-------------------------------------------------------------------------
p1(1)=(a1(j)+y1(1)-1)/(a1(j)+a2(j)+a3(j)+r(1)-3); %need to re-evaluate
p2(1)=(a2(j)+y2(1)-1)/(a1(j)+a2(j)+a3(j)+r(1)-3); %the x|a (& z) now
p3(1)=(a3(j)+r(1)-y1(1)-y2(1)-1)/(a1(j)+a2(j)+a3(j)+r(1)-3);
p1(2)=(a1(j)+y1(2)+y1(1)-1)/(a1(j)+a2(j)+a3(j)+r(2)+r(1)-3);
p2(2)=(a2(j)+y2(2)+y2(1)-1)/(a1(j)+a2(j)+a3(j)+r(2)+r(1)-3);
p3(2)=(a3(j)+r(2)-y1(2)-y2(2)+r(1)-y1(1)-y2(1)-1)/(a1(j)+a2(j)+a3(j)+...
    r(2)+r(1)-3);
```

Figure A.3: H-likelihood code for ARCH(3,2) (*continued*).

```
for t=3:n
    if (z(t)>0) p1(t)=(a1(j)+y1(t)+y1(t-1)-1)/(a1(j)+a2(j)+a3(j)+...
            r(t)+r(t-1)-3);
    elseif (z(t)<1) p1(t)=(a1(j)+y1(t)+y1(t-2)-1)/(a1(j)+a2(j)+a3(j)+...
            r(t)+r(t-2)-3);
    end
    if (z(t)>0) p2(t)=(a2(j)+y2(t)+y2(t-1)-1)/(a1(j)+a2(j)+a3(j)+...
            r(t)+r(t-1)-3);
    elseif (z(t)<1) p2(t)=(a2(j)+y2(t)+y2(t-2)-1)/(a1(j)+a2(j)+a3(j)+...
            r(t)+r(t-2)-3);
    end
    if (z(t)>0) p3(t)=(a3(j)+r(t)-y1(t)-y2(t)+r(t-1)-y1(t-1)-y2(t-1)-1)/...
            (a1(j)+a2(j)+a3(j)+r(t)+r(t-1)-3);
    elseif (z(t)<1) p3(t)=(a3(j)+r(t)-y1(t)-y2(t)+r(t-2)-y1(t-2)-y2(t-2)...
            -1)/(a1(j)+a2(j)+a3(j)+r(t)+r(t-2)-3);
    end
end

for t=1:n
    x1(t)=p1(t); x2(t)=p2(t); x3(t)=p3(t);
end
maxx1=x1; maxx2=x2; maxx3=x3;
%-------------------------------------------------------------------------
end %completes h-likelihood
mhlea=[mlea1; mlea2; mlea3] %displays mhles of party parameters
mlew %displays mhle of weight parameter
mlez %displays optimal dummy indicator
maxx=[maxx1; maxx2; maxx3] %displays optimal latent probabilities

Q = [-A -D -D 0; -D -B -D 0; -D -D -C 0; 0 0 0 -sum(d2ldv2)];
VC = inv(Q); %approx var-cov matrix
SC = [sqrt(VC(1,1)) VC(1,2)/(sqrt(VC(1,1))*sqrt(VC(2,2))) VC(1,3)/...
    (sqrt(VC(1,1))*sqrt(VC(3,3))) ...
    VC(1,4)/(sqrt(VC(1,1))*sqrt(VC(4,4))); ...
      VC(2,1)/(sqrt(VC(1,1))*sqrt(VC(2,2))) sqrt(VC(2,2)) VC(2,3)/...
      (sqrt(VC(3,3))*sqrt(VC(2,2))) ...
      VC(2,4)/(sqrt(VC(2,2))*sqrt(VC(4,4))); ...
      VC(3,1)/(sqrt(VC(1,1))*sqrt(VC(3,3))) VC(3,2)/(sqrt(VC(3,3))*...
      sqrt(VC(2,2))) sqrt(VC(3,3)) ...
      VC(3,4)/(sqrt(VC(4,4))*sqrt(VC(3,3))); ...
      VC(4,1)/(sqrt(VC(1,1))*sqrt(VC(4,4))) VC(4,2)/(sqrt(VC(4,4))*...
      sqrt(VC(2,2))) ...
      VC(4,3)/(sqrt(VC(4,4))*sqrt(VC(3,3))) sqrt(VC(4,4))] %approx.
A1CI = [mlea1-2.5758*SC(1,1) ; mlea1+2.5758*SC(1,1)]' %correlation matrix
A2CI = [mlea2-2.5758*SC(2,2) ; mlea2+2.5758*SC(2,2)]'
A3CI = [mlea3-2.5758*SC(3,3) ; mlea3+2.5758*SC(3,3)]'%99% approx confidence
WCI = [mlew-2.5758*SC(4,4) ; mlew+2.5758*SC(4,4)]' %intervals per mhle

for t=1:length(y1) %Deviance for relative goodness of fit test
    l1(t) = gammaln(r(t)+1) - gammaln(y1(t)+1) - gammaln(y2(t)+1) - ...
        gammaln(r(t)-y1(t)-y2(t)+1) + (y1(t))*log(x1(t)) + (y2(t))*...
        log(x2(t)) + (r(t)-y1(t)-y2(t))*log(1-x1(t)-x2(t));
end
loglik1=sum(l1);
Deviance = 2*(length(y1)-loglik1)
```

Figure A.4: H-likelihood code for ARCH(3,2) (*continued*).

```
for t=3:length(y1) %Computes the maximum hierarchical likelihood
    lik(t) = z(t)*(log(mlew) + gammaln(r(t)+1) - gammaln(y1(t)+1) -...
        gammaln(y2(t)+1) - gammaln(r(t)-y1(t)-y2(t)+1) + ...
        (mlea1+y1(t-1)-1+y1(t))*log(x1(t)) + (mlea2+y2(t-1)-1+y2(t))*...
        log(x2(t)) + (mlea3+r(t-1)-y1(t-1)-y2(t-1)-1+r(t)-y1(t)-y2(t))*...
        log(1-x1(t)-x2(t)) + gammaln(mlea1+mlea2+mlea3+r(t-1)) - ...
        gammaln(mlea1+y1(t-1)) - gammaln(mlea2+y2(t-1)) - gammaln(mlea3+...
        r(t-1)-y1(t-1)-y2(t-1)))+(1-z(t))*(log(1-mlew) + gammaln(r(t)+1)...
        - gammaln(y1(t)+1) - gammaln(y2(t)+1) - gammaln(r(t)-y1(t)-...
        y2(t)+1) + (mlea1+y1(t-2)-1+y1(t))*log(x1(t)) + (mlea2+y2(t-2)...
        -1+y2(t))*log(x2(t)) + (mlea3+r(t-2)-y1(t-2)-y2(t-2)-1+r(t)-...
        y1(t)-y2(t))*log(1-x1(t)-x2(t)) + gammaln(mlea1+mlea2+mlea3+...
        r(t-2)) - gammaln(mlea1+y1(t-2)) - gammaln(mlea2+y2(t-2)) -...
        gammaln(mlea3+r(t-2)-y1(t-2)-y2(t-2)));
end
    lik(2) = gammaln(r(2)+1) - gammaln(y1(2)+1) - gammaln(y2(2)+1) -...
        gammaln(r(2)-y1(2)-y2(2)+1) + (mlea1+y1(2-1)-1+y1(2))*log(x1(2))...
        + (mlea2+y2(2-1)-1+y2(2))*log(x2(2)) + (mlea3+r(2-1)-y1(2-1)-...
        y2(2-1)-1+r(2)-y1(2)-y2(2))*log(1-x1(2)-x2(2)) + gammaln(mlea1+...
        mlea2+mlea3+r(2-1)) - gammaln(mlea1+y1(2-1)) - gammaln(mlea2+...
        y2(2-1)) - gammaln(mlea3+r(2-1)-y1(2-1)-y2(2-1));
    lik(1) = gammaln(r(1)+1) - gammaln(y1(1)+1) - gammaln(y2(1)+1) -...
        gammaln(r(1)-y1(1)-y2(1)+1) + (mlea1-1+y1(1))*log(x1(1)) + ...
        (mlea2-1+y2(1))*log(x2(1)) + (mlea3-1+r(1)-y1(1)-y2(1))*log(1-...
        x1(1)-x2(1)) + gammaln(mlea1+mlea2+mlea3) - gammaln(mlea1) - ...
        gammaln(mlea2) - gammaln(mlea3);
  lo=sum(lik)
```

Figure A.5: H-likelihood code for ARCH(3,2) (*continued*).

```
function f=mhlesv32(param)  %param is a vector of all 'parameters'
phi(1)=param(1); v(1)=param(2);
%v(1) (weight parameter) is just a starting val for Newton-Raphson

global y1 y2 r d1 %define known data outside the code for ease overall

n=length(y1);


for t=2:n
    d2(t)=d1(t)+d1(t-1); %determine the 2nd lag time interval
end


for t=1:n %subsequent c_1 and c_2 inside the Dirichlet shape parameters
    c1(t)=1/(exp(phi(1)*d1(t))-1);
end
for t=2:n
    c2(t)=1/(exp(phi(1)*d2(t))-1);
end


w(1)=exp(v(1))/(1+exp(v(1))); %reparameterise the weight function so
                              %that lies in [0,1]


for t=1:n %initial values for x (do not matter after h-likelihood starts)
    x1(t)=(y1(t)+5)/(r(t)+5); x2(t)=(y2(t)-5)/(r(t)-5);x3(t)=1-x1(t)-x2(t);
end
%-----------------------------------------------------------------------------
for h=1:50 %h-likelihood starts here


for j=2:50 %Newton-Raphson to find phi, w | x


 for t=2:n
  c1(t) = 1/(exp(phi(j-1)*d1(t))-1); %derivatives of c_1 and c_2
  d1c1(t) = -(d1(t)*exp(phi(j-1)*d1(t)))/(exp(phi(j-1)*d1(t))-1)^2;
  d2c1(t) = (d1(t)^2*exp(phi(j-1)*d1(t))*(1+exp(phi(j-1)*d1(t))))/...
       (exp(phi(j-1)*d1(t))-1)^3;
  c2(t) = 1/(exp(phi(j-1)*d2(t))-1);
  d1c2(t) = -(d2(t)*exp(phi(j-1)*d2(t)))/(exp(phi(j-1)*d2(t))-1)^2;
  d2c2(t) = (d2(t)^2*exp(phi(j-1)*d2(t))*(1+exp(phi(j-1)*d2(t))))/...
       (exp(phi(j-1)*d2(t))-1)^3;
 end


 w(j-1)=exp(v(j-1))/(1+exp(v(j-1))); %derivatives of w
 dw=exp(v(j-1))/(1+exp(v(j-1)))^2;
 d2w=exp(v(j-1))*(1-exp(v(j-1)))/(1+exp(v(j-1)))^3;


 for t=3:n %relevant components of following partial derivatives
  Di0(t) = psi(0, (w(j-1)*c1(t)+(1-w(j-1))*c2(t)));
  Di1(t) = psi(0, (w(j-1)*c1(t)*y1(t-1)/r(t-1)+(1-w(j-1))*c2(t)*y1(t-2)/...
      r(t-2)));
  Di2(t) = psi(0, (w(j-1)*c1(t)*y2(t-1)/r(t-1)+(1-w(j-1))*c2(t)*y2(t-2)/...
      r(t-2)));
  Di3(t) = psi(0, (w(j-1)*c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))+...
      (1-w(j-1))*c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))));
  Tr0(t) = psi(1, (w(j-1)*c1(t)+(1-w(j-1))*c2(t)));
  Tr1(t) = psi(1, (w(j-1)*c1(t)*y1(t-1)/r(t-1)+(1-w(j-1))*c2(t)*y1(t-2)/...
```

Figure A.6: H-likelihood code for SV(3,2).

```
     r(t-2)));
 Tr2(t) = psi(1, (w(j-1)*c1(t)*y2(t-1)/r(t-1)+(1-w(j-1))*c2(t)*y2(t-2)/...
     r(t-2)));
 Tr3(t) = psi(1, (w(j-1)*c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))+...
     (1-w(j-1))*c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))));
end
 Di0(2) = psi(0, c1(2)); Di1(2) = psi(0, c1(2)*y1(1)/r(1));
 Di2(2) = psi(0, c1(2)*y2(1)/r(1)); Di3(2) = psi(0, c1(2)*(1-y1(1)/r(1)...
     -y2(1)/r(1)));
 Tr0(2) = psi(1, c1(2)); Tr1(2) = psi(1, c1(2)*y1(1)/r(1));
 Tr2(2) = psi(1, c1(2)*y2(1)/r(1)); Tr3(2) = psi(1, c1(2)*(1-y1(1)/r(1)...
     -y2(1)/r(1)));

 for t=3:n %main part
     d2ldphi2(t) = w(j-1)*d2c1(t)*((y1(t-1)/r(t-1))*log(x1(t)) + ...
         (y2(t-1)/r(t-1))*log(x2(t)) + (1-y1(t-1)/r(t-1)-y2(t-1)/...
         r(t-1))*log(1-x1(t)-x2(t)) + Di0(t) - Di1(t)*y1(t-1)/r(t-1) - ...
         Di2(t)*y2(t-1)/r(t-1) - Di3(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/...
         r(t-1))) + w(j-1)*d1c1(t)*(Tr0(t)*(w(j-1)*d1c1(t)+(1-w(j-1))*...
         d1c2(t)) - Tr1(t)*(y1(t-1)/r(t-1))*(w(j-1)*d1c1(t)*y1(t-1)/...
         r(t-1)+(1-w(j-1))*d1c2(t)*y1(t-2)/r(t-2)) - Tr2(t)*(y2(t-1)/...
         r(t-1))*(w(j-1)*d1c1(t)*y2(t-1)/r(t-1)+(1-w(j-1))*d1c2(t)*...
         y2(t-2)/r(t-2)) - Tr3(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))*...
         (w(j-1)*d1c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))+(1-w(j-1))*...
         d1c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2)))) +(1-w(j-1))*...
         d2c2(t)*((y1(t-2)/r(t-2))*log(x1(t)) + (y2(t-2)/r(t-2))*...
         log(x2(t)) + (1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))*log(1-x1(t)-...
         x2(t)) + Di0(t) - Di1(t)*y1(t-2)/r(t-2) - Di2(t)*y2(t-2)/...
         r(t-2) - Di3(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))) +...
         (1-w(j-1))*d1c2(t)*(Tr0(t)*(w(j-1)*d1c1(t)+(1-w(j-1))*d1c2(t))...
         - Tr1(t)*(y1(t-2)/r(t-2))*(w(j-1)*d1c1(t)*y1(t-1)/r(t-1)+...
         (1-w(j-1))*d1c2(t)*y1(t-2)/r(t-2)) - Tr2(t)*(y2(t-2)/...
         r(t-2))*(w(j-1)*d1c1(t)*y2(t-1)/r(t-1)+(1-w(j-1))*d1c2(t)*...
         y2(t-2)/r(t-2)) - Tr3(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))*...
         (w(j-1)*d1c1(t)*(1-y1(t-1)/r(t-1)-y2(t-2)/r(t-2))+(1-w(j-1))*...
         d1c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))));
  d2ldv2(t) = d2w*c1(t)*((y1(t-1)/r(t-1))*log(x1(t)) + (y2(t-1)/...
      r(t-1))*log(x2(t)) + (1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))*...
      log(1-x1(t)-x2(t)) + Di0(t) - Di1(t)*y1(t-1)/r(t-1) - Di2(t)*...
      y2(t-1)/r(t-1) - Di3(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))) + ...
     (dw*c1(t))*(Tr0(t)*(dw*c1(t)-dw*c2(t)) - Tr1(t)*(y1(t-1)/r(t-1))*...
     (dw*c1(t)*y1(t-1)/r(t-1)-dw*c2(t)*y1(t-2)/r(t-2)) - Tr2(t)*...
     (y2(t-1)/r(t-1))*(dw*c1(t)*y2(t-1)/r(t-1)-dw*c2(t)*y2(t-2)/...
     r(t-2)) - Tr3(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))*(dw*c1(t)*(1-...
     y1(t-1)/r(t-1)-y2(t-1)/r(t-1))-dw*c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/...
     r(t-2)))) -(d2w*c2(t)*((y1(t-2)/r(t-2))*log(x1(t))+(y2(t-2)/...
     r(t-2))*log(x2(t))+((1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2)))*log(1-...
     x1(t)-x2(t)) + Di0(t) - Di1(t)*y1(t-2)/r(t-2) - Di2(t)*y2(t-2)/...
     r(t-2) - Di3(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))) + (dw*c2(t))*...
     (Tr0(t)*(dw*c1(t)-dw*c2(t)) - Tr1(t)*(y1(t-2)/r(t-2))*(dw*c1(t)*...
     y1(t-1)/r(t-1)-dw*c2(t)*y1(t-2)/r(t-2)) - Tr2(t)*(y2(t-2)/r(t-2))*...
     (dw*c1(t)*y2(t-1)/r(t-1)-dw*c2(t)*y2(t-2)/r(t-2)) - Tr3(t)*(1-...
     y1(t-2)/r(t-2)-y2(t-2)/r(t-2))*(dw*c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)...
     /r(t-1))-dw*c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2)))));
   d2ldphidv(t) = d1c1(t)*dw*((y1(t-1)/r(t-1))*log(x1(t))+(y2(t-1)/...
    r(t-1))*log(x2(t))+(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))*log(1-x1(t)...
    -x2(t)) + Di0(t)-Di1(t)*y1(t-1)/r(t-1)-Di2(t)*y2(t-1)/r(t-1)-...
```

Figure A.7: H-likelihood code for SV(3,2) (*continued*).

```
     Di3(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))) + c1(t)*dw*(Tr0(t)*...
     (w(j-1)*d1c1(t)+(1-w(j-1))*d1c2(t))-Tr1(t)*(y1(t-1)/r(t-1))*...
     (w(j-1)*d1c1(t)*y1(t-1)/r(t-1)+(1-w(j-1))*d1c2(t)*y1(t-2)/r(t-2)) -...
     Tr2(t)*(y2(t-1)/r(t-1))*(w(j-1)*d1c1(t)*y2(t-1)/r(t-1)+(1-w(j-1))*...
     d1c2(t)*y2(t-2)/r(t-2))-Tr3(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))*...
     (w(j-1)*d1c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))+(1-w(j-1))*...
     d1c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2)))) -(d1c2(t)*dw*((y1(t-2)/...
     r(t-2))*log(x1(t))+(y2(t-2)/r(t-2))*log(x2(t))+(1-y1(t-2)/r(t-2)-...
     y2(t-2)/r(t-2))*log(1-x1(t)-x2(t)) + Di0(t)-Di1(t)*y1(t-2)/r(t-2)-...
     Di2(t)*y2(t-2)/r(t-2)-Di3(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))) + ...
     c2(t)*dw*(Tr0(t)*(w(j-1)*d1c1(t)+(1-w(j-1))*d1c2(t))-Tr1(t)*...
     (y1(t-2)/r(t-2))*(w(j-1)*d1c1(t)*y1(t-1)/r(t-1)+(1-w(j-1))*...
     d1c2(t)*y1(t-2)/r(t-2)) - Tr2(t)*(y2(t-2)/r(t-2))*(w(j-1)*d1c1(t)*...
     y2(t-1)/r(t-1)+(1-w(j-1))*d1c2(t)*y2(t-2)/r(t-2))-Tr3(t)*(1-y1(t-2)...
     /r(t-2)-y2(t-2)/r(t-2))*(w(j-1)*d1c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/...
     r(t-1))+(1-w(j-1))*d1c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2)))));
     dldphi(t) = w(j-1)*d1c1(t)*((y1(t-1)/r(t-1))*log(x1(t)) + (y2(t-1)/...
     r(t-1))*log(x2(t)) + (1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))*log(1-...
     x1(t)-x2(t)) + Di0(t) - Di1(t)*y1(t-1)/r(t-1) - Di2(t)*y2(t-1)/...
     r(t-1) - Di3(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))) + ...
     (1-w(j-1))*d1c2(t)*((y1(t-2)/r(t-2))*log(x1(t)) + (y2(t-2)/r(t-2))*...
     log(x2(t)) +(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))*log(1-x1(t)-x2(t)) +...
     Di0(t) - Di1(t)*y1(t-2)/r(t-2) - Di2(t)*y2(t-2)/r(t-2) - Di3(t)*...
     (1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2)));
     dldv(t) = c1(t)*dw*((y1(t-1)/r(t-1))*log(x1(t))+(y2(t-1)/r(t-1))*...
     log(x2(t))+(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))*log(1-x1(t)-x2(t))...
     + Di0(t)-Di1(t)*y1(t-1)/r(t-1)-Di2(t)*y2(t-1)/r(t-1)-Di3(t)*(1-...
     y1(t-1)/r(t-1)-y2(t-1)/r(t-1))) -c2(t)*dw*((y1(t-2)/r(t-2))*...
     log(x1(t))+(y2(t-2)/r(t-2))*log(x2(t))+(1-y1(t-2)/r(t-2)-y2(t-2)/...
     r(t-2))*log(1-x1(t)-x2(t)) + Di0(t)-Di1(t)*y1(t-2)/r(t-2)-Di2(t)*...
     y2(t-2)/r(t-2)-Di3(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2)));
  end
      d2ldphi2(2) = d2c1(2)*((y1(1)/r(1))*log(x1(2)) + (y2(1)/r(1))*...
          log(x2(2)) + (1-y1(1)/r(1)-y2(1)/r(1))*log(1-x1(2)-x2(2)) + ...
        Di0(2) - Di1(2)*(y1(1)/r(1)) - Di2(2)*(y2(1)/r(1)) - Di3(2)*(1-...
        y1(1)/r(1)-y2(1)/r(1))) + d1c1(2)^2*(Tr0(2)-Tr1(2)*(y1(1)/...
        r(1))^2-Tr2(2)*(y2(1)/r(1))^2-Tr3(2)*(1-y1(1)/r(1)-y2(1)/r(1))^2);
      dldphi(2) = d1c1(2)*((y1(1)/r(1))*log(x1(2)) + (y2(1)/r(1))*...
          log(x2(2)) + (1-y1(1)/r(1)-y2(1)/r(1))*log(1-x1(2)-x2(2)) + ...
     Di0(2) - Di1(2)*(y1(1)/r(1)) - Di2(2)*(y2(1)/r(1)) - Di3(2)*(1-...
     y1(1)/r(1)-y2(1)/r(1))); %the partial derivatives
A=sum(d2ldphi2); B=sum(d2ldv2); C=sum(d2ldphidv);D=sum(dldphi);E=sum(dldv);

  M = [phi(j-1) ; v(j-1)]; %last values of mhles
  M2 = M - [(B*D-C*E)/(A*B-C*C) ; (A*E-C*D)/(A*B-C*C)];
  phi(j)=M2(1); v(j)=M2(2); %latest values of mhles
end
mlephi=phi(j); %mhle of phi
phi(1)=phi(j); %start next iteration at latest mhle for phi
mlev=v(j);
v(1)=v(j); %start next iteration at latest mhle for v
w(j)=exp(v(j))/(1+exp(v(j))); %using invariance property
mlew=w(j) %mhle of w
%-----------------------------------------------------------------------
for t=1:n %now need to get x|phi,w
    c1(t)=1/(exp(phi(j)*d1(t))-1);
end
```

Figure A.8: H-likelihood code for SV(3,2) (*continued*).

```
for t=2:n
    c2(t)=1/(exp(phi(j)*d2(t))-1);
end
%sequentially evaluate each x
q1n=(y1(n)+c1(n)*w(j)*y1(n-1)/r(n-1)+c2(n)*(1-w(j))*y1(n-2)/r(n-2)-1)/...
    (r(n)+w(j)*c1(n)+(1-w(j))*c2(n)-3);
q2n=(y2(n)+c1(n)*w(j)*y2(n-1)/r(n-1)+c2(n)*(1-w(j))*y2(n-2)/r(n-2)-1)/...
    (r(n)+w(j)*c1(n)+(1-w(j))*c2(n)-3);


for m=1:n-3 %Newton-Raphson to find x
 q1(1)=x1(n-m);
 q2(1)=x2(n-m);
 for k=1:50
  A2 = -1*(y1(n-m)+c1(n-m)*w(j)*y1(n-m-1)/r(n-m-1)+c2(n-m)*(1-w(j))*...
      y1(n-m-2)/r(n-m-2)-1)/q1(k)^2 - (r(n-m)-y1(n-m)-y2(n-m)+c1(n-m)*...
      w(j)*(1-y1(n-m-1)/r(n-m-1)-y2(n-m-1)/r(n-m-1))+c2(n-m)*(1-w(j))*...
      (1-y1(n-m-2)/r(n-m-2)-y2(n-m-2)/r(n-m-2))-1)/(1-q1(k)-q2(k))^2;
  B2 = -1*(y2(n-m)+c1(n-m)*w(j)*y2(n-m-1)/r(n-m-1)+c2(n-m)*(1-w(j))*...
      y2(n-m-2)/r(n-m-2)-1)/q2(k)^2 - (r(n-m)-y1(n-m)-y2(n-m)+c1(n-m)*...
      w(j)*(1-y1(n-m-1)/r(n-m-1)-y2(n-m-1)/r(n-m-1))+c2(n-m)*(1-w(j))*...
      (1-y1(n-m-2)/r(n-m-2)-y2(n-m-2)/r(n-m-2))-1)/(1-q1(k)-q2(k))^2;
  C2 = -1*(r(n-m)-y1(n-m)-y2(n-m)+c1(n-m)*w(j)*(1-y1(n-m-1)/r(n-m-1)-...
      y2(n-m-1)/r(n-m-1))+c2(n-m)*(1-w(j))*(1-y1(n-m-2)/r(n-m-2)-...
      y2(n-m-2)/r(n-m-2))-1)/(1-q1(k)-q2(k))^2 ;
  D2 = (y1(n-m)+c1(n-m)*w(j)*y1(n-m-1)/r(n-m-1)+c2(n-m)*(1-w(j))*...
      y1(n-m-2)/r(n-m-2)-1)/q1(k) -(r(n-m)-y1(n-m)-y2(n-m)+c1(n-m)*w(j)*...
      (1-y1(n-m-1)/r(n-m-1)-y2(n-m-1)/r(n-m-1))+c2(n-m)*(1-w(j))*(1-...
      y1(n-m-2)/r(n-m-2)-y2(n-m-2)/r(n-m-2))-1)/(1-q1(k)-q2(k));
  E2 = (y2(n-m)+c1(n-m)*w(j)*y2(n-m-1)/r(n-m-1)+c2(n-m)*(1-w(j))*...
      y2(n-m-2)/r(n-m-2)-1)/q2(k) - (r(n-m)-y1(n-m)-y2(n-m)+c1(n-m)*...
      w(j)*(1-y1(n-m-1)/r(n-m-1)-y2(n-m-1)/r(n-m-1))+c2(n-m)*(1-w(j))*...
      (1-y1(n-m-2)/r(n-m-2)-y2(n-m-2)/r(n-m-2))-1)/(1-q1(k)-q2(k));
Ma = [q1(k) ; q2(k)];
Ma2 = Ma - [((B2*D2-C2*E2)/(A2*B2-C2*C2)) ; ((A2*E2-C2*D2)/(A2*B2-C2*C2))];
   q1(k+1)=Ma2(1);
   q2(k+1)=Ma2(2);
 end
 p1(n-m)=q1(k+1);
 p2(n-m)=q2(k+1);
end

 q1(1)=x1(2); %special case for t=2; similar method to above
 q2(1)=x2(2);
 for k=1:50
  A3 = -1*(y1(2)+c1(2)*y1(1)/r(1)-1)/q1(k)^2 - (r(2)-y1(2)-y2(2)+c1(2)*...
      (1-y1(1)/r(1)-y2(1)/r(1))-1)/(1-q1(k)-q2(k))^2;
  B3 = -1*(y2(2)+c1(2)*y2(1)/r(1)-1)/q2(k)^2 - (r(2)-y1(2)-y2(2)+c1(2)*...
      (1-y1(1)/r(1)-y2(1)/r(1))-1)/(1-q1(k)-q2(k))^2;
  C3 = -1*(r(2)-y1(2)-y2(2)+c1(2)*(1-y1(1)/r(1)-y2(1)/r(1))-1)/(1-q1(k)...
      -q2(k))^2;
  D3 = (y1(2)+c1(2)*y1(1)/r(1)-1)/q1(k) - (r(2)-y1(2)-y2(2)+c1(2)*(1-...
      y1(1)/r(1)-y2(1)/r(1))-1)/(1-q1(k)-q2(k));
  E3 = (y2(2)+c1(2)*y2(1)/r(1)-1)/q2(k) - (r(2)-y1(2)-y2(2)+c1(2)*...
      (1-y1(1)/r(1)-y2(1)/r(1))-1)/(1-q1(k)-q2(k));
Ma = [q1(k) ; q2(k)];
Ma2 = Ma - [((B3*D3-C3*E3)/(A3*B3-C3*C3)) ; ((A3*E3-C3*D3)/(A3*B3-C3*C3))];
```

Figure A.9: H-likelihood code for SV(3,2) (*continued*).

```
  q1(k+1)=Ma2(1);
  q2(k+1)=Ma2(2);
 end
 p1(2)=q1(k+1);
 p2(2)=q2(k+1);

p1(n)=q1n;p2(n)=q2n;p1(1)=x1(1);p2(1)=x2(1);


for t=1:n
 x1(t)=p1(t); x2(t)=p2(t); x3(t)=1-x1(t)-x2(t);
end
maxx1=x1; maxx2=x2; maxx3=x3;

end %h-likelihood complete


mlephi %displays optimal phi value
mlew %displays optimal weight
maxx=[maxx1;maxx2;maxx3] %displays maximised latent probabilities

VC = [-B/(A*B-C*C) C/(A*B-C*C) ; C/(A*B-C*C) -A/(A*B-C*C)];
%approx var-cov matrix
SC = [sqrt(-B/(A*B-C*C)) (C/(A*B-C*C))/(sqrt(-B/(A*B-C*C))*...
    sqrt(-A/(A*B-C*C))) ;(C/(A*B-C*C))/(sqrt(-B/(A*B-C*C))*...
    sqrt(-A/(A*B-C*C))) sqrt(-A/(A*B-C*C))] %approx correlation matrix
PCI = [mlephi-2.5758*SC(1,1) ; mlephi+2.5758*SC(1,1)]'
%approx 99% confidence interval
VCI = [mlev-2.5758*SC(2,2) ; mlev+2.5758*SC(2,2)]'

wse=sqrt(VC(2,2)*(exp(mlev)/(1+exp(mlev))^2)^2) %delta method to update
WCI=[mlew-2.5758*wse ; mlew+2.5758*wse]          %from v to w

for t=1:length(y1) %Deviance for goodness of fit test
    l1(t) = gammaln(r(t)+1) - gammaln(y1(t)+1) - gammaln(y2(t)+1) - ...
        gammaln(r(t)-y1(t)-y2(t)+1) +(y1(t))*log(x1(t)) + (y2(t))*...
        log(x2(t)) + (r(t)-y1(t)-y2(t))*log(1-x1(t)-x2(t));
end
loglik1=sum(l1);
Deviance = 2*(length(y1)-loglik1)

for t=3:length(y1) %Computes the maximum h-likelihood
    lik(t) = gammaln(r(t)+1) - gammaln(y1(t)+1)- gammaln(y2(t)+1) -...
        gammaln(r(t)-y1(t)-y2(t)+1) + (mlew*c1(t)*y1(t-1)/r(t-1)+(1-...
        mlew)*c2(t)*y1(t-2)/r(t-2)-1+y1(t))*log(x1(t)) + (mlew*c1(t)*...
        y2(t-1)/r(t-1)+(1-mlew)*c2(t)*y2(t-2)/r(t-2)-1+y2(t))*...
        log(x2(t))+ (mlew*c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))+(1-...
        mlew)*c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))-1+r(t)-y1(t)-...
        y2(t))*log(1-x1(t)-x2(t))+gammaln(mlew*c1(t)+(1-mlew)*c2(t)) -...
        gammaln(mlew*c1(t)*y1(t-1)/r(t-1)+(1-mlew)*c2(t)*y1(t-2)/...
        r(t-2)) -gammaln(mlew*c1(t)*y2(t-1)/r(t-1)+(1-mlew)*c2(t)*...
        y2(t-2)/r(t-2)) - gammaln(mlew*c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/...
        r(t-1))+(1-mlew)*c2(t)*(1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2)));
end
    lik(2) = gammaln(r(2)+1) - gammaln(y1(2)+1)- gammaln(y2(2)+1) -...
        gammaln(r(2)-y1(2)-y2(2)+1) +(c1(2)*y1(2-1)/r(2-1)-1+y1(2))*...
        log(x1(2)) + (c1(2)*y2(2-1)/r(2-1)-1+y2(2))*log(x2(2)) + (c1(2)*...
```

Figure A.10: H-likelihood code for SV(3,2) (*continued*).

```
(1-y1(2-1)/r(2-1)-y2(2-1)/r(2-1))-1+r(2)-y1(2)-y2(2))*log(1-...
x1(2)-x2(2)) +gammaln(c1(2)) - gammaln(c1(2)*y1(2-1)/r(2-1)) ...
- gammaln(c1(2)*y2(2-1)/r(2-1)) - gammaln(c1(2)*(1-y1(2-1)/ ...
r(2-1)-y2(2-1)/r(2-1)));
lo=sum(lik)
```

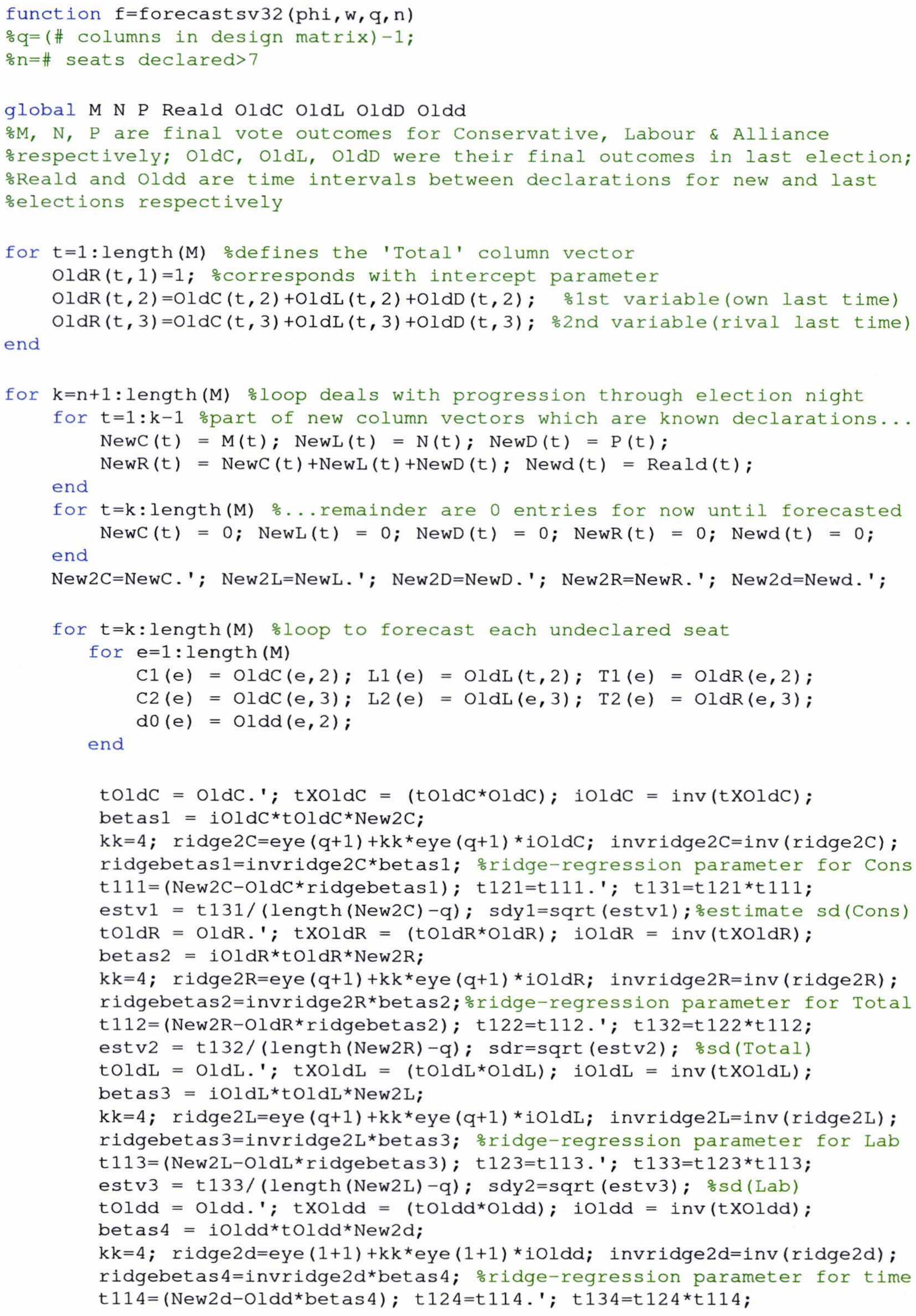
Figure A.11: H-likelihood code for SV(3,2) (*continued*).

```
function f=simsv32(param) %param is a vector of all 'parameters'
phi=param(1); x1(1)=param(2); x2(1)=param(3); w=param(4);
%each vector entry of parameter starting values

global d1 r
%enables us to define time interval and poll sample size outside the code

for t=2:length(r)
    d2(t)=d1(t)+d1(t-1); %time interval back to the 2nd lag
end

%Special case 1: t=1 (so no weight w appears here - 0 lags possible)
x3(1)=1-x1(1)-x2(1);
X = [x1(1) x2(1)];
if
    sum(X)<1 X=[X,1-sum(X)];       %simulates the x from Dirichlet
end
cump=cumsum(X);m=rand(r(1),1);cumy=zeros(1,3);
for j=1:3
    cumy(j) = sum(m<cump(j));
end
y=[cumy(1),cumy(2)-cumy(1),diff(cumy)];
y1(1)=cumy(1);
y2(1)=cumy(2)-cumy(1);      %simulates the y (multinomial) given the x

%Special case 2: t=2 (again no weight appears here - only 1 lag possible)
    c1(2)=1/(exp(phi*d1(2))-1);
    p1=gammasim(c1(2)*y1(1)/r(1),1);
    p2=gammasim(c1(2)*y2(1)/r(1),1);
    p3=gammasim(c1(2)*(1-y1(1)/r(1)-y2(1)/r(1)),1);
    x1(2)=p1 ./ (p1+p2+p3);
    x2(2)=p2 ./ (p1+p2+p3);
    x3(2)=1-x1(2)-x2(2);
    X = [x1(2) x2(2)];     %simulates the x from Dirichlet
    if
        sum(X)<1 X=[X,1-sum(X)];
    end
    cump=cumsum(X);m=rand(r(2),1);cumy=zeros(1,3);
    for j=1:3
        cumy(j) = sum(m<cump(j));
    end
    y=[cumy(1),cumy(2)-cumy(1),diff(cumy)];
    y1(2)=cumy(1);
    y2(2)=cumy(2)-cumy(1);   %simulates the y (multinomial) given the x

%Main part
for t=3:length(r)
    c1(t)=1/(exp(phi*d1(t))-1);
    c2(t)=1/(exp(phi*d1(t))-1);
    p1=gammasim((w*c1(t)*y1(t-1)/r(t-1)+(1-w)*c2(t)*y1(t-2)/r(t-1)),1);
    p2=gammasim((w*c1(t)*y2(t-1)/r(t-1)+(1-w)*c2(t)*y2(t-2)/r(t-1)),1);
    p3=gammasim((w*c1(t)*(1-y1(t-1)/r(t-1)-y2(t-1)/r(t-1))+(1-w)*c2(t)*...
                                   (1-y1(t-2)/r(t-2)-y2(t-2)/r(t-2))),1);
    x1(t)=p1/(p1+p2+p3);
    x2(t)=p2/(p1+p2+p3);
    x3(t)=1-x1(t)-x2(t);  %simulates the x from Dirichlet
    X = [x1(t) x2(t)];
    if
        sum(X)<1 X=[X,1-sum(X)];
    end
```

Figure A.12: Simulation code for SV(3,2).

```
    cump=cumsum(X);m=rand(r(t),1);cumy=zeros(1,3);
    for j=1:3
    cumy(j) = sum(m<cump(j));
    end
y=[cumy(1),cumy(2)-cumy(1),diff(cumy)];
y1(t)=cumy(1);
y2(t)=cumy(2)-cumy(1);    %simulates the y (multinomial) given the x
end

subplot(2,2,1)  %enables a series of small plots
m=1:1:length(r);
plot(m,x1,m,x2,':+',m,x3,'ok') %plots all party probabilities on one graph
ylim([0 1])   %defines limits of y axis to cover probability range [0,1]
subplot(2,2,2)
plot(m,x1)     %plots Conservative
ylim([0 1])
subplot(2,2,3)
plot(m,x2,':+')
ylim([0 1])   %plots Labour
subplot(2,2,4)
plot(x3,':ok')
ylim([0 1])   %plots Alliance
```

Figure A.13: Simulation code for SV(3,2) (*continued*).



```
function f=forecastsv32(phi,w,q,n)
%q=(# columns in design matrix)-1;
%n=# seats declared>7

global M N P Reald OldC OldL OldD Oldd
%M, N, P are final vote outcomes for Conservative, Labour & Alliance
%respectively; OldC, OldL, OldD were their final outcomes in last election;
%Reald and Oldd are time intervals between declarations for new and last
%elections respectively

for t=1:length(M) %defines the 'Total' column vector
    OldR(t,1)=1; %corresponds with intercept parameter
    OldR(t,2)=OldC(t,2)+OldL(t,2)+OldD(t,2);  %1st variable(own last time)
    OldR(t,3)=OldC(t,3)+OldL(t,3)+OldD(t,3); %2nd variable(rival last time)
end

for k=n+1:length(M) %loop deals with progression through election night
    for t=1:k-1 %part of new column vectors which are known declarations...
        NewC(t) = M(t); NewL(t) = N(t); NewD(t) = P(t);
        NewR(t) = NewC(t)+NewL(t)+NewD(t); Newd(t) = Reald(t);
    end
    for t=k:length(M) %...remainder are 0 entries for now until forecasted
        NewC(t) = 0; NewL(t) = 0; NewD(t) = 0; NewR(t) = 0; Newd(t) = 0;
    end
    New2C=NewC.'; New2L=NewL.'; New2D=NewD.'; New2R=NewR.'; New2d=Newd.';

    for t=k:length(M) %loop to forecast each undeclared seat
       for e=1:length(M)
           C1(e) = OldC(e,2); L1(e) = OldL(t,2); T1(e) = OldR(e,2);
           C2(e) = OldC(e,3); L2(e) = OldL(e,3); T2(e) = OldR(e,3);
           d0(e) = Oldd(e,2);
       end

        tOldC = OldC.'; tXOldC = (tOldC*OldC); iOldC = inv(tXOldC);
        betas1 = iOldC*tOldC*New2C;
        kk=4; ridge2C=eye(q+1)+kk*eye(q+1)*iOldC; invridge2C=inv(ridge2C);
        ridgebetas1=invridge2C*betas1; %ridge-regression parameter for Cons
        t111=(New2C-OldC*ridgebetas1); t121=t111.'; t131=t121*t111;
        estv1 = t131/(length(New2C)-q); sdy1=sqrt(estv1);%estimate sd(Cons)
        tOldR = OldR.'; tXOldR = (tOldR*OldR); iOldR = inv(tXOldR);
        betas2 = iOldR*tOldR*New2R;
        kk=4; ridge2R=eye(q+1)+kk*eye(q+1)*iOldR; invridge2R=inv(ridge2R);
        ridgebetas2=invridge2R*betas2;%ridge-regression parameter for Total
        t112=(New2R-OldR*ridgebetas2); t122=t112.'; t132=t122*t112;
        estv2 = t132/(length(New2R)-q); sdr=sqrt(estv2); %sd(Total)
        tOldL = OldL.'; tXOldL = (tOldL*OldL); iOldL = inv(tXOldL);
        betas3 = iOldL*tOldL*New2L;
        kk=4; ridge2L=eye(q+1)+kk*eye(q+1)*iOldL; invridge2L=inv(ridge2L);
        ridgebetas3=invridge2L*betas3; %ridge-regression parameter for Lab
        t113=(New2L-OldL*ridgebetas3); t123=t113.'; t133=t123*t113;
        estv3 = t133/(length(New2L)-q); sdy2=sqrt(estv3); %sd(Lab)
        tOldd = Oldd.'; tXOldd = (tOldd*Oldd); iOldd = inv(tXOldd);
        betas4 = iOldd*tOldd*New2d;
        kk=4; ridge2d=eye(1+1)+kk*eye(1+1)*iOldd; invridge2d=inv(ridge2d);
        ridgebetas4=invridge2d*betas4; %ridge-regression parameter for time
        t114=(New2d-Oldd*betas4); t124=t114.'; t134=t124*t114;
```

Figure A.14: Forecasting code.

```
        estv4 = t134/(length(New2d)-1); sdd=sqrt(estv4); %sd(time)

      y1(t) = ridgebetas1(1)+C1(t)*ridgebetas1(2)+C2(t)*ridgebetas1(3);
      y2(t) = ridgebetas3(1)+L1(t)*ridgebetas3(2)+L2(t)*ridgebetas3(3);
      r(t) = ridgebetas2(1)+T1(t)*ridgebetas2(2)+T2(t)*ridgebetas2(3);
      d1(t) = betas4(1)+d0(t)*betas4(2); %means of regression per case

      for j=1:10000          %how we estimate the predictive probabilities
           b1(j)=randn*sdy1+y1(t);
           b2(j)=randn*sdy2+y2(t);
           b3(j)=randn*sdr+r(t);  %1) sample from each distribution for
           b5(j)=randn*sdd+d1(t); %y1, y2, r and d; and ...
           b6(j)=b5(j)+d1(t-1);
           p1=gammasim(b1(j),1);
           p2=gammasim(b2(j),1);
           p3=gammasim(b3(j)-b1(j)-b2(j),1);
           b41(j)=p1 ./ (p1+p2+p3); b42(j)=p2 ./ (p1+p2+p3);
                 %2)...subsequently sample from the Dirichlet distribution
      end %repeat to get 10000 realisations of b41
       subplot(3,1,1) %enables multiple small plots rather than just one
       hist(b41)  %histogram of all Conservative sample probabilities
       xlim([0 1]) %plots whole of x range [0, 1]
       subplot(3,1,2)
       hist(b42) %histogram of all Labour sample probabilities
       xlim([0 1])
       subplot(3,1,3)
       hist(1-b41-b42) %histogram of all Alliance sample probabilities
       xlim([0 1])
       Egivenhist1=mean(b41); Egivenhist2=mean(b42);
       Predprob=[Egivenhist1; Egivenhist2; 1-Egivenhist1-Egivenhist2].';
       %estimates of the predictive probabilities of voting
       New2C(t)=r(t)*Egivenhist1;
       New2L(t)=r(t)*Egivenhist2;
       New2D(t)=r(t)-New2C(t)-New2L(t);
       New2R(t)=r(t); %subsequently predictive votes per party, as required
       Predvote=[New2C(t);New2L(t);New2D(t)].';
       New2d(t)=d1(t);
    end

    for t=1:length(M) %sorts out winners of each seat
        if ((NewC(t) > NewL(t)) && (NewC(t) > NewD(t))) s1(t)=1;
        else s1(t)=0;
        end
        if ((NewL(t) > NewC(t)) && (NewL(t) > NewD(t))) s2(t)=1;
        else s2(t)=0;
        end
    end
    Cseats=sum(s1); Lseats=sum(s2); Dseats=length(M)-Cseats-Lseats;
    CvsLvsD = [Cseats; Lseats; Dseats].';
end %now the next seat has been declared so we revise all undeclared left
```

Figure A.15: Forecasting code (*continued*).

Appendix B

Computational Statistics

B.1 Maximum Likelihood Estimation

Consider a family of probability distributions parameterized by an unknown parameter vector θ, associated with either a known probability density function (continuous case) or probability mass function (discrete case). We will denote this by f_θ. We obtain a sample $\{y_1, y_2, \ldots, y_n\}$ of n values from this distribution, and then use f_θ to compute the multivariate probability density associated with our observed data, $f_\theta(y_1, y_2, \ldots, y_n)$. As a function of θ with $y_1, y_2, \ldots, y_n$ fixed, we call this the likelihood function $L(\theta)$. The method of maximum likelihood estimates θ by finding the value of θ, $\hat{\theta}$, which maximises $L(\theta)$. This is the maximum likelihood estimate (MLE) of θ. The maximum likelihood estimator may not be unique, or may not even exist. Commonly, we assume that the data from a particular distribution are independent, identically distributed with unknown parameters. This considerably simplifies the problem, as the likelihood may then be written as a product of n univariate probability densities:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(y_i), \tag{B.1}$$

and, to simplify computation, we may take the logarithm of (B.1), since maxima are unaffected by monotone transformations such as $a = \ln(b)$:

$$l(\theta) = \sum_{i=1}^{n} \ln f_\theta(y_i). \tag{B.2}$$

Then, the maximum of (B.2) may be found numerically using various optimization algorithms, such as the Newton-Raphson method. The MLE may be supplemented by its approximate covariance matrix, derived from the likelihood function. The likelihood function itself may be used to construct improved versions of confidence intervals compared to those obtained from the approximate covariance matrix.

This contrasts with seeking an unbiased estimator of θ, which may not necessarily yield the MLE but which will yield a value that (on average) will neither tend to overestimate nor underestimate the true value of θ.

B.1.1 Properties

Invariance: If $\hat{\theta}$ is the MLE for θ, and g is any function of θ, then the MLE for $g(\theta)$ is $g(\hat{\theta})$. It maximizes the so-called **profile likelihood**.

Bias: For small samples, this may be substantial.

Asymptotics: Often, estimation is performed using a set of i.i.d. measurements, such as distinct elements from a random sample or repeated observations. Here, we are interested in the behaviour of an estimator as the number of measurements increases to infinity (asymptotic behaviour).

Under the regularity conditions stated below, the MLE has several characteristics such that we say that it is 'asymptotically optimal'. These include:

- Asymptotic unbias, i.e., bias tends to zero as $n \to \infty$;
- Asymptotic efficiency, i.e., achieves the Cramér-Rao lower bound when $n \to \infty$ (minimum mean squared error)[1]; and
- Asymptotic normality. As $n \to \infty$, the distribution of the MLE tends to the Gaussian distribution with mean θ and covariance matrix equal to the inverse of the Fisher information matrix

Some **regularity conditions** which govern this behaviour are:

1. Defined first and second derivatives of the log-likelihood;
2. A non-zero Fisher information matrix and continuous as a function of the parameter; and
3. A consistent maximum likelihood estimator.

[1] However, the Cramér-Rao bound only speaks of unbiased estimators while the MLE is usually biased.

There may exist other (nearly) unbiased estimators with much smaller variance. On the other hand, it may be that among all regular estimators, whose asymptotic distribution is not dramatically disturbed by small changes in the parameters, the asymptotic distribution of the maximum likelihood estimator is the best possible.

Cases where asymptotic behaviour does *not* hold:

1. The MLE is on its boundary of values.

2. The data boundary is parameter dependent, for example when estimating θ from a set of i.i.d. $U[0, \theta]$. Here, the MLE exists and has some good behaviour, but the asymptotics are not as outlined above.

3. Nuisance parameters. There may exist many, but this number should not increase with n.

4. Increasing information. Where i.i.d. observations does *not* hold, the information in the data must increase with n. This may not be met if there is too much dependence in the data (for example, if new observations are essentially identical to existing observations) or new independent observations have an increasing observation error.

B.2 EM Algorithm

B.2.1 Motivation

Optimisation iterative methods such as Newton Raphson, steepest ascent and simplex do not necessarily converge. In most cases, the EM algorithm converges to a stationary point of the likelihood surface. The more parameters on the likelihood surface the more local maximum points it has. As with other optimisation methods, we cannot guarantee that the EM algorithm provides the *global* maximum, but it will usually be a local maximum. In quite general cases, the EM algorithm procedure results in an increasing sequence of values for the likelihood, which ultimately converge to a stationary value.

B.2.2 Definition

To start, we will write the likelihood as

$$L(\theta; \mathbf{x}) = f(\mathbf{x}|\,\theta),$$

where θ is the parameter vector and $\mathbf{x}$ is data which are incomplete in some way. Then, we somehow choose to *augment* $\mathbf{x}$ to get $\mathbf{y}$, which is now complete data, and get subsequently a revised likelihood

$$L(\theta;\, \mathbf{y}) = g(\mathbf{y}|\,\theta).$$

Suppose that we start the iteration from $\theta^{(0)}$, that is, we choose the starting point intuitively. The EM algorithm procedure involves generating a sequence of $\{\theta^{(m)}\}$, for $m = 1, 2, \ldots$, in which each iteration has *two* stages: firstly, the expectation stage (the E-step), in which we require the general expression of

$$Q(\theta,\, \theta^{(\mathbf{m})}) = \mathbb{E}(\ln g(\mathbf{y}|\,\theta)|\,\mathbf{x},\, \theta^{(\mathbf{m})}) \tag{B.3}$$

and secondly, the maximisation stage (the M-step), at which we must obtain the value of θ, $\theta^{(\mathbf{m+1})}$, which maximises (B.3).

B.2.3 Remarks

A property of the method is that the sequence is monotonic increasing, although not always strictly, that is, $L(\theta^{(\mathbf{m+1})}) \geq L(\theta^{(\mathbf{m})})$.

Often, it is difficult to obtain expressions to go on to maximise; also integrals at the E-step are often difficult if not impossible to solve. Extensions to the EM algorithm include the generalised EM (GEM) algorithm and the stochastic EM (SEM) algorithm, which offer ways to tackle such problems (Morgan, 2000, [87]).

A complete account of the EM algorithm, including derivation, proof of convergence and numerous applications, is detailed in Dempster, Laird and Rubin (1977, [34]) and Figueiredo (2004, [44]).

Appendix C

Opinion Poll Data

Harvey and Shephard (1990, [55]) mention that opinion polls from panel surveys have been excluded, as they violate the assumption that the measurement errors are independent through time. Also excluded are surveys carried out in marginal constituencies and then reweighted to give an impression of the national vote, as they view the practice as potentially highly inaccurate. Further notes about data cleansing is in Harvey and Shephard (1990).

The data do not always sum to 100% since nationalist parties receive significant support in polls and due to rounding error by individual pollsters.

For 1983, fieldwork for the first poll started on 10^{th} May and the election was held on the 9^{th} June. For 1987, fieldwork for the first poll occurred between 6^{th} and 11^{th} May and the election was held on the 11^{th} June.

We stress the point made by Harvey and Shephard (1990) that it would have been useful had any fieldwork spread over a number of days been broken down into individual day results, thus giving us more data in order to help overcome the problem of having less data. Instead, the average of the period is what is recorded. Nevertheless, in our two illustrations the vector lengths are large enough for our programs/models to work, but our programs/models would have struggled for example in the October 1974 election, when only fifteen polls were recorded.

Poll no. (t)	Time (days)	δ (days)	Con. (%)	Lab. (%)	Lib. (%)	Poll size (r_t)
1	0.0	0.5	46.0	31.0	21.0	1047
2	2.0	2.0	49.0	34.0	15.0	964
3	3.5	1.5	46.0	33.0	19.0	946
4	6.0	2.5	44.0	37.0	17.0	1090
5	6.0	0.5	46.0	31.0	21.0	1154
6	6.5	0.5	49.0	31.0	19.0	1584
7	7.0	0.5	44.0	33.0	21.0	507
8	7.5	0.5	47.0	30.0	21.0	960
9	9.0	1.5	46.0	37.0	16.0	1100
10	9.5	0.5	45.0	36.0	18.0	1052
11	10.0	0.5	47.0	34.0	18.0	1250
12	11.5	1.5	48.0	33.0	18.0	1700
13	13.0	1.5	45.0	32.0	20.0	1071
14	13.0	0.5	51.0	33.0	15.0	1068
15	13.0	0.5	52.0	33.0	14.0	1100
16	14.0	1.0	47.5	32.5	19.0	1422
17	14.0	0.5	45.0	32.0	21.0	557
18	14.5	0.5	46.0	30.0	23.0	1023
19	15.0	0.5	49.0	31.5	18.0	2015
20	16.0	1.0	51.0	29.0	19.0	1088
21	16.5	0.5	47.0	30.0	21.0	1029
22	17.0	0.5	49.5	31.0	19.0	1325
23	17.5	0.5	47.5	28.0	23.0	918
24	20.0	2.5	41.0	30.0	24.0	1056
25	20.5	0.5	47.0	30.0	22.0	1276
26	21.0	0.5	44.0	32.0	21.0	1026
27	21.0	0.5	44.0	29.0	25.0	504
28	21.5	0.5	46.0	28.0	24.0	1038
29	22.0	0.5	45.5	31.5	22.0	1989
30	22.5	0.5	45.0	28.0	25.0	942
31	23.0	0.5	43.0	32.0	23.0	1067
32	23.5	0.5	47.0	28.0	23.0	1041

Table C.1: Election poll data in the run up to the general election of 1983.

Poll no. (t)	Time (days)	δ (days)	Con. (%)	Lab. (%)	Lib. (%)	Poll size (r_t)
33	24.0	0.5	44.0	27.0	27.5	1311
34	24.0	0.5	47.0	29.0	23.0	1074
35	26.0	2.0	45.0	24.0	28.0	1038
36	27.0	1.0	47.0	26.0	25.0	1337
37	27.5	0.5	46.0	28.0	24.0	1040
38	28.0	0.5	46.0	23.0	29.0	1100
39	28.5	0.5	45.5	26.5	26.0	2003
40	29.0	0.5	46.0	26.0	26.0	1335
41	29.0	0.5	47.0	25.0	26.0	1013
42	29.0	0.5	48.0	28.0	26.0	1101

Table C.2: Election poll data in the run up to the general election of 1983 (*continued*).

Poll no. (t)	Time (days)	δ (days)	Con. (%)	Lab. (%)	All. (%)	Poll size (r_t)
1	0.0	0.5	46.0	28.0	25.0	1735
2	0.5	0.5	39.0	28.0	30.0	1085
3	1.5	1.0	43.0	29.0	25.0	1445
4	1.5	0.5	44.0	31.0	23.0	1934
5	4.5	3.0	41.0	30.0	26.0	1020
6	5.5	1.0	42.0	33.0	23.0	1040
7	8.0	2.5	42.0	32.0	24.0	1058
8	9.5	1.5	41.0	33.0	24.0	1072
9	11.0	1.5	42.0	33.0	23.0	2640
10	11.0	0.5	45.0	36.0	20.0	1079
11	12.0	1.0	41.0	34.0	22.0	1066
12	12.5	0.5	41.0	33.0	21.0	1517
13	15.0	2.5	42.0	37.0	21.0	1075
14	17.5	2.5	42.0	35.0	20.0	1035
15	18.0	0.5	44.5	36.0	18.0	2506
16	19.0	1.0	41.0	37.0	21.0	1072
17	19.0	0.5	45.0	32.0	22.0	1067
18	19.5	0.5	44.0	32.0	21.0	1553
19	22.5	3.0	42.0	36.0	20.0	1573
20	22.5	0.5	44.0	33.0	21.0	1063
21	24.0	1.5	40.5	36.5	21.5	2553
22	25.0	1.0	44.0	33.0	21.0	1087
23	25.5	0.5	44.0	34.0	24.0	1576
24	26.0	0.5	43.0	33.0	22.0	2102
25	26.5	0.5	43.0	35.0	21.0	1065
26	29.5	3.0	45.0	32.0	21.0	1575
27	30.0	0.5	42.0	35.0	21.0	2122
28	30.0	0.5	41.0	34.0	23.5	2005
29	30.5	0.5	43.0	35.0	21.0	1086
30	30.5	0.5	43.0	34.0	21.0	1702
31	31.0	0.5	44.0	32.0	22.0	1668
32	31.5	0.5	42.0	35.0	21.0	1633

Table C.3: Election poll data in the run up to the general election of 1987.

Bibliography

[1] Abramowitz, Milton & Stegun, Irene A., *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, (New York, Dover Publications, ISBN 978-0-486-61272-0, 1972).

[2] Aigner, D. J. and Goldberger, A. S., *Latent Variables in Socio-economic Models*, (Contributions to Economic Analysis, 103, North-Holland Publishing Company, 1977).

[3] Akaike, Hirotugu, *A new look at the statistical model identification*, (IEEE Transactions on Automatic Control, 19 (6), 1974, pp716-723).

[4] Barndorff-Nielsen, O. E. and Cox, D. R., *Asymptotic Techniques for Use in Statistics*, (Monographs on Statistics and Applied Probability, Chapman and Hall, 1989).

[5] Bartolucci, Francesco, De Luca, Giovanni, *Maximum likelihood estimation of a latent variable time-series model*, (Applied Stochastic Models in Business and Industry, 17, 2001, pp5-17).

[6] Bayarri, M. J., DeGroot, M. H. and Kadane, J. B., *What is the likelihood function? (with discussion)*, (Statistical Decision Theory and Related Topics IV, Vol. 1, New York: Springer, 1988).

[7] BBC website for UK election results:
http://news.bbc.co.uk/1/hi/ukpolitics/vote2005/constituencies/default.stm.

[8] Bean, Louis H., *How to Predict Elections*, (Greenwood Press, Publishers, Westport, Connecticut, 1948).

[9] Bierens, Herman J., *Information Criteria and Model Selection*, (Pennsylvania State University, 2006).

[10] Bilmes, Jeff A., *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*,

(International Computer Science Institute, University of California, Berkeley, 1998).

[11] Bjørnstad, J. F., *On the generalization of the likelihood function and likelihood principle*, (Journal of the American Statistical Association, 91, 1996, pp791-806).

[12] Blumenthal, M., *More pollsters interviewing by telephone*, (Pollster.com, 2008).

[13] Blumenthal, M., *New Pew data on cell phones*, (Pollster.com, 2008).

[14] Blumenthal, M., *Questions About Exit Polls*, (Pollster.com, 2008).

[15] Bollen, Kenneth, A., *Structural Equations with Latent Variables*, (John Wiley & Sons, 1989).

[16] Bollerslev, Tim, *Generalized Autoregressive Conditional Heteroskedasticity*, (Journal of Econometrics, 31, 1986, pp307-327).

[17] Bondon, Pascal, *Influence of Missing Values on the Prediction of a Stationary Time Series*, (Journal of Time Series Analysis, Vol. 26, 4, 2005, pp519-525).

[18] Box, George and Jenkins, Gwilym, *Time series analysis: Forecasting and control*, (San Francisco: Holden-Day, 1970).

[19] Box G. E. P. & Pierce, D. A., *Distribution of residual autocorrelations in autoregressive-integrated moving average time series models*, (Journal of the American Statistical Association, Vol. 65, No. 332, 1970, pp1509-1526).

[20] Box, M. J., Draper, N. R. and Hunter, W. G., *Missing values in multiresponse nonlinear data fitting*, (Technometrics, 12, 1970, pp613-620).

[21] Brockwell, P. and Davis, R., *Introduction to Time Series and Forecasting*, (Second Edition, Springer, New York, 2002).

[22] Brown, P. J., Firth, D. and Payne, C. D., *Forecasting on British election night 1997*, (Journal of the Royal Statistical Society, A, 162, Part 2, 1999, pp211-226).

[23] Brown, Philip and Payne, Clive, *Election Night Forecasting*, (Journal of the Royal Statistical Society, Series A (General), Vol. 138, No. 4, 1975, pp463-498).

[24] Brown, Philip and Payne, Clive, *Forecasting the 1983 British general election*, (The Statistician, 33, 1984, pp217-228).

[25] Brown, Philip, J. and Payne, Clive, D., *Aggregate Data, Ecological Regression, and Voting Transitions*, (Journal of the American Statistical Association, Vol. 81, No. 394, Theory and Methods, 1986, pp452-460).

[26] Brubacher, S. R. and Wilson, G. T., *Interpolating time series with application to the estimation of holiday effects on electricity demand*, (Journal of Applied Statistics, 25(2), 1976, pp107-116).

[27] Burnham, Anderson, *Model Selection and Inference - A practical information-theoretic approach*, (ISBN 0-387-98504-2, 1998).

[28] Burnham, K. P. and Anderson D. R., *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*, (Second Edition, Springer-Verlag, ISBN 0-387-95364-7, 2002).

[29] Cantril, Hadley and Strunk, Mildred, *Public Opinion, 1935-1946*, (Princeton University Press, 1951).

[30] Cox, D. R. and Miller, H. D., *The Theory of Stochastic Processes*, (Chapman & Hall, 1977).

[31] Crowder, M. J., *Beta-binomial Anova for proportions*, (Applied Statistics, 27, 1978, pp34-37).

[32] Curtice, John and Firth, David, *Exit polling in a cold climate: the BBC-ITV experience in Britain in 2005*, (Journal of the Royal Statistical Society, Series A, Vol. 171, Part 2, 2008, pp1-25).

[33] Cuzán, Alfred G., Armstrong, J. Scott and Jones, Jr, Randall J., *Combining Methods to Forecast the 2004 Presidential Election: The Pollyvote*, (2004).

[34] Dempster, A. P., Laird, N. M. and Rubin, D. B., *Maximum Likelihood from Incomplete Data via the EM Algorithm*, (Journal of the Royal Statistical Society, Series B (Methodological), Vol. 39, No. 1., 1977, pp1-38).

[35] Dickey, D. A. and Fuller, W. A., *Distribution of the Estimators for Autoregressive Time Series with a Unit Root*, (Journal of the Americal Statistical Association, 74, 1979, pp427-431).

[36] Durbin, J. and Koopman, S. J., *Monte Carlo maximum likelihood estimation for non-Gaussian state space models*, (Biometrika, 84, 3, 1997, pp669-684).

[37] Durbin, J. and Koopman, S. J., *Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian persepectives*, (Journal of the Royal Statistical Society, B, 62, Part 1, 2000, pp3-56).

[38] Electoral Calculus website: http://www.electoralcalculus.co.uk/index.html.

[39] Enders, W., *Applied Econometric Time Series*, (John-Wiley & Sons, New York, 1995).

[40] Engle, Robert F., *Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation*, (Econometrica, 50, 1982, pp987-1008).

[41] Engle, Robert F., *GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics*, (Journal of Economic Perspectives, 15(4), 2001, pp157-168).

[42] Engle, R. F., *ARCH: selected readings*, (Oxford University Press, ISBN 0-19-877432-X, 1995).

[43] Ferreiro, O., *Methodologies for the estimation of missing observations in time series*, (Statistics and Probability Letters 5, 1987, pp65-69).

[44] Figueiredo, Mario A. T., *Lecture Notes on the EM Algorithm*, (www.stat.duke.edu/courses/Spring06/sta376/Support/EM/EM.Mixtures. Figueiredo.2004.pdf, 2004).

[45] Findley, David F., Monsell Brian C., Bell, William R., Otto, Mark C., Chen, Bor-Chung, *New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program*, (Journal of Business and Economic Statistics, 1998, 16, pp127-177).

[46] Gelman, Andrew, Carlin, John B., Stern, Hal S. and Rubin, Donald B., *Bayesian Data Analysis*, (London: Chapman and Hall, First Edition, 1995, Chapter 11).

[47] Geyer, Charles, J. and Thompson, Elizabeth, A., *Constrained Monte Carlo Maximum Likelihood for Dependent Data*, (Journal of the Royal Statistical Society, Series B (Methodological), Vol. 54, No. 3, 1992, pp657-699).

[48] Goldstein, H., *Multilevel Statistical Models*, (Arnold, London, 1995).

[49] Goodman, Leo, A., *Analyzing qualitative/categorical data: log-linear models and latent-structure analysis*, (Cambridge, MA.: Abt Books, 1978).

[50] Griliches, Zvi, *Errors in variables and other unobservables*, (Econometrica, Vol. 42, 1974, pp971-998).

[51] Ha, I. D. and Lee, Y., *Estimating frailty models via Poisson hierarchical generalized linear models*, (Journal of Computational and Graphical Statistics, 12, 2003, pp663-681).

[52] Ha, I. D., Lee, Y. and Song, J. K., *Hierarchical likelihood approach for frailty models*, (Biometrika, 88, 2001, pp233-243).

[53] Ha, I. D., Lee, Y. and Song, J. K., *Hierarchical likelihood approach for mixed linear models with censored data*, (Lifetime Data Analysis, 8, 2001, pp163-176).

[54] Hagenaars, Jacques, A., *Loglinear models with latent variables*, (Sage Publications, Inc., 1993).

[55] Harvey, A. C. and Shephard, N. G., *Tracking the Level of Party Support During General Election Campaigns*, (Department of Statistical and Mathematical Sciences, London School of Economics and Political Science, 1990).

[56] Hawkes, A. G., *An Approach to the Analysis of Electoral Swing*, (Journal of the Royal Statistical Society, Series A, Vol. 132, 1969, pp68-79).

[57] Henderson, C. R., *Best linear unbiased estimation and prediction under a selection models*, (Biometrics, 31, 1975, pp423-447).

[58] Hong, Han and Preston, Bruce, *Bayesian Averaging, Prediction and Nonnested Model Selection*, (NBER Working Paper Series, Working Paper 14284, National Bureau of Economic Research, 2008).

[59] Hurvich, C. M. and Tsai, C.-L., *Regression and time series model selection in small samples*, (Biometrika, Vol. 76, 1989, pp297-307).

[60] Hyndman, R. J., *Time Series Data Library*,
(http://www.robhyndman.info/TSDL, data accessed in 2006).

[61] Johansen, Adam, Doucet, Arnaud and Davy, Manuel, *Maximum Likelihood Parameter Estimation for Latent Variable Models Using Sequential Monte Carlo*, (Acoustics, Speech and Signal Processing, Vol. 3, 2006, pp640-643).

[62] Johansen, Adam M., Doucet, Arnaud and Davy, Manuel, *Particle Methods for Maximum Likelihood Estimation in Latent Variable Models*, (Statistics and Computing, Vol. 18, No. 1, 2008, pp47-57).

[63] Kendall, Maurice and Ord, J. Keith, *Time Series*, (Edward Arnold, Third Edition, 1990).

[64] Kuk A. Y. C and Cheng Y. W, *Pointwise and functional approximations in Monte Carlo maximum likelihood estimation*, (Statistics and Computing, Vol. 9, 1999, pp91-99).

[65] Lee, Yoon Dong, Yun, Sungcheol and Lee, Youngjo, *Analyzing weather effects on airborne particulate matter with HGLM*, (Environmetrics, Vol. 14, 7, 2003, pp687-697).

[66] Lee, Youngjo and Nelder, John A., *Conditional and marginal models: another view (with discussion)*, (Statistical Science, 19, 2004, pp219-238).

[67] Lee, Youngjo and Nelder, John A., *Fitting via alternative random-effect models*, (Statistics and Computing, Vol. 16, 2006, pp69-75).

[68] Lee, Youngjo and Nelder, John A., *Hierarchical Generalized Linear Models: A synthesis of generalised linear models, random effect model and structured dispersion*, (Biometrika, 88, 2001, pp987-1006).

[69] Lee, Youngjo and Nelder, John A., *Hierarchical Generalized Linear Models (with discussion)*, (Journal of the Royal Statistical Society, B Met, Vol. 58, 1996, pp619-678).

[70] Lee, Youngjo and Nelder, John A., *Likelihood for random-effect models*, (Statistics and Operations Research Transactions, Vol 29 (2), 2005, pp141-164).

[71] Lee, Youngjo, Nelder, John A. and Noh, Maengseok, *H-likelihood: problems and solutions*, (Statistics and Computing, Vol. 17, No. 1, 2007, pp49-55).

[72] Lee, Youngjo, Nelder, John A. and Pawitan, Yudi, *Generalized Linear Models with Random Effects, Unified Analysis via H likelihood*, (Monographs on Statistics and Applied Probability 106, Chapman and Hall/CRC, 2006).

[73] Lee, Youngjo, Noh, Maengseok and Ryu, Keunkwan, *HGLM modelling of dropout process using a frailty model*, (Computational Statistics, Vol. 20, No. 2, 2005, pp295-309).

[74] Lewis-Beck, Michael S., *Election Forecasting: Princples and Practice*, (The British Journal of Politics and International Relations, Vol. 7, 2005, pp145-164).

[75] Lewis-Beck, M., *Election forecasts in 1984: how accurate were they?*, (PS: Political Science and Politics, 18: 1, 1985, pp53-62).

[76] Lindsey, J. K., *On h-likelihood, random effects, and penalised likelihood*, (Biostatistics, Limburgs Universitair Centrum, Diepenbeek).

[77] Lindsey, J. K. and Lambert, P., *On the appropriateness of marginal models for repeated measurements in clinical trials*, (Statistics in Medicine, 17, 1998, pp447-469).

[78] Little, Roderick J. A. and Rubin, Donald B., *Statistical Analysis with Missing Data*, (John Wiley and Sons, Inc. 2002).

[79] Liu, Q. and Pierce, D. A., *Heterogeneity in Mantel-Haenszel-type models*, (Biometrika, 80, 1993, pp543-556).

[80] Ljung, G. M., *A note on the estimation of missing values in time series*, (Communications in Statistics. Simulation and Computation, 18(2), 1989, pp459-465).

[81] Ljung, G. M. and Box, G. E. P., *On a Measure of a Lack of Fit in Time Series Models*, (Biometrika, 65, 1978, pp297-303).

[82] McQuarrie, A. D. R. and Tsai, C.-L., *Regression and Time Series Model Selection*, (World Scientific, 1998).

[83] Mena, Ramses H. and Walker, Stephen G., *Stationary Mixture Transition Distribution (MTD) models via predictive distributions*, (Journal of Statistical Planning and Inference, Vol. 137, 10, 2007, pp3103-3112).

[84] Mena, Ramses H. and Walker, Stephen G., *Stationary Autoregressive Models via a Bayesian Nonparametric Approach*, (Journal of Time Series Analysis, Vol. 26, Issue 6, 2005, pp789-805).

[85] Miller, Irwin, Miller, Marylees, *John E. Freund's Mathematical Statistics Sixth Edition*, (Pearson Education, 1999).

[86] Miller, W., *Measures of electoral change using aggregate data*, (Journal of the Royal Statistical Society, A, 135, 1972, pp122-142).

[87] Morgan, Byron J. T., *Applied Stochastic Modelling*, (Arnold Publishers, 2000).

[88] Nelder, J. A. and Wedderburn, R. W. M., *Generalized linear models*, (Journal of the Royal Statistical Society, Ser. A, 135, 1972, pp370-384).

[89] Nelson, D. B., *Conditional heteroskedasticity in asset returns: A new approach*, (Econometrica, 59, 1991, pp347-370).

[90] Noh, M. and Lee, Y., *REML estimation for binary data in GLMMs*, (Journal of Multivariate Analysis, 98, 2007, pp896-915).

[91] Payne, Clive, Brown Philip and Hanna, Vincent, *By-election Exit Polls*, (Electoral Studies, 5:3, 1986, pp277-287).

[92] Perron, P., *The great crash, the oil price shock and the unit root hypothesis*, (Econometrica, 57, 1989, pp1361-1401).

[93] Phillips, P. C. B. and Perron P., *Testing for a Unit Root in Time Series Regressions*, (Biometrika, 75, 1988, pp335-346).

[94] Pitt, Michael, K., Chatfield, Chris and Walker, Stephen G., *Constructing First Order Stationary Autoregressive Models via Latent Processes*, (Scandanavian Journal of Statistics, 29, 2002, pp657-663).

[95] Pitt, Michael, K. and Walker, Stephen G., *Constructing Stationary Time Series Models Using Auxillary Variables with Applications*, (Journal of the Americal Statistical Association, Vol. 100, No. 470, 2005, pp554-564).

[96] Pitt, Michael, K. and Walker, Stephen G., *Extended constructions of stationary autoregressive processes*, (Statistics & Probability Letters, 76, 2006, pp1219-1224).

[97] Pourahmadi, M., *Foundations of Time Series Analysis and Prediction Theory*, (New York: Wiley, 1976).

[98] Rose, Richard, *Opinion Polls as Feedback Mechanisms: From Cavalry Charge to Electronic Warfare*, (Britain at the Polls 1983, Durham: Duke University Press, 1985, pp108-138).

[99] Rozanov, Yu. A., *Stationary Random Processes*, (San Francisco: Holden-Day, 1967).

[100] Sanders, D., *Government popularity and the next general election*, (Political Quarterly, 62, 1991, pp235-261).

[101] Schwarz, G., *Estimating the dimension of a model*, (Annals of Statistics, 6(2), 1978, pp461-464).

[102] Sims, C. A., *Macroeconomics and Reality*, (Econometrica 48, 1980, pp1-48).

[103] Singh, A. C. and Roberts, G. R., *State space modelling cross-classified time series and counts*, (International Statistical Review, 60, 1992, pp321-335).

[104] Smith, T. M. F., *Principles and problems in the analysis of repeated surveys*, (Survey Sampling and Measurement, New York, Academic Press, 1978).

[105] Thoning, Kirk, W., Tans, Pieter, P. and Komhyr, Walter, D., *Atmospheric carbon dioxide at Mauna Loa Observatory 2. Analysis of the NOAA GMCC data, 1974-1985*, (Journal of Geophysical Research, Vol. 94, 1989, pp8549-8565).

[106] UK Elect website: http://www.ukelect.com/index.html.

[107] UK Election Results website: http://www.election.demon.co.uk/.

[108] Whiteley, P., *Electoral forecasting from poll data: the British case*, (British Journal of Political Science, Vol. 9, 1979, pp219-236).

[109] Whiteley, Paul F., *Forecasting Series from Votes in British General Elections*, (The British Journal of Politics and International Relations, Vol. 7, 2005, pp165-173).

[110] Yates, F., *The analysis of replicated experiments when the field results are incomplete*, (Empire Journal of Experimental Agriculture, 1, 1933, pp129-142).

[111] Yule, G. Udny, *On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers*, (Philosophical

Transactions of the Royal Society of London, Ser. A, Vol. 226, 1927, pp267-298).

[112] Yun, Sungcheol and Lee, Youngjo, *Comparison of hierarchical and marginal likelihood estimators for binary outcomes*, (Computational Statistics & Data Analysis, Vol. 45, Issue 3, 2004, pp639-650).