Data Mining of Early Day Motions and Multiscale Variance Stabilisation of Count Data

Daniel John Bailey

University of BRISTOL

School of Mathematics

September 2008

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science

Abstract

This thesis consists of two parts: an exploration of new measures of backbench opinion in the UK House of Commons, and an exploration of variance stabilising transformations of count data.

In the first part, we consider the use of Early Day Motions (EDMs) as a means of gauging opinions of Members of Parliament (MPs) over a range of issues. A much used measure of opinion is that of cohesion; how similar MPs from each political party are to each other. We define a new cohesion measure using the signatories of Early Day Motions and explore this measure over a moving time period for each of the main political parties.

We then use Early Day Motions for feature selection. We first identify issues which cause individual parties to be more or less cohesive with one another, before setting out methodology to distinguish which issues cause the major political parties to differ in opinion.

We then turn our attention to methods of variance stabilisation of count data. Using data of the number of deaths of coalition forces in Iraq, we demonstrate the good variance stabilisation which the data-driven Haar-Fisz transform possesses. We then modify this transformation so that data with negative counts can be variance stabilised. We show its good performance for simulated data and demonstrate its practical use on the central England temperature data set.

Finally, we set about incorporating a transformation parameter into the Haar-Fisz methods, so that through the use of maximum likelihood techniques, the transformation primarily attempts to normalise the data, rather than variance stabilise it.

Dedication

Dedicated to the memories of Cecil Bailey, Nachama Sadeh and Alexey Blokonenkov.

Acknowledgements

I would like to thank my supervisor, Professor Guy Nason for his help and encouragement over the past three and a half years.

Thanks to my family for their love and support and to my friends for their welcome distractions. In particular, thanks to Jaymie for his patience in proof reading and Matt for help and advice on just about everything!

A special thanks to Laura Dennis, for sticking by me.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the Regulations of the University of Bristol. The work is original except where indicated by special reference in the text. No part of the dissertation has been submitted for any other academic award. All views expressed in the dissertation are those of the Author.

SIGNATURE

DATE

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is in two parts: the first, Chapters 2–4 consider the use of Early Day Motions to gauge backbench opinion in the UK House of Commons. Chapters 5–8 consider methods of variance stabilisation of count data.

1.1 Early Day Motions

Early Day Motions (EDMs) have been a much used tool by British politicians to convey an opinion or to support the view of other Members of Parliament (MPs). Although popular amongst MPs, the use of EDMs to statistically gauge opinion has been somewhat contentious. Their vast subject area and *cheap-talk* nature drew concerns over their applicability and reliability, with critics often overlooking the wealth of information contained within the data.

In Chapter 2 we review the means by which an MP can cast a vote on a particular issue. Divisions are introduced and their *whipped* nature discussed. We review the role of divisions in calculating the *cohesion* of the major political parties, as well as the usefulness (if not limited number) of *unwhipped* divisions. We then introduce EDMs, detailing them historically and reviewing their use in measuring backbench opinion in the House of Commons. We end by reviewing some recent techniques developed to model legislators using roll call data.

The work in Chapters 3 and 4 makes extensive use of the EDM data set. Acquiring this data was key to this research and the detailed processes of downloading, converting and coding the EDMs is only briefly described within this thesis. Having carried out this task, the data has been made publicly available to allow for further research into the EDM data set.

We propose the use of EDMs in calculating the cohesion of major political parties in Chapter 3. We give a detailed review of the criticism and support which EDMs have received in the past. We bring this debate up to date and show that although one signature on an EDM may be cheap-talk, hundreds of signatures on thousands of EDMs constitute a rich body of information. A new cohesion measure is then defined and applied to EDMs over a moving time window. We use this cohesion measure for feature selection, picking out issues which cause the parties to be more, or less cohesive.

Chapter 4 details further data mining applications with EDMs. In contrast to work in Chapter 3, we suggest looking for issues which cause the political parties to be less similar to each other, essentially discovering the issues which cause division *between* parties. We also include a brief investigation into how an MPs propensity to sign EDMs manifests itself within the data.

1.2 Variance Stabilisation

The remaining chapters of this thesis concentrate on the variance stabilisation of count data. This data is often ‘Poisson like’, and periods of high signal intensity are often coupled with a higher degree of variability. This can cause problems with many smoothing methods, which assume a certain degree of Gaussianity within the data. We turn our attention to transformations which Gaussianise, and variance stabilise data.

In Chapter 5 we review literature used in the remaining chapters of this thesis. We introduce the discrete wavelet transform, as well as some methods of smoothing data. We summarise some models for time series data where the counts are assumed to be drawn from a Poisson distribution. Some variance stabilising, and Gaussianising transformations are then reviewed. In particular, the Box & Cox (1964) transformation is given in detail, along with methods to estimate the transformation parameters. The recently developed Haar-Fisz transformation by Fryzlewicz & Nason (2004), and the data-driven Haar-Fisz transform by Fryzlewicz *et al.* (2007) for variance stabilisation are described in detail. These are used extensively in the following chapters.

We use the data-driven Haar-Fisz transform (DDHFT) in Chapter 6 to stabilise the variance of counts of the mortality levels of coalition forces in Iraq and use a range of smoothing methods to estimate the underlying intensity. We compare the transform to that of Box & Cox (1964) in terms of variance consistency and Gaussianity of residual variance. We find the DDHFT outperforms the Box-Cox transform, and results in better intensity estimates than the ‘running-mean’ techniques currently being used.

Chapter 7 considers modifications of the DDHFT for when the data includes negative counts. Whereas it is common practice to add a constant to the data for use with the Box-Cox transform, this is not always appropriate for the DDHFT. Furthermore, the choice of this constant for the Box-Cox transform can be problematic. We suggest two modifications of the DDHFT, depending on assumptions about the data, and suggest a bootstrap test for deciding the most appropriate of these two transforms. We show our methods perform better than the Box-Cox transform over a range of test signals and then apply them to the central England temperature dataset — annual temperature measurements often used in climatological studies.

The work in Chapter 8 further modifies the Haar-Fisz transforms so that it’s primary goal is good Gaussianisation, rather than variance stabilisation. A general Haar-Fisz transform is defined in which a transformation parameter is to be estimated to optimise Gaussianisation. Similar to the Box-Cox transform, we use maximum likelihood techniques to estimate this parameter. The work outlined is initial and many possible extensions are left as future work. Nevertheless, the methods described show the potential for these transforms to be effective Gaussianisers.

Finally, Chapter 9 summarises the work of this thesis and outlines future work and extensions of the methods and applications presented within.

Chapter 2

Literature Review I

2.1 Introduction

This chapter reviews literature which involves the quantitative and qualitative analysis of Members of Parliament (MPs) in the UK House of Commons in terms of their voting behaviour on parliamentary roll calls and other such devices. These devices come in the form of parliamentary *divisions* and the less formal *Early Day Motion* (EDM).

Our review of divisions focuses mainly on their use for the measure of *cohesion*, either between individual MPs or between political parties in the UK. Other measures of cohesion which have been applied to non-UK legislators are described in Chapter 3.

Although we do not use the division lists as a source of data in this thesis, they play an important role in terms of work which has been carried out using them, and giving the reader an understanding of the freedom of expression British politicians have.

We then review the use of EDMs as a measure of backbench opinion in the House of Commons. We give an account of the rise in popularity amongst MPs to use EDMs as well as a brief history of these relatively unknown parliamentary devices. We detail the characteristics of EDMs which make them interesting to use (spontaneous signing, unwhipped nature), and review work which use EDMs to analyse backbench opinion.

Finally, we review some recent techniques using spatial models which have been applied to legislators from many parliaments around the world.

2.2 Divisions

A *division* is the term given when the House of Commons votes on a particular issue. These divisions give MPs the chance to cast their vote on laws and various pieces of legislation. Party discipline, however, plays a huge role in how MPs vote, with the *party whips* having a tight control over MPs. Parliamentary Factsheet P9, (House of Commons Information Office (2003b)) gives full details as to the history and procedure of divisions.

Division lists were analysed by Lowell (1919) to investigate the decline of independent voting from 1836 to the end of the 19th century. Cox (1987) derived various tables of figures from those reported in Lowell's original work, and examined party discipline over the divisions. He showed that by the end of the study period, the number of divisions which had the party whip had doubled to around 90 per cent. Cox also used Lowell's data to calculate a cohesion measures for the political parties during the period of study. An index of cohesion was defined to compare the intra-party cohesion of each party on both whipped and unwhipped divisions. The measure used was expressed algebraically by McLean (1995) and the cohesion of MPs from party i is given by

$$C_i = \frac{\sum_{k=1}^{n} 2[(V_{i,k}^{\mathrm{maj}}/V_{i,k}) - 0.5]}{n}, \tag{2.2.1}$$

where k is a division in the session, with $k = 1, 2, \ldots, n$. It is assumed that the parties are labelled $1, \ldots, i, \ldots, m$, although i in (2.2.1) is arbitrary. $V_{i,k}^{maj}/V_{i,k}$ is the ratio of votes cast by the majority of MPs within party i on division k to all votes cast by MPs of party i on division k. The ratio can thus range from 0.5 (when half the MPs make up the majority) to 1, when all MPs vote the same way. C_i is thus scaled to range from 0 to 1, with 1 being perfect cohesion.

For whipped votes, Cox (1987) found that cohesion increased markedly over the study period whereas unwhipped votes showed no such trend.

Berrington (1968) pointed out the frailties of this cohesion measure: not infrequently, the front benches did not vote in the same way as the majority of their party and instead relied on support of the 'opposition' backbenchers over their own. This meant that those making up the value of $V_{i,k}^{\mathrm{maj}}$ in (2.2.1) may not be voting according to the (front bench) party line, as implied.

In recent years, with vast data sets from the Houses of Parliament being available over the internet, websites such as `www.publicwhip.org.uk` have been created to automatically download division lists and present the user with this information in a more transparent form. Their aim is to make MPs more accountable for their actions by allowing the public to identify how a particular MP has voted. A feature of their analysis involves identifying 'rebellious' MPs, that is, when an MP votes against the party whip. The site has also used multidimensional scaling to obtain an idea of party structure and see where MPs lie in relation to each other in terms of dissimilarity of voting.

The work on the site is informative and potentially useful to someone researching voting patterns by MPs in general or on particular issues, yet the voting patterns revealed are not that surprising. With the party whips having such strong hold over the voting of MPs, any rebellion would likely to already be known to them. High profile rebels may also promote their stance by telling journalists and other MPs, making the results published largely known beforehand.

2.2.1 *Unwhipped* Divisions

Divisions which are known to be free from the party whip are known as *free votes*. Free votes generally occur when voting on issues such as the running of parliament, issues of individual conscience or when the whips are no longer able to enforce a party line. Free votes allow an MP to vote independently and in line with how they truly believe. Although informative, free votes are nowadays rare. They have, however, been subject to quantitative analysis.

We look in particular at work which uses free votes to calculate cohesion of the parties. Read *et al.* (1994) used free votes to look at how MPs from the three main parties voted on the issues of homosexuality and capital punishment. The measure, for a single free vote is defined by:

$$C_i' = \frac{V_i^{\mathrm{maj}} - V_i^{\mathrm{min}}}{m}, \tag{2.2.2}$$

where C_i' is the cohesion for MPs in party i and V_i^{maj} and V_i^{min} are the number of votes which make up the majority and minority of the party respectively. By excluding non-voters, we have $V_i^{\mathrm{maj}} + V_i^{\mathrm{min}} = m$, allowing for the cohesion, or the *Index of Party Unity* (IPU) to range from 0 to 1.

Compared to (2.2.1), the IPU is the cohesion on a specific division, rather than a mean cohesion over a session. The IPU does not include non-voters, so makes less assumptions about the behaviour of those MPs, whereas (2.2.1) classifies them separately and allow for the possibility for $V_{i,k}^{maj}$ to be the comprised of MPs who abstained. Thus, if we were to consider a session with only one division (i.e n = 1 in (2.2.1)), and where all MPs voted without abstention, we have $C_i = C_i'$.

Cowley & Stewart (1997) used the IPU to calculate the cohesion of the main parties when they voted in free votes between 1979 and 1996. They defined a cohesive party as one which has an IPU of 0.80; a divided party to be one with an IPU of below 0.80 and a party to be considered seriously split if their IPU falls below 0.33. The free votes which are analysed are considered by the author to be *conscience* votes and they conclude that they *followed* party lines (i.e. each party behaves independently) and that it is rare for all parties, within themselves, to be split on an issue.

Free votes, by their very nature are different to other divisions in the House of Commons. Although they give MPs a chance to vote without the party whip, they are limited in number and generally restricted in content. Furthermore, as previously stated, Cox (1987) found that there was no observable trend of cohesion as the number of unwhipped divisions rose. Cox also believed that even for unwhipped divisions, party pressures may still have been present and affecting the cohesion (reasons for this are given by Cox (1987, page 25)). The unknown element of unwhipped divisions leads the author to concentrate only on those division which were known to be whipped.

We discuss some further measure of cohesion, which were applied to non-UK legislators in Section 3.1.1.

2.3 Early Day Motions

Early Day Motions (EDMs) are spontaneous, unwhipped motions which MPs can table and support free from the party pressures which are associated with divisions. Much of the information about the history of Early Day Motions is detailed in the Parliamentary Factsheet P3 (House of Commons Information Office (2003a)). Full details of the procedure governing how an MP proposes, signs or amends an EDM are given, along with details of types, signature levels and even the cost to the tax payer of printing and publishing EDMs. Finer *et al.* (1961) briefly details the history of EDMs and discusses in more detail the reasons for an MP to sign, or not to sign a given motion. Here, we review the history of EDMs and their *unwhipped* nature in comparison to Divisions. Details of the procedure, types of EDMs and reasons for MPs to sign them are further discussed in Chapter 3.

The current procedure for tabling an Early Day Motion has been in place since 1943, although the idea of proposing a motion with no fixed date for debate started to evolve nearly 100 years before. Prior to this, there had been ample time for Members to raise matters of interest to the House in order for debate, usually by the member simply announcing that they were to raise such a question. In the 1850s and 60s, at the end of a session, when it was impossible to set a date for such a debate, the practice of informing Members that they wished to raise such matters in the future developed. These had no fixed date, but would be intended to be debated in the next session, or at an early opportunity.

By 1865, the daily Notice (or Order) Paper which Members received would commonly have a separate section headed Notices of Motions. Some were intended for debate, others just an expression of opinion. At this time, other Members would submit the same motion as a sign of support to the original. The process evolved so that a new name to a motion did not warrant resubmitting the entire motion (although a new number was attached to the name). By the 1940s, EDMs were sometimes seen to attract hundreds of signatures and for ease of reference, the number was attributed to the motion, and not to the names of each supporter.

The phrase "For An Early Day" was appended to such motions in the 1940s, with the notion (or indeed fiction), that the motion was for serious debate at the earliest opportunity in the future. This was the origin of their modern name, which became the name of the section where they were printed in the Notice Paper.

The rise of popularity of EDMs stemmed from more time being taken up in House of Commons by government. In the 1940s this was never moreso; the implications of war meant that time for Private Members' Motions and Bills was no longer available. It was around this time that the popularity of EDMs, as a means of expressing an opinion started to soar. It is reported that in the 1950s there were approximately 100 EDMs each Session, rising to about 400 in the 70s and 700 by the early 80s. The thousand mark was first broken in 1983 and by the end of the century there were around 1400 per session. It is now common to see in excess of 2000 EDMs tabled per session.

Unlike Divisions, EDMs are *unwhipped*, that is, there is no pressure put on an MP by their party to submit or sign a given motion. EDMs are, in their very nature cheap-talk and Finer *et al.* (1961) give reasons, (apart from actually agreeing or disagreeing) for why an MP would choose to sign, or not to sign a given motion. This has been the cause of much criticism into the use of EDMs to gauge political opinion and is discussed in more detail in Section 3.1.2

2.3.1 Early Day Motions as a Measure of Backbench Opinion

Backbench Opinion in the House of Commons 1955-59, by Finer *et al.* (1961) introduced the idea of using EDMs to gauge the opinion of the backbenches. The follow-up study by Berrington (1973), looks at the earlier period of 1945–55 and also utilises other information, such as floor revolts, open letters and free votes. The vast majority of data and analysis still come from EDMs and the other forms of backbench expression afford Berrington, where appropriate, to confirm findings drawn from the EDMs.

The authors look at the two main political parties, Labour and Conservative to investigate where Members stand on certain issues of the day, such as foreign affairs, social welfare and penal reform. Information about MPs' backgrounds were used to investigate such issues, comparing views of the MPs to information such as their educational background, occupation before entering the Commons and whether they had served national duty. Finer *et al.* (1961), and Berrington (1973), used cross-tabulations and Guttman scaling to test the significance of the variables and to look for a single dimension running through certain EDMs (i.e., relationship between advocates of European Unity with those who agree with the 'wider' concept of world federation).

Finer *et al.* (1961) also produced a list of the '50 most left Labour MPs', which proved to be a controversial element of the study. Critics viewed the list as almost absurd, with names appearing who were clearly known to be be less 'left' than others not appearing on the list. We detail the reviews which Berrington's work received in Section 3.1.2, summarising both the positive and negative feedback which the work provoked.

Following from the work by Berrington (1973), Franklin & Tappin (1977) further investigate the use of EDM as an unobtrusive measure of backbench opinion. The authors consider some of the criticism which Finer *et al.* (1961) received and discuss the issues surrounding the use of EDMs in such a study. The work uses answers from 72 MPs given during an interview in 1964. A wide variety of questions were asked, ranging from the respondent's backgrounds and attitudes to political life, as well as questions involving pertinent issues of that time. EDMs on the same subject, which were signed by the MPs in question, were used to compare answers to signatures.

Given this information, the authors looked for similarities and differences between the two expressions of opinion (interview and EDM response), and defined an error rate of using EDMs to predict a Members opinion. This, they comment, "is surely a great deal lower than would have been expected by those commentators who have criticized the use of EDMs as indicators of opinion".

The authors then consider how EDMs could be used to gauge the opinion of non-signers. In doing so, they define two models which account for the way in which an EDM gains support. The first model categorises EDMs which are readily signed by those MPs in favour of the motion. An example given of such types of EDMs are those covering the topic of nuclear disarmament. The second type of EDM defined in the study are those in which the signatures appear to be gained more randomly. EDMs on the common market and government control of the economy are given by the authors as likely topics to attract such random support. For such random signing topics, the authors conclude that a larger number of EDMs would be needed to judge the opinion of non-signers.

Leece & Berrington (1977) used EDMs to study the Labour party during the 1968–69 parliamentary session. Similar to earlier work by Berrington, Guttman scales are used to scale attitudes within the Labour Party. The authors are careful in choosing which EDMs to analyse, and set out criteria for an EDM to be included in the study.

EDMs are considered in pairs. They either represented opposing views, or more commonly a more 'extreme' view along a scale (i.e. to see if those MPs who thought family allowance should be 40 pence a week also agreed that 25 pence should be given). For each pair of EDMs, an association table is produced to look at the numbers of MPs signing each motion. The table is as follows:

	EDM 1 signed	EDM 1 not signed
EDM 2 signed	a	b
EDM 2 not signed	c	d

Where a is the number of MPs who signed both EDMs 1 and 2, b is the number who didn't sign EDM 2 but did sign EDM 1 and so on. A similar tabulation is later used in Section 3.1.3 to compare pairs of MPs based on the EDMs they signed.

Given this representation of two EDMS, they were both judged suitable for inclusion into the study based upon two measurements: the Yule Q score, $(ad - bc)/(ad + bc)$ and the similarity ratio, $a/(a+b+c)$. The former, which ranges from -1 to 1, is an association measure between EDMs and measures how likely a supporter of an 'extreme' EDM will sign the less extreme EDM (or indeed, if they sign neither). A high value is obtained if MPs consistently voted for or against a given viewpoint, so a and d would be large. This indicates that the MPs are concordant. For a given EDM, if the Yule's Q score of those who signed it was greater than 0.8, it was considered for the study.

The similarity ratio between EDMs based on their signature levels is based upon the Jaccard measure of dissimilarity, as later defined in Section 3.1.3. It is measure of similarity of EDMs, ranging from 0 to 1 which does not include MPs who signed neither EDM. The higher the value, the more similar the EDMs are. If this is greater than 0.25 for a pair of EDMs, then they were considered for the study.

Out of those EDMs considered for the study based on the Yule's Q and similarity scores, those on a similar topic and which had similar numbers of signatures as others considered were then actually included. This careful selection of EDMs allowed the authors to validate scales with the voting pattern of MPs in the division lobbies. They found that the MPs investigated had rebelled against the party via their voting behaviour on various occasions.

Similar to previous studies, biographic attributes were also investigated in relation to the scales produced. Results included showed how trade-union sponsored members had become more 'left wing' since 1959 (this was explained by how the party was selecting candidates, rather than the candidates themselves changing their views).

The increased use of computer power differentiates the work of Leece & Berrington (1977) to earlier studies. They indicate using multidimensional scaling and cluster analysis to group EDMs and that they perform their statistical calculations on computer. Scaling solutions are used in Chapter 3 to show party structure, and in Chapter 4 as part of feature selection of divisive issues between parties.

The only further publication using this original data set was Berrington (1982), who looks in particular at how the 'left' of the Labour party has changed over the years, using methodology introduced in previous work.

Nason (2001) examined EDMs from a modern standpoint, utilising advances in computational power to mine the data. The work is exploratory and the author admits that the focus is on answering interesting questions, rather than on statistical significance of the results. That said, the work is an insightful reintroduction to the use of EDMs along with modern statistical applications. Data visualisation software is used to display multidimensional scaling plots over a moving time period, focusing on the interaction between the three main parties. These plots are used, for example, to focus on the position of the Liberal Democrat leaders, Charles Kennedy and his predecessor Paddy Ashdown and their relationship with the Labour Party. Also shown are how classification trees can be used to classify MPs (or would-be voters) into a political party, given their signing (or non signing) of given EDMs.

The most recent published work using EDMs is that by Childs & Withey (2004) which studies the differences in signing patterns between the sexes and whether women are more likely to sign 'women's' EDMs. The authors use chi-squared tests on the response of MPs to either ‘sign’ or ‘not sign’ and conclude that female MPs did indeed act for women, by signing for women.

2.4 The OC and NOMINATE Procedures

We end this chapter with a review of some of the most recent techniques of modelling parliamentary voting, in terms of spatial positions of legislators, called the OC and NOMINATE procedures.

Poole (2005) describes both the non-parametric optimal classification (OC) method for spatial modelling of legislators in parliament and an approach to parametric classification which is dubbed the NOMINATE procedure (*NOMINA*l *T*hree-Step *E*stimation). Other models based on the same approach are also described (namely the D-, W- and DW-NOMINATE procedures). This work brings together a multitude of research papers, as well as building on the framework set out by Poole & Rosenthal (1997).

Both the OC and NOMINATE algorithms analyse parliamentary data, from which it is inferred that such data is the outcome of a set of legislators voting either Yea or Nay on a given number of roll calls. An error is introduced into a legislators choice by using a *random utility model*, which assumes that their utility for a Yea or Nay vote is the sum of a deterministic utility function and a random error. Legislator i’s utility for the Yea outcome (denoted by y) on roll call j is given as:

$$U_{ijy} = u_{ijy} + \varepsilon_{ijy}, \tag{2.4.3}$$

where u_{ijy} is the deterministic portion of the utility function and ε_{ijy} is the random portion.

If there were no error, the legislator votes Yea if $U_{ijy} > U_{ijn}$, i.e. if the difference $U_{ijy} - U_{ijn}$ is positive. With random error, this difference is given by,

$$U_{ijy} - U_{ijn} = u_{ijy} - u_{ijn} + \varepsilon_{ijy} - \varepsilon_{ijn},$$

so the legislator votes Yea if,

$$u_{ijy} - u_{ijn} > \varepsilon_{ijy} - \varepsilon_{ijn}.$$

The OC method only assumes that legislators have *symmetric single-peaked* utility functions. That is, if a legislator ideally votes on the 'centre' ground, they are equally likely (with probability defined by the utility function) to vote in favour of roll calls which are the same *distance* either *left* or *right* of their ideal spatial (central) position.

With no error present, a spatial map of the legislators can be obtained simply by using multidimensional scaling.

When error is introduced into the decision making of the legislators, this scaling solution is only one of many possible representations. Therefore the scaling solution may not be the best set of coordinates to represent the data, and the OC method is developed to tackle this problem.

Given an initial set of spatial coordinates (or *ideal points*) of the legislators (given their votes on the roll calls), the OC method first finds a *cutting point* or *plane* for each roll call. The cutting points or planes split the ideal points of the legislators with Yea votes on one side and Nay on the other. Given the known voting patterns of the legislators and their fixed ideal points, the cutting plane is such that the number of erroneously classified legislators is a minimum.

The second step is to estimate new ideal points of the legislators given the cutting point (or plane). That is, given that the cutting plane remains fixed, new spatial coordinates of the legislators are found to further reduce error of classification. This process is repeated until convergence of cutting planes and ideal points. The OC method therefore assumes there is error present, but does not attempt to model it.

The NOMINATE procedures considers the distribution of the utility function in (2.4.3) and attempts to estimate the functional form of both the deterministic portion, u_{ijy} and the random (or stochastic) portion ε_{ijy}.

The deterministic part of the utility function is assumed to follows a Gaussian distribution, although Poole (2005) also derives models for the simpler case of it following a quadratic distribution. The random part of the utility function is also assumed to follow a Gaussian distribution, although two other models, the uniform and the logit have also been used (and are fully referenced by Poole (2005)).

Given the distributional assumptions, the distribution of the difference between the utility for Yea and the utility for Nay for the ith legislator on the jth roll call is derived and shown to be Gaussian with constant variance, given by,

$$U_{ijy} - U_{ijn} \sim N(u_{ijy} - u_{ijn}, \sigma^2).$$

The NOMINATE procedure works as follows. Firstly, a reasonable set of ideal points are generated with which an initial set of roll call parameters (which determine the position of cutting plane) are found, *given* the ideal points. Next, *given* estimates of the roll call parameters, better estimates of the ideal points are found. The third set of parameters to be found are the utility function parameters, which are estimated *given* both the legislator ideal points and the roll call outcome points. Estimation of the three sets of parameters are cycled through until convergence.

These two methods can be thought of doing the following. Say we have coordinates to represent MPs (i.e. the scaling solution) over a range of issues and for a given issue, we wish to draw a line to separate those who were for or against the issue. We can then see which MPs have been erroneously classified on either side of the line (and possibly subject those MPs to further scrutiny). There is a question, however, that if given another vote on the same set of issues, would the MPs vote the same way as previously? If not, their position in the scaling solution and the cutting plane could change. The error term accounts for this by allowing small ‘movements’ of the MP positions on the scaling solution. This may be such that they change sides of the cutting line. The process is repeated to minimise the number erroneously classified.

We consider the use of these procedures in relation to EDMs and detail why they are not directly suitable, or usefully adaptable to use in the British political system in Section 3.3.

Techniques involving the notion of categorising MPs into parties, and using the number of erroneously classified within an optimising criteria is developed in Chapter 4.

Chapter 3

Cohesion of Major Political Parties

3.1 Introduction

Cohesion of political parties or groups of legislators is of key interest to political analysts and commentators. These measures have been largely used to investigate legislators in the US Congress and Senate, and more recently have been developed for use within the European Parliament (EP). The application of such measures to political parties in the UK is limited due to the strong political pressures which are put on Members of Parliament (MPs) to vote according to their party line. Even *free votes* (where MPs vote according to their true beliefs) are not without their critics. In this chapter, which is based on the paper by Bailey & Nason (2008), we revisit the idea of using Early Day Motions as a measure of backbench opinion and review the criticism to which they have been subjected to in the past. We argue that although there is a degree of ‘uncertainty’ in the reason for an individual’s signature of an EDM, the effect of this in the analysis of EDMs diminishes as the number of motions studied increases.

We develop a cohesion measure based on the *asymmetric* signing of EDMs and use this to investigate the cohesion of major political parties in the UK. Finally, we use modern statistical techniques and utilise computational power to investigate the issues which are associated with cohesion and separation within political parties, via an exploratory method which highlights the modern statistical method of ‘data-mining’.

3.1.1 Measures of Cohesion

The vast literature on party cohesion and discipline mostly involves analysis of roll calls in the US House of Congress. The work is extensive, and we direct the reader to the comprehensive overview of the literature by Owens (2003) and the more recent book by Hazan (2005) (in which the Owens article forms a chapter). We give a brief review of some cohesion measures in the literature (in particular those not covered by Owens (2003), or detailed in Chapter 2), and focus on attempts to calculate the cohesion of political parties in the UK.

A much used cohesion measure is introduced by Rice (1928). This *index of voting likeness* within a political party is defined as

$$100\sum_{j=1}^{n}\left|\frac{NYeas_j - NNays_j}{NYeas_j + NNays_j}\right|, \tag{3.1.1}$$

where *NYeas*$_j$ is the number of voters in a given party who vote *Yea* on vote j and *NNays*$_j$ the number of those voting *Nay*. An index of voting likeness of 1 indicates that all votes within a party voted the same way across all n votes. A value of 0 indicates that the party was split, over all votes, with half voting *Yea* and the other half voting *Nay*.

Many measures are similar to, or based upon Rice's measure, such as the *Agreement Index* (AI) by Hix *et al.* (2005) which further makes allowance for the legislator to abstain from a vote. This is a rescaling of the *Index of agreement* by Attiná (1990), such that cohesion values range from 0 to 1. As with the work by Attiná, the cohesion measure is used to investigate the European Parliament.

Rahat (2007) defines a cohesion measure also based on Rice's but which treats an abstention as a 'halfway' vote between Yes or No. Furthermore, the measure only includes the number of abstentions in the numerator of the cohesion measure if this was the majority vote. The measure is thus defined by two formulas; the first when the majority of a party votes for or against a bill, the second when the majority abstains. These two measures are defined respectively as:

$$\frac{|NYeas_j - NNays_j|}{N} \qquad \text{and} \qquad \frac{NAbs_j}{N},$$

where the notation is the same as Rice's measure in (3.1.1) and where $NAbs_j$ is the number of voters abstaining and N is the total number of votes (including abstentions). This measure is used to analyse the cohesion of the Israeli Parliament.

There have been many attempts to calculate cohesion scores for members of parliament (MPs) in the UK House of Commons. The difference with MPs compared to their American counterparts is the party discipline exerted by the *party whips* to force them to vote according to the party line. This was not always the case, as we described in Section 2.2

Unwhipped divisions, also known as *free votes*, have been of some interest to political researchers. They allow an MP to vote independently (usually on issues concerning the running of parliament or issues of individual conscience), but although informative, are nowadays rare. Section 2.2.1 gives more details on these divisions, as well as cohesion measures used to analyse them.

Is it possible to gauge cohesion levels in the House of Commons? Free votes do not have the problems associated with their *whipped* counterparts, but their number and subject matters are both limited. Furthermore Cox (1987, page 25), felt that although unwhipped, party pressures were still evident in these divisions. For other divisions, cohesion levels merely inform us as to how well the party whips are doing their job. Low cohesion may signify unrest within the party, but this would already be known by the party whips. We instead reopen the case for using Early Day Motions (EDMs) as a source of information on backbench opinion: a much used device by MPs that allow them spontaneous and unwhipped opinions on a variety of subject matters.

Section 2.3 introduced EDMs and their historical context. We next give further information about EDMs and consider the criticism which has been levelled at them in the past, in reference to the works by Finer *et al.* (1961) and Berrington (1973).

3.1.2 Early Day Motions

An Early Day Motion (EDM) is traditionally a motion put down by a Member of Parliament (MP) calling for a debate on a particular subject. The number of EDMs has increased in recent years, however, they are rarely debated (see Appendix A.1.1 for an example of a debated EDM and Appendix A.1.2 for a recent EDM). The modern-day purpose of EDMs is to allow MPs to express their opinion on a subject and to canvass support for their views by inviting other members to add their signatures in support of the motion.

An EDM takes the form of a single sentence, no more that 250 words long and beginning "That this house..." as it must be of the form of a resolution (House of Commons Information Office, 2003a). EDMs are submitted to the House by an MP on a specially printed form with space for six main sponsors and 50 further names. Any MP can initiate an EDM, although Ministers, whips, the speaker and deputies generally do not. Recall that unlike most votes in the House of Commons, EDMs are unwhipped; that is, there is no pressure put on an MP *by their party* to sign it. EDMs could therefore be viewed as useful by political researchers as they give an objective indication of what that MP truly believes.

EDMs fall into several groups. Opposition EDMs are put down by the opposition against a government policy, Rebellion EDMs may be put down by members of a party which express a view different from that of the party concerned, and "all-party" EDMs express views across party divides (often on social issues which have been promoted by one party but attract signatures from MPs of different political allegiance). Factsheet P3, House of Commons Information Office (2003a), gives more detail on these types of EDMs and gives examples of EDMs which may not fall into such groups (for example, an EDM criticising another Member of the House, or the House of Lords).

An EDM will remain current throughout the entire parliamentary session. An amendment EDM can be made by a different MP other than the initiator of the original. Amendments can be made at any time during the session, and can either oppose or strengthen the view offered by the original EDM. If an MP wishes to table an amendment for an EDM which they have already signed, they must first withdraw their name from the main motion.

The first major works involving the statistical exploration of EDMs was by Finer *et al.* (1961) and Berrington (1973). The original work attracted much attention due to its provocative aims of using EDMs to gauge backbench opinion in the House of Commons. Initial reviews of the work, as gauged by both Bromhead (1962) and Lloyd (1977), were written by journalists and journalistically inclined politicians. Three such reviews were those by Crossman (1961), Fellows (1962) and Howard (1962), whose hostile reactions left a stain on the work. In reply to these critics, Berrington (1973) devoted the entire first chapter of the second book on this subject to a defence of the work.

The majority of reviews published in peer reviewed journals were more positive. None completely dismiss the issues which arise with the collecting of signatures of EDMs and thus their cheap-talk nature, but as both Bromhead (1962) and Richards (1962) comment, the authors "fully recognise the issues which arise" and they concur that the cheap-talk nature does not rule out EDMs as an important source of information. This view is neatly summarised by Turner (1963), who believes that "some of the patterns that emerge stand out too clearly to be ignored". These patterns, including the strong party structure shown by Nason (2001), indicate the irrefutable wealth of information this data contains. Although a single EDM may be cheap-talk, it is certainly the case that several thousand EDMs collectively contain important and discoverable information. In the most recent review of Berrington's work, McLean (1995) discusses the contribution made to political analysis over Berrington's career. Work of a similar nature is explored, as well as how the work and ideas in *Backbench Opinion* were derived. The article discusses the hostility of the initial reviews toward the work and Berrington's reaction to them. Also McLean (1995) noted that the battles that Berrington had fought over the use of statistical methodology had now been won and that opportunities now exist to do far more with his data than was easily possible in the 1960s and 1970s.

As mentioned in Section 2.3.1, Finer *et al.* (1961) comments on possible reasons for MPs not to sign EDMs. These reasons include how the original sponsor collects signatures, as an EDM with an active 'business-like' sponsor who asks MPs for their signatures is more likely to receive a large number of signatures. That said, the sponsor may be after the signature of certain, influential MPs, rather than a large number. Furthermore, some MPs will be of the type who sign few, or no EDMs. With the current trend in numbers of EDMs per session and the time an MP has to read them all, this latter reason is still pertinent.

The cheap-talk nature of EDMs was the main cause of concern for reviewers who questioned their validity to reveal information about the British political system. It is true that there are many factors which influence whether an MP will show support for a particular motion and, as McLean (1995) comments, these factors may indeed be frivolous. They may not, however, be clear. As an example, consider a motion which congratulates a particular football team's success. This seemingly *cross-party* EDM could be called to question because an MP whose constituency forms the fan base of a rival team may not sign it to avoid increasing sporting tensions and, in some cases, sectarian violence. This, however, is not always the case; MPs may sign or propose such congratulatory motions as a statement to show that the rivalry should remain only on the football pitch (this can be seen in EDMs signed by the Everton MP, Louise Ellman, following Liverpool Football Club's success in the 2005 Champions League finals). We use this example to demonstrate the complexity of EDMs and to highlight that perceived reasons for signing may be wrong and indeed opposite to the perceived truth. There are many reasons for MPs to sign a particular motion, but it is not valid to dismiss their content entirely.

The non-signing of EDMs may be as frivolous as the signing of them. As discussed in Section 2.3, there are many reasons for an MP to sign (or not sign) an EDM. If an MP has signed a particular EDM, it is reasonable to assume that they are committed to that point of view. The absence of a signature on an EDM does not, however, imply that a given MP disagrees with that EDM. As Finer *et al.* (1961) point out, reasons for this may be that the MP was not canvassed for their signature (if they do not usually sign EDMs at will), or indeed that the MP may not sign any EDMs regardless of opinion. Finer *et al.* (1961) test whether certain type-classes of MPs (for example, with particular educational or occupational backgrounds) are more likely to sign particular types of EDMs, rather than the process being at random. Assuming this random signing pattern, the authors calculate the distribution of signatures (and non-signatures) of MPs of different type-classes across the EDMs. These expected frequencies were then compared to the observed frequencies and the chi-squared test was used to establish statistical significance of the signing patterns. The authors report a significant, and in some cases highly significant association between the substance of the Motion and the type-class of MP signing it and conclude that a lack of signature is not (statistically) due to the canvassing of that particular EDM.

We do not delve into type-classes for this study. As our work uses dissimilarity measures between MPs, we instead ensure that MPs who do not sign any EDMs are not included in the study. For other patterns of signing, we ensure that MPs are not compared on EDMs which *neither* sign.

We further consider the cheap-talk nature of the signing of EDMs by viewing it as adding *noise* to the data. If this ‘noise’ has a large effect, we would not expect there to be strong party structure within the data. To briefly illustrate the party structure, we perform classical multidimensional scaling on a dissimilarity matrix of MPs for EDMs tabled during the 2005/06 parliamentary session (details of methods are found in Chatfield & Collins (1996) and in Section 3.1.3 of this chapter). Figure 3.1 shows the first two dimensions of the scaling solution, wherein we see the strong grouping which exists within the data. The first dimension appears to somewhat split the Conservatives from the other parties whilst Labour and the Liberal Democrats are similar yet still distinct. Other minority parties have more in common with the Liberal Democrats than either of the other two main parties. Investigating higher dimensions of the scaling solution further supports our findings that political parties are split. The plot in Figure 3.1 shows strong structure within the data; MPs are not forced to sign motions along party lines but nevertheless often do so (note that we discuss the ‘horseshoe’ effect within this plot in Section 4.2). Within the vast amount of data on EDMs there is definitely useful information to be found; modern statistics and computational power can assist the process of discovery. Although one signature on an EDM may be cheap-talk, hundreds of signatures on thousands of EDMs constitute a rich body of information.

3.1.3 Obtaining and Analysing EDM Data

All EDMs signed since the start of the 1989/90 parliamentary session, including amendments can be found on the EDM website `www.edm.ais.co.uk`. The site contains information on all EDMs proposed including their content, date tabled and their supporters. Unfortunately, the list of names of signers for EDMS between 1989–1992 are incomplete and thus not used for analysis within this thesis.

Having downloaded the relevant web pages, the data is converted into matrix form, with columns and rows representing each EDM (by number) and MP names respectively (see Appendix A.2 for details of downloading and conversion of data). Data on MPs signing EDMs from each parliamentary session are stored in $n \times p$ matrices, with binary entries:

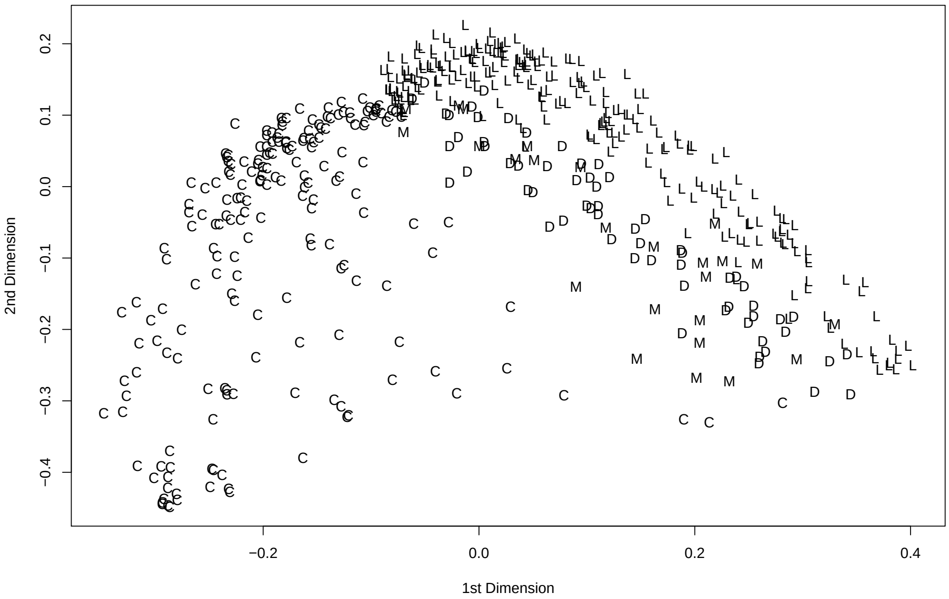


Figure 3.1: Classical multidimensional scaling solution of 05/06 EDM data. L = Labour, C = Conservative, D = Liberal Democrat, M = other Minority parties.

$$m_{ij} = \begin{cases} 1, & \text{if MP } i \text{ signed EDM } j \text{ for } i = 1, \ldots, n; j = 1, \ldots, p. \\ 0, & \text{otherwise,} \end{cases} \tag{3.1.2}$$

Table 3.1 shows the top left of such a matrix. We have produced several such matrices, one for each session from 1992 to 2005 which can be found on the website

`www.maths.bris.ac.uk/∼db0797/Research.html`.

Nason (2001) investigated the relationship between MPs and between political parties by defining a dissimilarity coefficient between pairs of MPs. The Jaccard coefficient of dissimilarity was used (see Chatfield & Collins (1996, page 195)) as it reflects the important feature of EDMs of having to ‘opt-in’ to agree with the motion. Failure to sign does not necessarily indicate disagreement with that motion (Finer *et al.* (1961, pages 9–10), discuss many reasons for the varying levels of signatures that an EDM receives). We modify this measure to create a separate dissimilarity coefficient for each EDM type. Our dissimilarity coefficient between MPs (i,j), denoted by D_{ij} will then be a weighted average of the respec-

	EDM number								
MP names	1	2	3	4	5	6	7	8	...
Abbott/Diane	1	0	0	0	0	0	0	0	...
Afriyie/Adam	0	0	1	0	0	0	0	1	...
Ainsworth/Bob	0	0	0	0	0	0	0	0	...
Ainsworth/Peter	0	0	0	0	0	0	0	1	...
Alexander/Danny	1	0	0	1	1	1	0	1	...
...	...	...	...	...	...	...	...	...	

Table 3.1: Top left of data matrix for 2005/6 session

tive EDM type distances (see Friedman & Meulman (2004)). Our measure, which we call a weighted average Jaccard coefficient is defined as follows.

Let each MP i be categorised by N EDMs, ($\mathbf{x}_i = x_{i1}, x_{i2}, \ldots, x_{ik}, \ldots, x_{iN}$). The EDMs are classified into T different types with the weight of each EDM type being denoted by $\mathbf{w} = \{w_k\}_1^T$. We define the Jaccard coefficient between any pair of MPs on EDMs of type k as follows: given two MPs let a_k be the number of EDMs of type k that *both* MPs sign and define b_k, c_k and e_k according to the following table:

	MP1 signs	MP1 doesn't sign
MP2 signs	a_k	b_k
MP2 doesn't sign	c_k	e_k

The Jaccard coefficient $d_{ijk} = (b_k + c_k)/(a_k + b_k + c_k)$ is then used to measure the dissimilarity between MPs i and j for EDMs of type k.

The dissimilarity coefficient between MPs i and j over all T types of EDMs is denoted by D_{ij} and defined as the sum of the product of the Jaccard dissimilarity coefficients for each EDM type, and its weighted average:

$$D_{ij} = \sum_{k=1}^{T} w_k d_{ijk}, \tag{3.1.3}$$

with

$$\{w_k \geq 0\} \text{ and } \sum_{k=1}^{T} w_k = 1,$$

for EDMs of type $k : 1 \ldots T$.

Hence a dissimilarity matrix derived from the EDM data is obtained, with entry D_{ij} representing the dissimilarity between MP i and MP j, based upon their averaged weighted dissimilarity over each EDM type.

We next consider a party of MPs of size n. The overall level of similarity, or cohesion for those MPs can be calculated by first defining the mean dissimilarity that MP i has with all other MPs in the party:

$$M_i = \frac{1}{n-1} \sum_{j \in \{1,\ldots,n\} \setminus \{i\}} D_{ij}. \tag{3.1.4}$$

The mean dissimilarity of all MPs with each other, within the party, can then be calculated by further taking the mean of the M_i's over all n MPs:

$$\overline{M} = \frac{1}{n} \sum_{i=1}^{n} M_i \qquad \text{and let} \qquad \hat{C} = 1 - \overline{M}. \tag{3.1.5}$$

The quantity $\overline{M}$ in equation (3.1.5) is a measure of overall separation. We therefore define $\hat{C}$ as the cohesion measure, which takes values between 0 and 1, the larger the number, the stronger the coherence between MPs within that party.

Note that MPs who did not sign any EDMs over each period of interest were not included in these calculations as, by definition, they would have perfect dissimilarity with all other MPs who signed at least one EDM (and have an undefined dissimilarity coefficient with other non-signing MPs). We also do not remove any EDMs which may be considered by some as irrelevant, for example sporting EDMs. There are many reasons for this which we touched upon toward the end of Section 3.1.3. It is the wealth of subject matter, and vast number of EDMs, which reduces the impact of the cheap-talk nature of individual EDMs. We direct the reader ahead to table 3.4, which shows the number of EDMs of different types for each session of the previous Parliament. This shows the vast number of different types of EDMs which make up the dataset and how they are split into different *types*. For a so-called 'irrelevant' type of EDM to be tabled numerous times within a 100 EDM time window and for it to have a large effect on cohesion, it is very unlikely to have be irrelevant in the first place.

3.2 Cohesion of Political Parties

The overall cohesion measure for each of the three main political parties and all parties combined (from the 1992 parliamentary session to date) is shown in Figure 3.2. Note that a parliamentary session (the time between state opening and dissolution) commonly covers an entire year, starting around November, with many short recesses during the year and a longer summer recess around late July. A short session, caused by an early dissolution for a general election is common (as with the May 2005 election) and is often followed by a longer than usual session (such as the 2005/06 session). Other general elections during the time period of our data are April 1992, May 1997 and June 2001.

We use the dissimilarity measure from equation (3.1.3) with all EDMs but with $k = 1$ (i.e. all EDMs are the same type). This is the standard Jaccard coefficient of dissimilarity of the coefficient). From the plot, we can see that the Liberal Democrats are more cohesive overall than the other two parties. This might be expected: with far fewer MPs it is much easier for them to agree with each other (we discuss this in more detail later). The Conservative and Labour parties have more comparable cohesion measures and which of the two is more cohesive alternates throughout the study period. From the closely fought general election of 1992, both party's cohesion levels have fluctuated. Generally Labour have decreased in annual cohesiveness whilst the Conservatives have increased. Both show variation in cohesion levels during Labour's first term following the 1997 election victory.

The cohesiveness measure gives a static feel of the situation for each parliamentary year in its entirety. To see how this measure changes throughout a session, we look at the cohesion averaged over a moving 'time window' of EDM's. For example we calculate cohesion for EDMs $1 - 100$, $2 - 101$ and so on over the entire session. This idea was used by Cromwell (1982) to analyse MP behaviour on division lists over a period of time.

As EDMs remain open over the whole of the parliamentary session, the use of a moving time window may seem inappropriate. A time window of EDMs during the beginning of the session may contain signatures which were only been placed during the last day of the

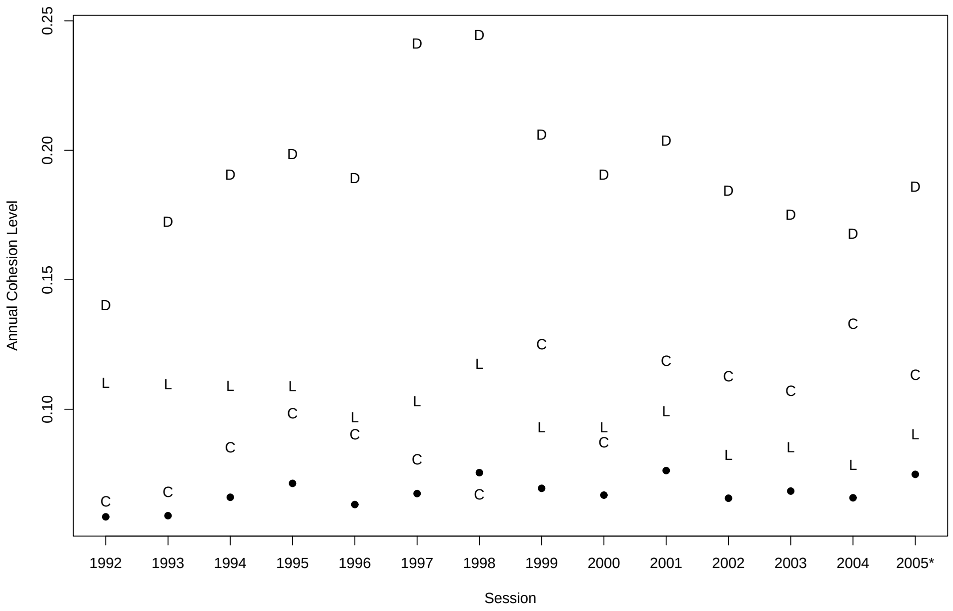


Figure 3.2: Annual coherence of main political parties: L = Labour; C = Conservative; D = Liberal Democrat; ● = All parties (*up to 25/7/06).

session. This occurrence, however, is seldom seen and most EDMs receive the majority of their signature over the first few weeks of them being tabled. Using a large enough time window (of 100 EDMs) ensures the effect of any ‘slow’ EDM is small and considering time periods in months, rather than weeks or days will insure that any misrepresentation of the data is at a minimum. Some information will inevitably be analysed as ‘belonging’ to a different time period but this will have minimal effect on our findings.

As well as the size of the time window, or bandwidth, we can also adjust the overlap between consecutive windows. This step size is the number of EDMs we ‘step over’ each time to get to our next time window. We focus on the 2005/06 session. The moving cohesion plots with a time window of 100 EDMs can be seen in Figure 3.3 and are discussed in Section 3.2.3.

We present each party on separate axes for clarity. This also allows the reader to compare the *trend* of the cohesion rather than the actual value. This is desired due to the possibility of the coherence value being affected by party size. We investigated the relationship between the cohesion of a party and the number of EDMs they signed for each time window

	Labour (264)		Conservative (195)		Liberal Democrat (63)	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
$N_{\text{MP}}/5$	0.094	0.010	0.115	0.015	0.200	0.020
$2N_{\text{MP}}/5$	0.093	0.007	0.118	0.010	0.197	0.011
$3N_{\text{MP}}/5$	0.095	0.004	0.117	0.007	0.199	0.016
$4N_{\text{MP}}/5$	0.094	0.003	0.117	0.004	0.199	0.004

Table 3.2: Mean and standard error (S.E.) of cohesion of 100 random samples from main political parties (N_{MP} in brackets). The mean propensity to sign remains constant (to 2 decimal places) over all sample sizes for each of the parties (Labour = 0.10, Conservatives = 0.06 and Liberal Democrat = 0.16).

(the number signed is a proxy for both the party size and propensity to sign). The Labour party had a weak correlation of 0.1 but there appeared to be evidence of slight positive correlation for the Conservatives (0.25) and a stronger relationship for the Liberal Democrats (0.4). For this parliamentary session, the number of (actively signing) members in a party appears to increase the chance of correlation between the party size and cohesion levels. We must therefore ensure that any analysis of cohesion levels between parties takes into account this possible underlying structure. We achieve this by developing *calibration* levels in order to gauge if a party's cohesion is higher or lower than may otherwise be expected.

3.2.1 Simple Random Sampling of MPs

We briefly investigate how the cohesion of each of the political parties is related to the number of MPs within that part which the measure is based upon. For each of the three political parties, of size N_{MP} we randomly sample $\lfloor n \rfloor$ MPs, where $n = \frac{N_{\text{MP}}}{5}, \frac{2N_{\text{MP}}}{5}, \frac{3N_{\text{MP}}}{5}, \frac{4N_{\text{MP}}}{5}$ (where $\lfloor n \rfloor$ is the largest integer less than or equal to n). We then calculate the cohesion (3.1.5) of these n MPs, given their support of EDMs. For each value of n, we take 100 different random samples from the data and calculate the mean cohesion and standard errors as given in table 3.2. For each party, the cohesion remains almost constant (to 2 decimal places) over all sample sizes. We also note that the mean propensity to sign (Number of Signatures / (Number of MPs × Number of EDMs) also remains constant for each party, no matter the sample size.

We next turn our attention to the effect that party size and propensity to sign EDMs has on the cohesion measure. We test these by simulating a time window of 100 EDMs. Let

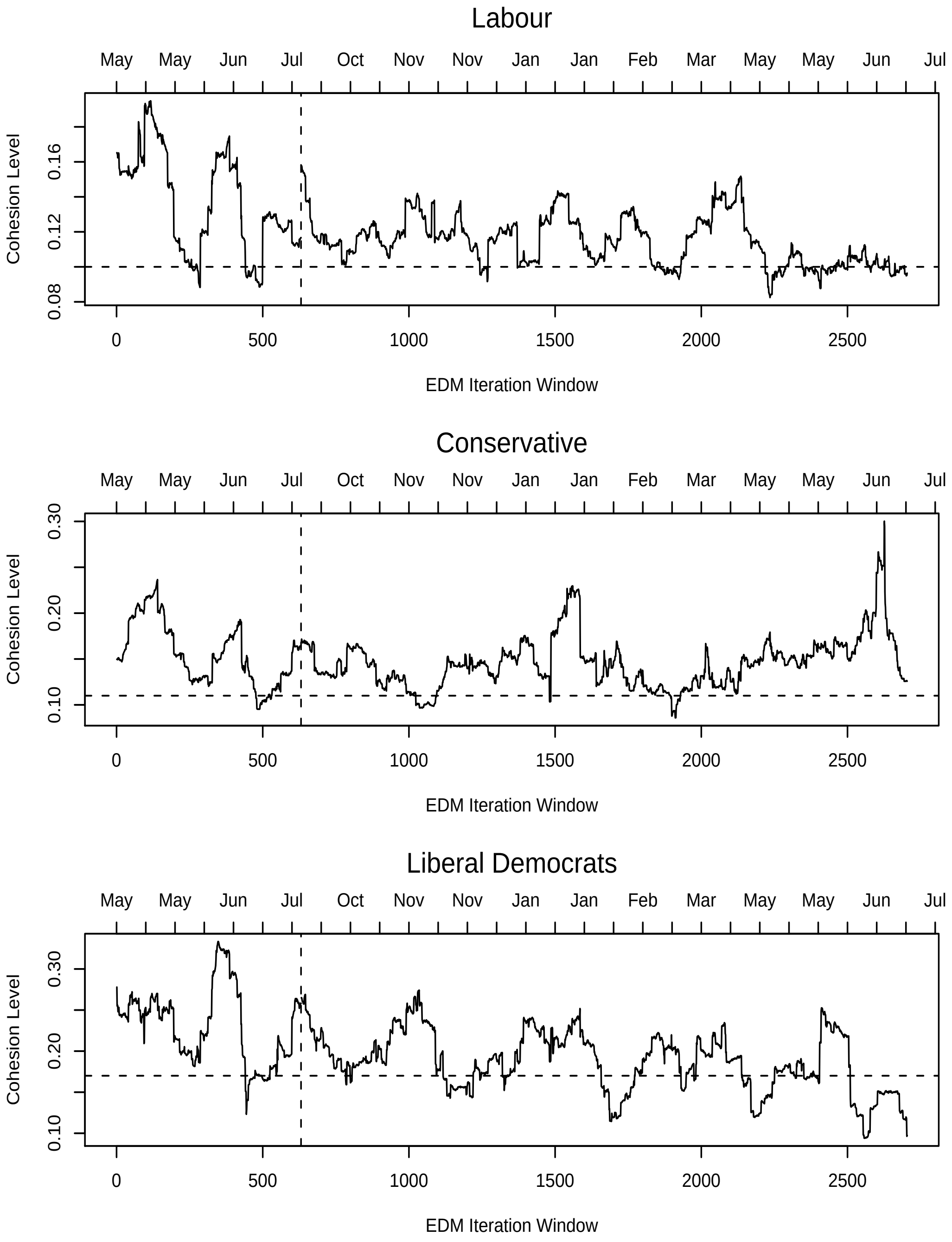


Figure 3.3: Cohesion of main political parties with a moving time window of *bandwidth* 100 and *step size* 1, during the 2005/06 session (up to summer recess 25/7/06). Month on top axis taken from ‘middle’ EDM in time window. Horizontal dashed lines represent the signature based simulated cohesion level, as described in Section 3.2.2, vertical dashed line is the 2005 summer recess.

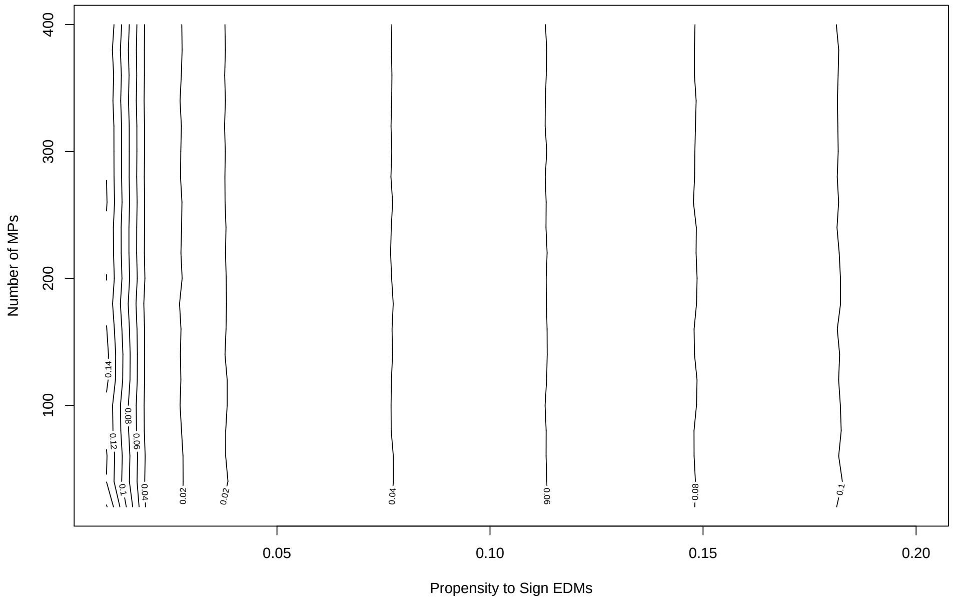


Figure 3.4: Contour plot of mean cohesion measure of simulated incidence matrices.

m_{ij}, represent whether MP i signs EDM j as defined in (3.1.2). We model m_{ij} by:

$$m_{ij} \sim Bernoulli(p), \tag{3.2.6}$$

where p is the propensity for MPs to sign EDMs. Setting p equal to 0.5 implies that each EDM has an equal chance of being signed or not by a given MP and would thus be expected to gain support from roughly half of the MPs.

For this study we set $j = 1, \ldots, 100$ and $i = 1, \ldots, N_{\text{MP}}$ where N_{MP} ranges from 20 to 400 to represent different party sizes. For each of these party sizes, we allow a different propensity to sign, p, ranging from 0.01 to 0.1 (these cover the range of propensities which we later report for the main political parties). For each combination of propensity to sign and number of MPs, we create 100 simulated incidence matrices and take the mean of the cohesion measure from (3.1.5). A contour plot of the mean cohesion measures is given in figure 3.4. We clearly observe that the (mean) cohesion depends on the propensity to sign, and not the number of MPs within the party.

We further note that as the propensity to sign increases, the cohesion quickly decreases

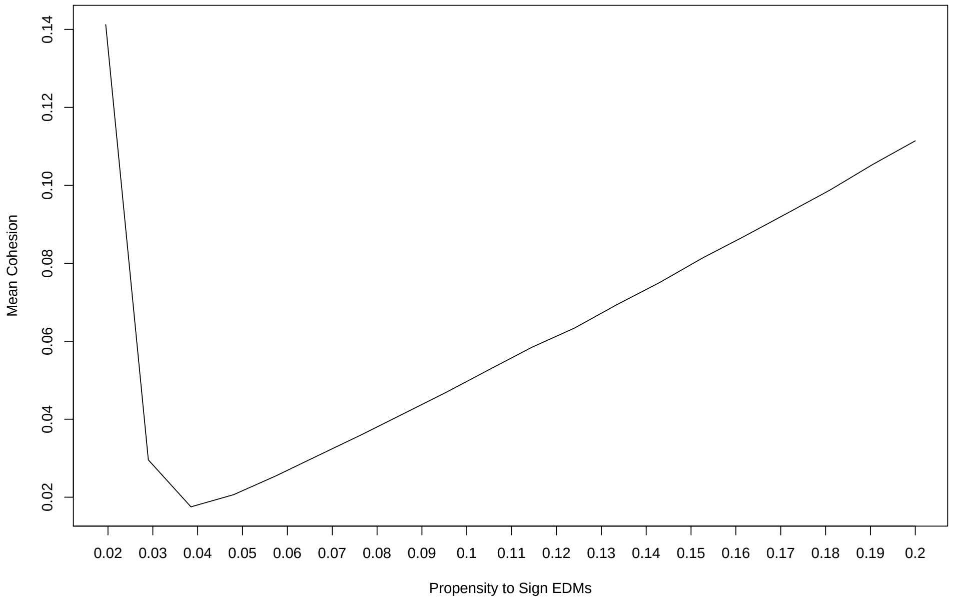


Figure 3.5: Cohesion measure of simulated incidence matrix for 100MPs with changing propensity to sign.

before slowly increasing. An example of this for the simulated study with 100 MPs is shown in figure 3.5. We leave further investigations of the relationship between propensity to sign and cohesion measure as future work.

This small study indicates that the propensity to sign, rather than the number of MPs within a party which effects the cohesion measure. Our simulation did not, however, take into account the way EDMs attract signatures: most receive few signatures and only some attract large support. We thus create a simulated cohesion measure to be able to compare the observed cohesion with a ‘random’ cohesion, which takes into account propensity to sign, but also the popularity of certain EDMs. We start with a simple Bernoulli model, similar to that in the above simulation study to motivate the need for further complexity.

3.2.2 Simulated Cohesion Levels

Cohesion levels are derived from the opinions of MPs and are not signed at random. To attempt to calibrate our measures we wish to simulate the signing of EDMs to replicate our data set but with the addition of the MPs signing EDMs at random. We simulate these incidence matrices (as in table 3.1) by developing simple stochastic models of the rate at which EDMs are signed by MPs. We can then calculate coherence on these simulated matrices using our cohesion measure.

Our simulated models must take into account that each of the political parties have a different propensity to sign EDMs. To more accurately model the way in which each EDM attract signatures, we also wish to factor in that different EDMs receive differing levels of support. We next define our model, starting with the most simple Bernoulli model and then factoring in the above properties which we wish to account for.

Let m_{ij}, represent whether MP i signs EDM j as defined in (3.1.2) and recalled above in (3.2.6). We estimate the value of p from our data, for each political party.

Define p, the propensity to sign, as Number of Signatures / (Number of MPs × Number of EDMs) — the 'mean' of the incidence matrix. We calculate p from the 2005/06 data set as 0.05 for both Labour and Liberal Democrat, and 0.03 for Conservative (similar to the proportions calculated for the 1997–2000 data set by Nason (2001)).

Table 3.3 shows the simulated cohesion levels for each party using this model with a time window of 100 EDMs. When compared with true cohesion levels, it is clear that this model produces cohesion levels lower than observed. Moreover, from experimentation, the resulting cohesion level from the simulated model does not depend that greatly on the size of the time window or the number of MPs within the party being modelled. The model also assumes that each EDM is identical in the level of support it attracts from MPs. This can be shown to be untrue, simply from observing the range and distribution of the total number of signatures per EDM over the session.

Signature-based model. We fit a new model which uses bootstrap resampling (Davison & Hinkley, 1997) which will more accurately model how many signatures each EDM receives but is tailored around each political party. The model, based on a time window of 100 EDMs, works as follows:

Let Y_g be the number of signatures EDM g receives from MPs in a particular party, for $g = 1, \ldots, N_{EDM}$ and N_{EDM} the number of EDMs in the data set. Take a random sample, with replacement, of size 100 from the Y_g to represent our time window of EDMs. Denote the elements of this sample by X_j, with $j = 1, \ldots, 100$.

We want the X_j to influence the level of signatures our simulated EDMs receive. For each X_j, define a corresponding probability p_j by:

$$p_j = \frac{X_j}{N_{MP}}, \qquad \text{for } j = 1, \ldots, 100 \text{ EDMS},$$

where N_{MP} is the number of MPs in our data set. Thus $p_j \in [0, 1]$, with the extreme values being met when none or all of the party sign EDM j.

Define m_{ij} as in equation (3.1.2), and model m_{ij} by:

$$m_{ij} \sim Bernoulli(p_j),$$

to obtain a simulated data matrix.

Calculate the (unweighted) cohesion of the simulated data matrix using equation (3.1.5) and denote by C_B, for $B = 1, 2, \ldots, n$. Repeating the entire process, we then define the simulated cohesion level by:

$$C_{SIM} = \frac{1}{n} \sum_{B=1}^{n} C_B.$$

For results shown here, we used $n = 10,000$.

Table 3.3 shows the cohesion levels calculated using this ‘signature based’ model. The cohesion values are much more realistic compared to the simpler model, accounting for features within the data more accurately and therefore suitable as a ‘calibration’ level.

	Party		
Method	Labour	Conservative	Liberal Democrats
Simple Bernoulli model	0.05	0.05	0.03
‘Signature’ based model	0.10	0.11	0.17

Table 3.3: Simulated cohesion levels for time window of 100 EDMs

A wide variety of other models could be used in the calibration of cohesion levels. A simplistic model would be to use the mean of the cohesion calculated over the session. This however would give a calibration level closely related to the actual cohesion and without the element of ‘random signing’ of EDMs by MPs.

The ‘signature’ based model used repeated simulations to get an overall cohesion level for the entire session. It is feasible that during different periods, for example before and after a parliamentary recess, the cohesion level is affected. A model could be used which sampled EDM signature levels by recent trends or by analysis of previous sessions. Development of these time-varying cohesion calibrations are left as future work.

3.2.3 05/06 Session Interpretation

Figure 3.3 shows the cohesion plots for the three main parties for the 2005/06 session, which started after the general election on May 5. The plots use a bandwidth of 100 EDMs and a step size of one. The three parties are shown along with horizontal lines representing the 'signature' based cohesion calibration levels. We use this calibration level as an indication of whether the party has more or less cohesion than if they were signing EDMs at random (but taking into account number of signatures varying over EDMs).

The plots show changing cohesion within the parties over time. Overall, the cohesion of the Liberal Democrat party is far more variable than Labour or Conservative, intuitively due to the far fewer MPs in the party making the cohesion level more sensitive to differences in opinion. As we do not know the exact relationship between party size and cohesion for all the parties, we only compare cohesion between parties in terms of *trend* rather than actual cohesion level and also take care to ensure that any fluctuations in cohesion due to EDMs receiving very few signatures is reported. Finally, we compare cohesion to that of the simulated level.

All three parties at the start of the session show a level of similarity in trend, increasing and decreasing cohesion at similar times. This trend stops shortly after the summer break (vertical line), and by the end of October all three parties appear to behave independently

Conservatives: Comparing cohesion levels to that of the calibration level, the Conservatives, unlike the other parties, are mostly above what would be expected if signing randomly (or rather with our 'signature' based model). The middle plot of Figure 3.3 shows that in late November the cohesion of the Conservative party appears to slump and then suddenly rises. Following the general election in May that year, Michael Howard resigned as party leader but did not step down immediately. The leadership campaign lasted all of November and continued into December, with David Cameron winning the leadership election on the 6th of December (see White (2005)).

A further low in the cohesion of the Conservatives can be seen during late March. This was an eventful period in politics. The new leader was starting to show the direction the party was heading in and a controversial education bill was narrowly passed with Conservative support.

A further significant period of the cohesion plot where the cohesion of the Conservatives is less than our ‘signature’ based model is during July 2005 and corresponds to a period in which there is a local minimum in the number of EDMs the Conservatives were signing. We believe this to be the reason although we have not found any other apparent link between cohesion and number of signatures at other periods during the session. We discuss this variation in the number of signatures briefly in the next section.

Liberal Democrats: The lower cohesion plot in Figure 3.3 also reflects interesting variability within the Liberal Democrat party during this session. From February onwards it remains at a low level compared to earlier in the session. Following what many considered to be a disappointing General Election result, and despite gaining seats, the leadership came under a lot of pressure. Activists felt the party had not taken advantage of a weakening government and opposition and criticised the leader, Charles Kennedy, for his policies and election campaign. It was also known within the party that he was battling with alcoholism (see Hurst (2006, Chapters 1 and 10)). After a period of intense pressure by high profile party members, Charles Kennedy admitted having a problem with alcohol and resigned as party leader on 9th of January (see Hurst (2006, page 23)). Comparing to simulated cohesion levels, the Liberal Democrats fluctuated above and below and more recently remained less cohesive than if EDMs were signed at random. Following the leadership election, Sir Menzies Cambell’s Liberal Democrats did not achieve the cohesion levels seen during 2005.

Labour: The Labour party generally has a lower cohesion than the other main parties. Following the summer break (indicated by a vertical line), there appears to be no dramatic change in Labour’s cohesion level, which is at a level suggesting that they regularly disagree with other members of the party. One period of interest is that of March 2006. As with the Conservatives at this time, the cohesion of the party dropped. The education bill which was passed during this time split the Labour party, with mass rebellion from the Labour backbenches, see Wintour (2006). Further problems for both Labour and the Conservatives at this time was the news of secret loans that both major parties had received in the run up to the general election, see Hencke (2006). These continuing revelations about party funding were a blow to both major parties.

The 2005/06 session is unique in that the cohesion of all three main parties are at a comparable level. Previous sessions have exhibited vast differences in the level of cohesion. Possible explanations of this are the number of MPs signing EDMs or changes in propensity to sign EDMs. Previous session plots from the 1997/98 session to date can be found online at `www.maths.bris.ac.uk/~db0797/Research.html`.

3.2.4 Volatility of Cohesion

When considering the cohesion plots with bandwidths of 100 EDMs or less, cohesion levels tend to be much greater for the more recently tabled EDMs. As the EDMs in the time windows become older, the coherence generally reduces. Some EDMs will receive all of their signatures quickly, whereas others may take longer to obtain support. Some possible reasons for this are discussed by Franklin & Tappin (1977) who consider EDMs to either obtain many signatures quickly, due to 'business-like' canvassing of MPs by the proposer of the motion, or to gain support in a more random fashion.

If we were considering a dynamic cohesion measure, which was updated on, say, a daily basis, the addition of signatures to EDMs causes the coherence of a given time window to change. Further, it is of interest to see how the measure M_i, defined in (3.1.4), varies within a given time window (recall that M_i is the mean dissimilarity that MP i has with all other MPs, and is used to define the cohesion measure). The variation of the M_i during each time window is a measure of the range of similarities that the MPs have towards each other. To measure this volatility we define the variation measure $V = \mathrm{var}(M_i)$ for MPs $i = 1, \ldots, n$.

Similar to the cohesion plots, we look at the measure V_i over a moving time window of 100 EDMs with fixed step size of one. Figure 3.6 shows the plot for the 2005/06 session. We observe that the variance of the M_i for all three parties remain low and fairly stable up to mid November (indeed, the Liberal Democrats and Labour have a low and stable variation of M_i throughout). This is expected. With no major changes in the signing patterns by MPs, the spread of the M_i would reasonably be expected to remain fairly constant given

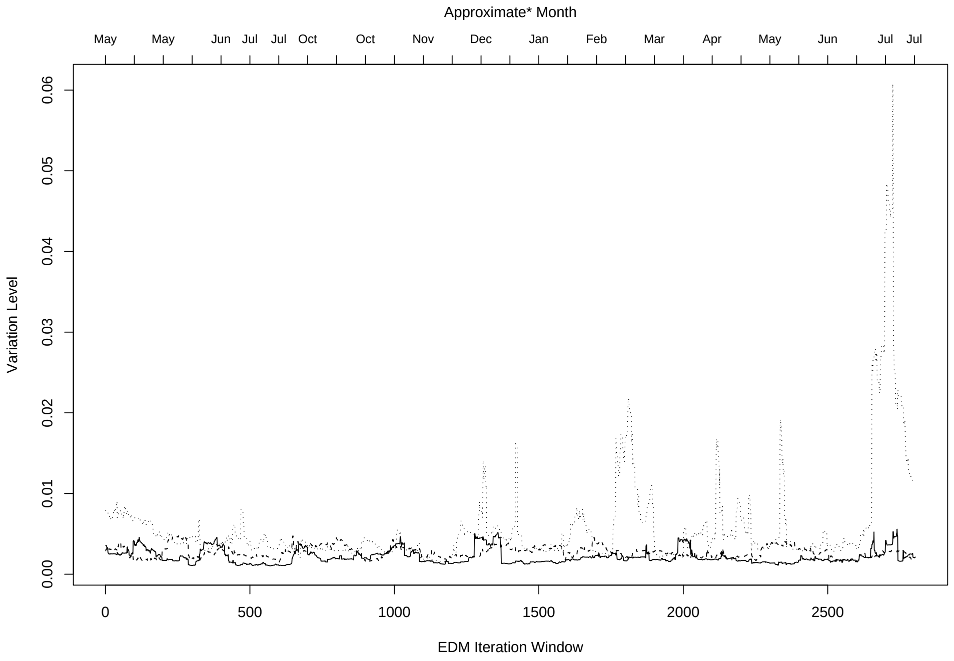


Figure 3.6: Variation of mean dissimilarity of MPs with a moving time window during 2005/06 session: Solid = Labour; Dotted = Conservative; Dashed = Liberal Democrat.

that the EDMs have obtained all, or most of their signatures. For very recent time windows the variation of the Labour party, and more so the Conservatives are a lot higher, indicating the large range of similarities each MP has with members of their own party. This is due to the more recent EDMs having yet to receive all of their signatures. As EDMs receive signatures, they have a changing influence in the calculations of the M_i and cause a larger range in the values of M_i. This range in values of M_i becomes more constant (and smaller) in the long run when the EDMs no longer attract additional support from more MPs.

Two large changes in this pattern occur during this session for the Conservative party, the first during December and January, following David Cameron's election to leader of the party. This plot shows the variability of similarity amongst MPs during this time, indicating the differences within the party. Considering that the cohesion at this point is also rising, there is 'unrest' within the party with some MPs within the party in agreement with each other and others who are not. To a certain but lesser extent Labour also show unrest during this period, which highlights the impact a new leader can have on the other parties.

The second major disruption comes more recently, during February and March 2006.

This coincides with a period of low signing of EDMs by the Conservatives and this may be the explanation rather than any political events. Further increases in the variance are towards the beginning of April and are at the same time as a drop in cohesion (and the party loan scandal), see Hencke (2006). More recently there is a change in variation in July 2006. This variation may settle over time, with the relationship being temporary, as not all signatures will have been received for those recent EDMs.

3.3 Feature Selection using Cohesion Levels

We next propose to use our cohesion measure in a technique which utilises computational power and a combination of statistical techniques. We first consider the work by Poole & Rosenthal (1997) and Poole (2005) who developed the non-parametric optimal classification (OC) method for the spatial modelling of legislators in parliament and parametric methods based on the NOMINATE procedure (*NOMINA*l *T*hree-Step *E*stimation) which were described in Section 2.4.

We refer the reader back to the scaling solution of EDMs in Figure 3.1. Recall that the OC and NOMINATE techniques use an initial set of spatial coordinates of legislators (such as the scaling solution) and finds cutting planes which best separate the legislators between those who vote Yea or Nay on a given roll call. The method iterates between finding cutting planes, spatial coordinates, and in the case of NOMINATE, parameters for a legislators *utility* function. The results are used to see where the legislators lie on a given roll call and to investigate cohesion and discipline within the legislators.

The OC (and NOMINATE) procedures are not easily adaptable to the analysis of EDMs. The methods rely on either a Yea or Nay vote in order to *cut* the legislators into two groups and, as previously discussed it would be wrong to construe a lack of signing EDM as a ‘Nay’ vote. A possible method could be to investigate EDMs with opposite opinions, or to use amendments with conflicting views of the original EDM to obtain a data set where MPs vote for either sides of an opinion. This would create a much smaller subset of MPs where a firm Yes/No could be worked out for the given subject matter. Not only would the data set be very small, there would still be difficulty regarding the subjective nature of EDMs.

Spirling & McLean (2006) attempted to use the OC procedure on divisions in the House of Commons and concluded that the method was unsuitable for the data due to the divisions being whipped and strategic voting by MPs. Even if the methods outlined here were adapted for use with EDMs, their cheap-talk nature would again be an issue when ‘cutting’ MPs on a single motion.

We instead look to other methods to further investigate EDMs by turning our attention to using feature selection to pick out important information from the data.

Feature selection is a valuable tool used in many applications, reducing dimensionality and allowing for easier subsequent analysis and interpretation of results. Our cohesion measure is dependent on which EDMs MPs sign. Some of these EDMs are likely to influence the cohesion measure more than others. They may have a higher popularity amongst MPs or they may be on issues which regularly divide MP opinion. In contrast to narrowly searching individual topics for a reaction amongst the legislators, we wish to investigate whether, over the vast data set, there are types of EDM that MPs agree on, or EDMs which cause disagreement within a political party. This has a further advantage over the OC and NOMINATE procedures in that we use all EDMs on a given topic, not just one of many to use as a cutting plane.

We first classify EDMs from the 2001–2005 parliament into different issues or types. By using average weighted dissimilarity measures between MPs, we can allow certain types of EDM to have a greater or lesser effect on the cohesion of the MPs. Minimising and maximising the cohesion measure by adjusting these weights, we discover the cause of most agreement and disagreement within the main political parties during the last parliament.

3.3.1 Optimisation using Weighted Cohesion

EDMs are used by MPs to give an opinion on any subject matter. There are a huge range of topics which EDMs can cover and we categorise EDMs into a number of different types or issues. EDM topics can range from education to the environment; they can give an opinion against a political party or their policy, such as criticising public spending, or be such that all members may wish to sign.

For each session in the 2001—05 parliament, each EDM has been categorised into one

	Session				
Type	01/02	02/03	03/04	04/05	All
Health	212	246	233	101	792
Trade/Business	87	84	163	81	415
Social Issues	143	133	83	37	396
Abuse/Humanitarian	87	136	**55**	38	316
Congratulatory	**54**	90	87	44	275
Foreign Issues	**50**	98	93	**33**	274
Policy/Legislation	70	81	82	**31**	264
Transport Issues	**63**	88	66	43	260
Environmental	**20**	82	82	51	235
Sport	75	**55**	65	**29**	224
Education	**51**	**61**	**62**	46	**220**
Media	75	**55**	**57**	**29**	**216**
Arts/Culture/History	**50**	**23**	81	**18**	**172**
Employment	64	**49**	**49**	**32**	**194**
Defence/Armed Forces	65	**34**	**47**	45	**191**
Congratulations in Sport	70	**29**	65	**20**	184
Iraq	**31**	93	**42**	**15**	**181**
Food/Agriculture/Farming	**48**	**61**	**51**	50	**210**

Table 3.4: Number of EDMs on different issues. EDMs are listed in descending order of total number over all sessions. Plain type means that number for that issue in a particular session is in top 10, bold means that it is in top 10 for another session.

of 50 different types. Table 3.4 shows the most popular 10 types for each session (along with other types referring to top 10 from other sessions). Care was taken to classify an EDM into its primary subject matter. When more than one category was possible all were recorded; but only one has been used per EDM for this analysis (see Appendix A.2 for details of our classification process).

We are interested in which EDMs influence cohesion. We consider the EDMs which form the 30 most popular types of EDM and give each type a weight so as to regulate their influence or importance. For a set of weights, the weighted cohesion can be calculated using equations (3.1.4)-(3.1.5). We wish to adjust the weights so as to maximise the cohesion measure. The corresponding weights will identify features (or types of EDM) which bring members of the parties *closer* together. Minimising the weighted cohesion will pick out features which cause most disagreement within the parties.

Many classical numerical optimisation routines assume smoothness of the optimisation function and depend heavily on starting values. Due to the nature of our optimisation criteria and size of the problem, genetic algorithms have been selected for this optimisation task.

A good introduction to genetic algorithms can be found in Mitchell (1998). Briefly, genetic optimisation uses the principle of natural selection to obtain (locally) optimal values. For a starting population of weights, each set of weights is considered in binary form along with their corresponding weighted cohesion level. Sets of weights which are found to be more optimal progress to the second generation, and combine to produce new sets of offspring (or weights) also in this generation. Analogous to survival of the fittest, this process aims to produce a more optimal second generation of weights. Picking a suitable initial population size allows for a desired level of 'genetic diversity' and the number of iterations (or generations) will control the level of convergence of the optimal value (with an obvious trade-off between the two).

Genetic optimisation was used for feature selection by maximising and minimising coherence levels as described below. Initial population sizes of 1000 were used over 500 generations to allow values to converge suitably. Other variables used within the procedure is the mutation chance, which is the probability that a given entry in the binary number will change. This was set at 0.001. Finally, for each iteration, 200 of the population were allowed to proceed to the next generation.

We note that the ordering of the weights within the procedure may affect convergence, with weights ordered close together being more likely to have either combined, or not combined with another set of weights. So, for example, with 30 different weights, weights 1 and 2 are more likely to have the same outcome (combine with another set of weights, or not) compared to weights 1 and 30. Repeated optimisation routines with different ordering would give more information as to the severity and effect of this, but is left as future work.

Finally, we do not rule out other methods of optimisation procedures to solve these problems. Methods based on estimating the gradient of the output function were tried (for example, those by Nelder & Mead (1965) and Fletcher & Powell (1963), but results were slow and often did not converge.

3.3.2 2001–2005 Parliament Results

Table 3.5 shows the results of feature selection for individual sessions during the 2001–05 parliament. Cohesion values were maximised to select features which make parties more cohesive and minimised to find issues which separate the parties. When the optimising criteria is maximised, we find that the results mostly contain only a single feature (EDM type) of the data. This is somewhat expected, with the solution putting maximal weight on the attribute which causes maximal cohesion and minimal weight on all others. Any other combination of weights would generally produce less optimal results. This is treated more formally, and more generally (considering subsets of attributes) in Friedman & Meulman (2004) (we leave any adaptation of our methods to subsets of attributes as future work). When minimising cohesion, results would not be expected on just a single attribute.

Given that we would reasonably expect single issues to be given maximal weight in our procedures, we can compare the resulting (weighted) cohesion level with the cohesion calculated from considering each of the different 30 types of EDMs individually. When more than one issue is found to give maximal importance, this will act as a check that the results are, indeed, more optimal.

For all sessions and parties except for the Conservatives during 2003/04, maximum cohesion is a result of a single issue. Comparing this to the cohesion calculated when all other issues individually are given maximal weighting, confirms that the issue found using our optimisation techniques is the most pertinent. In many cases, other issues would cause near optimal results and this is discussed in the next section. Maximal disagreement was caused by combinations of different issues. In all cases, the level of cohesion caused by a combination of the issues was less than that from any single issue.

Cohesion: During the 2001/02 session, foreign issues bring all three major parties together. This was a very eventful parliamentary year, covering the attacks on the World Trade Centers in New York (The Poynter Institute, 2001) and subsequent invasion of Afganistan (Wintour *et al.*, 2001). These occurrences are a possible reason for the solidarity within the parties.

Energy issues maximise the Liberal Democrat cohesion for all sessions except for 2001/02

01/02 Session					
Labour		Conservative		Lib Dems	
Coh.(0.72)	Sep.(0.95)	Coh.(0.75)	Sep.(0.96)	Coh.(0.63)	Sep.(0.91)
Foreign	Europe Environmental Statutory Instruments Congrats. in Sport (6.1)	Foreign	Disability Issues Middle East Environment Employment Congratulatory(2.3)	Foreign	Congrats. in sport Europe
02/03 Session					
Labour		Conservative		Lib Dems	
Coh. (0.84)	Sep. (0.95)	Coh.(0.79)	Sep.(0.96)	Coh.(0.67)	Sep.(0.92)
Energy Issues	Northern Ireland Defence/Armed Forces Immigration Statutory Instruments Foreign	Social Issues	Northern Ireland Congrats. in Sport Transport Safety Employment Immigration	Energy Issues	Northern Ireland Sport Congrats. in Sport Arts/culture
03/04 Session					
Labour		Conservative		Lib Dems	
Coh. (0.83)	Sep. (0.96)	Coh. (0.78)	Sep. (0.94)	Coh.(0.66)	Sep.(0.91)
Arts/Culture/History	Northern Ireland Congrats in Sport Religious Issues (0.5)	Tax Policy/Legislation	Middle East Abuse/Humanitarian Disability Issues (1.1)	Energy Issues	Northern Ireland Sport Europe
04/05 Session					
Labour		Conservative		Lib Dems	
Coh.(0.82)	Sep.(0.97)	Coh.(0.76)	Sep.(0.94)	Coh.(0.65)	Sep.(0.93)
Abuse/Humanitarian	Northern Ireland Arts/Culture/History Prisons	Foreign	Middle East Transport Safety Disability Issues Congratulatory (1.3) Northern Ireland Employment (0.1)	Energy Issues	Northern Ireland Iraq Local Govt.
All Session					
Labour		Conservative		Lib Dems	
Coh.(0.84)	Sep.(0.94)	Coh.(0.79)	Sep.(0.93)	Coh.(0.67)	Sep.(0.91)
Energy Issues Pensions Sport	Disability Issues Congrats. in Sport Policy/Legislation (5.2) Education (3.0)	Policy/Legislation	Disability Issues Defence/Armed Forces	Energy Issues	Northern Ireland Sport Congrats. in Sport Regional Issues

Table 3.5: Optimal issues for cohesion and separation of parties 2001/05 parliament, with cohesion value. (Issues have maximal weighting (10) unless stated and each line corresponds to a different issue.)

(where, in fact, the cohesion is just 0.01 less than for foreign issues alone). This topic, covering all aspects of renewable and non-renewable energy sources, was deemed important by political parties during this period. Labour published The Energy Review (Performance and Innovation Unit, 2002) and the Liberal Democrats had clear policies on such issues, with targets of reducing climate change emissions depending hugely on renewable energy expansion, see Liberal Democrats (2001). They have been outspoken over the issues of energy, in particular regarding renewable sources. In an interview on the subject the Liberal Democrat shadow Environment Secretary Norman Baker stated that “proper investment in renewables together with energy conservation and efficiency measures would eliminate the need to rely on nuclear power to meet Britain’s greenhouse gas commitments”, see Baker (2004).

Separation: Interpreting the results from the minimisation of cohesion to find separation is more complicated than the maximisation. There could be an underlying structure in the way certain MPs sign EDMs which are in favour or oppose a given subject. There could also be a pattern in how MPs consistently sign no EDMs of a certain type. We leave the details of this to further work and here present an overview of results and *possible* interpretations.

Congratulatory and sporting EDMs are often selected amongst issues which cause maximal separation within the parties. They are generally considered to be ‘all-party’ EDMs and obtain signatures from MPs regardless of political affiliation. Many MPs may choose not to sign these EDMs at all, whereas others will consistently sign them, causing them to be a source of separation within the parties.

Northern Ireland issues (by which we mean those relating to the peace process and its implications for devolution) are fairly common in separating all the parties during different sessions. The period of interest was an eventful time in the Northern Ireland Assembly, with the first minister, David Trimble, resigning and later returning to power only to resign for a second time. Following Britain resuming direct rule in 2002, a bank raid and ‘brutal murder’ blamed on the IRA delayed any progress on restoring the Assembly, see The Economist (2006). These events appear to have continually divided the opinion of MPs in Westminister.

Our results also show that disability issues as well as the Middle East separate the Conservative Party in all but one of the sessions. The latter is potentially caused by MPs criticising government policy regarding the Middle East (Watt, 2003) during a time of hostility and uncertainty following the invasion of both Afganistan and Iraq (Conte, 2005) and the death of the Palastinian leader Yassar Arafat in November 2004 (Whitaker, 2004). It is plausible that the separation is caused by a small number of MPs agreeing on certain EDMs, whilst all others show little or no opinion.

The Conservatives were the only main political party to have a manifesto during the 2001 General Election which had pledges aimed at the needs of the disabled, see Conservative Party (2001). Our results indicate that over the following years the party was divided on the issue. In 2004 the Conservatives started a nationwide consultation process in disability legislation, see Conservative Research Department (2004), possibly as a result of divided opinion within the party.

Other interesting (and possibly expected) results include immigration issues dividing Labour and Conservative in 2002/03 when asylum procedures came under heavy scrutiny, see Robinson (2003).

Entire Parliament: Table 3.5 also shows results from when the entire 2001—05 parliament was considered as one data set. Only MPs who were present over the entire period were considered for the analysis. The results show which issues are consistently causing cohesion and separation over the entire parliament. As expected given results from individual sessions, Energy Issues and Northern Ireland cause the Liberal Democrats to be more and less cohesive respectively. Sport is the only issue to appear to cause opposite reactions, being a minor cause of separation for Labour during 2001/02 yet overall causing separation over the entire parliament. Issues regarding policy and legislation separate the Labour party overall, whereas they make the Conservatives more cohesive. Possible causes of this would be EDMs which are against government policy being highly signed by Conservatives and also highlighting rifts within the Labour camp.

3.4 Conclusions and Future Work

In this chapter we have proposed a new cohesion measure to analyse the behaviour of Members of Parliament in the UK. Our measure uses the signatures of Early Day Motions, a large and rich data source.

We apply our measure to a moving time window of EDMs over the recent 2005–06 parliamentary session. Due to the complexity and uncertainty in the signing of EDMs, we constructed a simulated cohesion measure for each of the main political parties. We use these measures to assess whether a party's cohesion is high or low, and highlight changes in cohesion which can then be linked to political events. These comparisons showed evidence to suggest that the level of our cohesion measure is broadly indicative of perceived party unity.

We further use our cohesion measure to investigate which 'issues' cause the parties to unite and separate in opinion. This was achieved by classifying each EDM by primary topic and using a range of statistical techniques to 'weight' groups of EDMs on similar issues. Data from the entire 2001–05 parliament were used and pertinent issues, which caused maximal and minimal cohesion for each party during each session, were discovered. Comparing these results once again to political events over the sessions there is further evidence to suggest that the tabling and signing of EDMs reflects current political climates on both national and international issues.

We emphasise that our analysis of results and comparisons to political events are to some extent conjectural, and our analysis is not exhaustive. We have indicated possible reasons for some of the behaviour shown within EDMs where we believe there is an interpretation which gives a new insight or interest into EDMs and MPs.

We also highlight the subjective nature of classification of EDMs and the difficulty of assigning each to just one category. Careful steps were taken to ensure a 'good' classification (described in Appendix A.2), and for the vast majority of EDMs, a primary category for each was agreed upon by all coders.

This article shows how EDMs can be used to measure political cohesion in the UK. Without the influence of the party whips, there are many extensions to their use. A natural extension to the moving cohesion measure would be one of forecasting behaviour of the cohesion of the political parties. Identification of those MPs who are the main cause of the lack of unity within a party could also give an indication of the structure in the party. Consideration of the hierarchy of MPs from a centroid point of a party is a possible way of carrying out such a task.

As discussed, single issues were the main cause of the parties showing maximal cohesion. The second most important issue could also be of interest and could be obtained by introducing an entropy measure to our optimisation criteria. Furthermore, by classifying EDMs into more than one issue, a more detailed and accurate picture of the content of each EDM could be used for feature selection.

Chapter 4

Divisive Issues Between Political Parties

4.1 Introduction

In Chapter 3 we defined a cohesion measure for political parties in the UK. EDMs were classified by *issue* and an optimisation procedure was defined which used *weights* assigned to each issue to maximise and minimise the cohesion measure. That is, we identified the issues on which MPs within a party agreed on, and issues which caused disagreement.

In this chapter we extend the idea of using weighted dissimilarity matrices along with feature selection to pick out attributes of the EDM data set which are of interest. Previously, each of the main political parties was considered separately, with no interaction between them. Here we develop an optimisation criterion to investigate the interaction *between* political parties and to identify issues which differentiate the main political parties, by causing them to be less similar to each other.

Briefly, we wish to pick out issues which cause political parties to differ in opinion. Our measure for this is the number of MPs who are erroneously classified as being in a particular political party. The smaller this number, the more disjoint the parties are. By altering the influence that EDMs of a particular *type* have on the classification procedure, we can identify the issues which cause the most separation.

The OC and NOMINATE techniques, by Poole (2005), were described in Section 2.4

and use a scaling solution of legislators based on parliamentary voting. These techniques classify legislators by finding 'cutting lines' or planes which best separate different parties. Both the OC and NOMINATE methods induce changes in the scaling solution by assuming errors in the voting pattern. We considered the use of these procedures for EDM data in Section 3.3 and pointed out why they are not suitable for such analysis (asymmetric signing and strategic voting were amongst the reasons).

We next consider the scaling solution of MPs given EDM signing, as well as a method of classification given this scaling solution. We use data from the 2001/02 parliamentary session to give an example of these procedures before incorporating them into an optimisation routine to identify key issues.

4.2 Scaling Solution of EDM Data

The dissimilarity matrix of MPs can be represented spatially using classical multidimensional scaling (MDS) to reduce dimensionality of the of the data (see Chatfield & Collins (1996) for details). We use classical scaling, rather than other types of scaling for consistency, simplicity and ease of computation, as we will compute many multidimensional scalings. Further, good results using classical scaling have been previously shown using the EDM data set (see Nason (2001)). We do not investigate the use of the many other scaling methods here. References to scaling solutions in the remainder of this chapter refer to classical multidimensional scaling. The scaling solution plays an important role in our methodology as we use it to categorise MPs into political parties.

We use the EDM data for the 2001/02 session to aid the explanation of our methodology. We use the unweighted Jaccard coefficient to calculate the dissimilarity between MPs (defined in Chatfield & Collins (1996) and given in general form in Section 3.1.3). Figure 4.1 shows the first two dimensions (or principal components) of the scaling solution for the three main parties during the 2001/02 parliamentary session. We see the strong party structure which exists within the signing of EDMs. The Conservatives, although fairly spread out are clearly separated from the other parties. The Liberal Democrats are strongly grouped but appear attached to the edge of the Labour party. Further investigations show that for

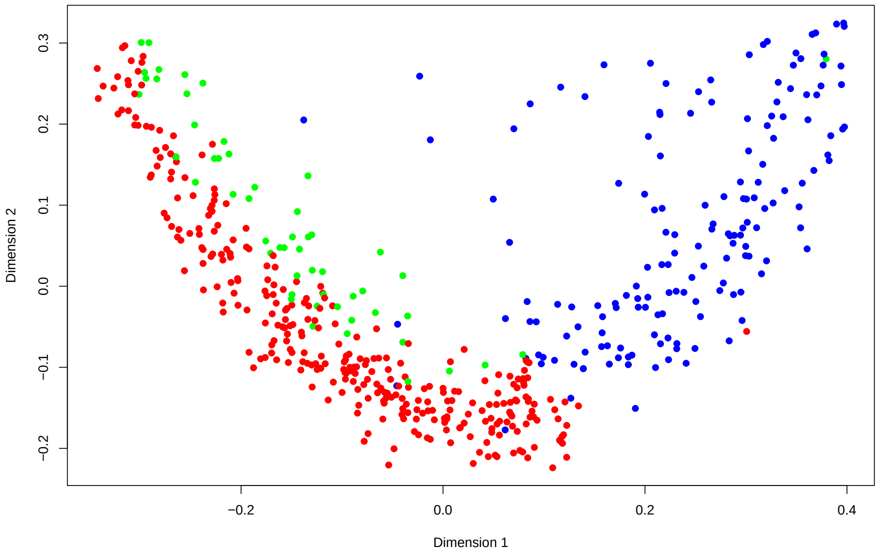
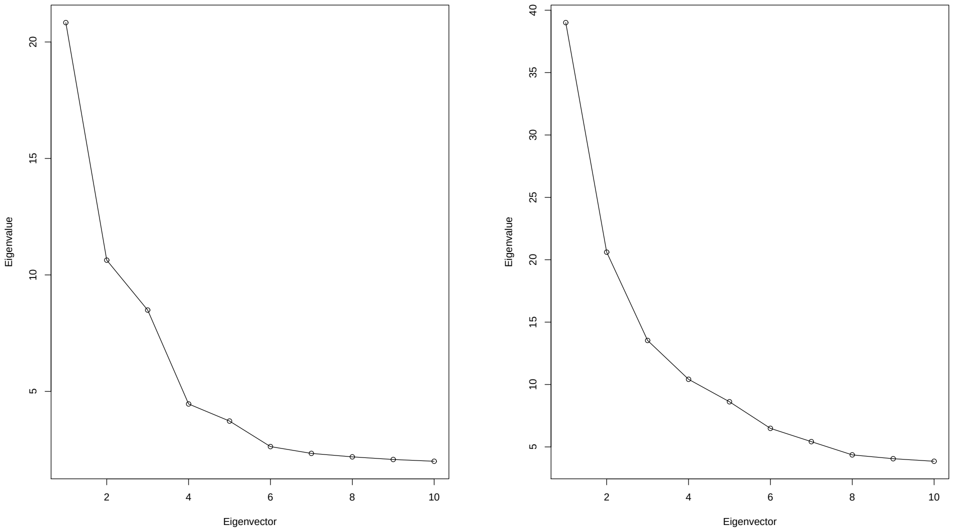


Figure 4.1: Scaling solution for 2001/02 data. ● = Labour, ● = Conservative, ● = Liberal Democrat.

higher dimensions, the Liberal Democrats are more disjoint from Labour. This raises the question of how many dimensions one should use to represent the data in the scaling solution. Common practice is to look at the *scree* plot of the eigenvalues of each principal component of the scaling solution, and to use the number of dimensions indicated by the *elbow* point on the plot. Figure 4.2(a) shows the scree plot of the first 10 eigenvalues (out of a possible 529) from the scaling solution of the 2001/02 data. The elbow point seems to be at the fourth eigenvalue, so we use up to and including this in the subsequent analysis.

We note at this point the 'horseshoe' effect (see Kendall (1971)) which is apparent in the scaling solution in figure 4.1. This phenomenon occurs in classical multidimensional scaling (as well as other linear scaling methods and non-metric multidimensional scaling) and is observed for large datasets where there are many dissimilarity coefficients between pairs of observations which are close to maximum. At a certain point, it is no longer possible to plot the observations any further from each other and this causes extreme points to be positioned in a curve. It implies that the second (or subsequent) axes may be dependent upon the first axis (although they are linearly uncorrelated).

Podani & Milklos (2002) observe the horseshoe effect using a variety of dissimilarity



(a) (b)

Figure 4.2: Scree plot of first 10 eigenvalues. (a) 2001/02 Session, (b) First 200 EDMs from 2001/02 Session.

coefficients and classical scaling. The Jaccard coefficient is not one which is considered within this study, but it has been observed in other studies (in particular, see Hoiland *et al.* (2004)). We refer the reader to the list of references provided by Podani & Milklos (2002) and to the work by Hill & Gauch (1980) for many methods of accounting for horseshoe effects.

We will not incorporate such techniques within this study. For the data presented in this chapter we note that although the horseshoe effect is observed for the second dimension, it is not apparent in higher dimensions. Further, although we use data from the entire session for motivation, our actual analysis will be performed on a much smaller data set. We direct the reader ahead to figure 4.3(a) for the first two dimensions of the scaling solution for just 200 EDM; an example of the data which will be used in within this chapter. We note that the horseshoe effect is not as apparent over these dimensions, nor is it for higher dimensions. We thus make no effort to adjust for any possible ‘horseshoeing’ at this point, but note that it should be considered for any future analysis using data from the entire session.

4.2.1 Classifying MPs

We wish to use the scaling solution to predict party affiliation of each MP. To group the data, we use the simple method of linear discriminant analysis (LDA), proposed by Fisher (1936). This seeks to find a linear combination $\mathbf{a}^T\mathbf{x}$ of the n variables (or dimensions), $\mathbf{x} = (x_1, \ldots, x_n)$, that would best separate the data into groups (see Krzanowski & Marriott (1995) for a detailed overview). Note that this scoring technique has been chosen for its speed and simplicity and alternative methods, such as clustering procedures, could be used instead within this procedure. This method has been used (although with non-metric scaling) by Cox & Ferry (1993). Work by Chang (1983) and Gnanadesikan *et al.* (2007) are examples of studies which are are similar, but using principal component analysis with the aim of clustering, rather than classification.

We apply LDA to the first 4 dimensions of the scaling solution for the three main political parties. The LDA calculates discriminant values with which to best predict party affiliation of all MPs (given that there are three party choices). Given the *known* party affiliation of the MPs, we can then identify the number of erroneously classified MPs in each group. The number of misclassified MPs can be thought of as a measure of overlap between parties.

We return to the data from the 2001/02 parliamentary session and perform LDA on the first 4 dimensions of the scaling solution. We predict each MP's party given the LDA results. The number predicted for each party, along with the known party affiliation of the MPs is shown in Table 4.1. We see that the most erroneously classified MPs are Liberal Democrat as Labour (50 times) and Conservatives classified as Labour (21 times).

	Classification		
Actual Party	Con	Lab	Lib
Con	139	21	1
Lab	1	315	0
Lib	1	50	2

Table 4.1: LDA Classification for 2001/02 session.

With equal EDM weights, the total number of erroneously classified MPs is 74. This number will be used later as our optimisation criteria. Note that for the remaining analysis we concentrate on the total number of erroneously classified MPs and not on the behaviour of individual MPs or parties. Our justification of this is similar to the cheap-talk nature of EDMs, discussed in Section 3.1.2. We do not pick out (and name) MPs due to the ‘noise’ which may be present. We treat misclassification of an MP as a sign that they are far from their own party, rather than having beliefs and opinions belonging to another.

The example we have used so far is for an entire session. In practice, looking at an entire session is not computationally feasible and instead we look at a moving time window of 200 EDMs. This is the same idea as the moving cohesion measure in Chapter 3, but here the EDM time window is shifted by 100 EDMs each time.

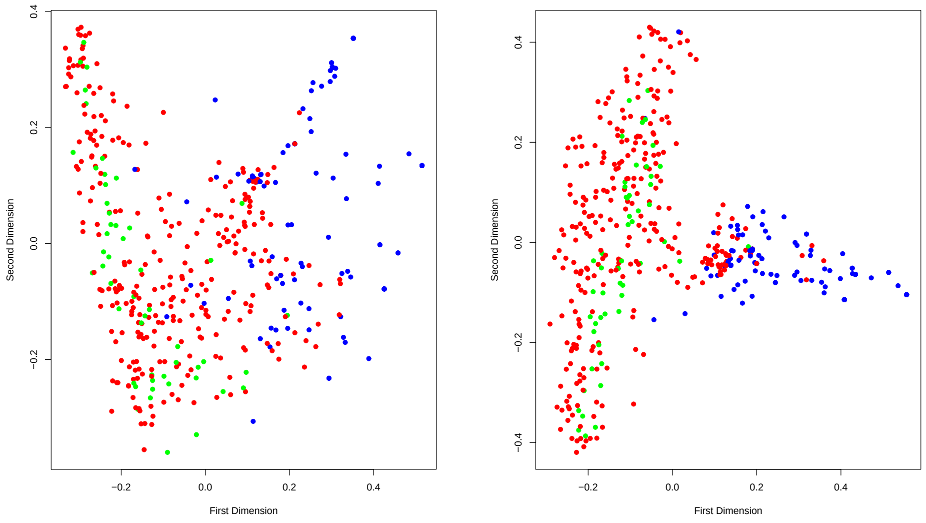
The scree plot of the first time window for the 2001/02 session is given in Figure 4.2(b). For this example, the scree plot does not exhibit such defining features as the full session and is more of a curve than a definitive *scree* shape. We continue to use 4 dimensions in the subsequent analysis of the data, based on the full data set. Furthermore, as will become clear from the methodology, it is not a priority to have a scaling solution which is near perfect as the measure of overlap between parties (misclassified MPs) will be used as our optimisation criteria.

4.3 Feature Selection using Linear Discriminant Analysis

We next describe our methodology for identifying key issues which cause separation between political parties. We introduce variables into the techniques previously described in the form of EDM weights. These enable us to alter the importance of each EDM type to minimise the number of erroneously classified MPs.

As with the cohesion measure of Section 3.1.3, we classify EDMs into different *types* to which we assign a weight w_k. We use this to calculate an average weighted dissimilarity D_{ij} between MPs (i, j) defined in equation (3.1.3) as

$$D_{ij} = \sum_{k=1}^{T} w_k d_{ijk}, \tag{4.3.1}$$



(a) (b)

Figure 4.3: First two dimensions of scaling solution on first time window of 2001/02 session. (a) Unweighted. (b) Weighted. 🔴 = Labour, 🔵 = Conservative, 🟢 = Liberal Democrat.

with

$$\{w_k \geq 0\} \text{ and } \sum_{k=1}^{T} w_k = 1,$$

for attributes (EDMs) $k : 1 \rightarrow T$. As with (3.1.3), d_{ijk} is the Jaccard dissimilarity measure between MPs (i, j) on EDM type k, with weight w_k.

We use the average weighted dissimilarity matrix to form a scaling solution, as described by example in Section 4.2. We then classify MPs using linear discriminant analysis (Section 4.2.1) and find the number of MPs who have been erroneously classified. We can then modify the weights assigned to each EDM *type* to attempt to reduce the number of misclassified MPs: the smaller the number, the larger the difference between the parties. The corresponding weight values of the optimal solution will identify the importance of given issues in separating the main political parties, and answer questions as to which issues distinguish the parties from each other.

We once again use genetic algorithms, described in Section 3.3.1, to minimise the number of erroneous MPs and find optimal EDM issue weights. We run the procedure with a population size of 500 over a total of 750 generations.

Returning to our example, Figure 4.3(a) shows the first two dimensions of the unweighted scaling solution for the first time window of 200 EDMs for the 2001/02 session. Note that with fewer EDMs, the structure does not appear as strong compared with the full data set in Figure 4.1. Figure 4.3(b) shows the first two dimensions of the scaling solution for the same time window but with the average weighted dissimilarity matrix using the optimal weights from our procedure. We see that on these dimensions the weighting has, to a certain extent, brought MPs of a similar party together. The Conservatives and Labour are more compact on these dimensions. Further, for higher dimensions, the Liberal Democrats are more disjoint from the other parties. Overall, 45 MPs are misclassified.

Preliminary tests using this method indicate that despite a large starting population, the results found appear only locally optimal – the genetic optimisation for repeat trials on the same data converged with *different* optimal weights. Instead of increasing the population size, we choose to repeat the process 10 times with new starting populations to get repeat results (this allows each 'run' to be computed in parallel). We then include these 10 optimal solutions in the starting population of a final procedure to get the results which we present in this chapter. For the example presented above, this method reduced the number of misclassified MPs to 36.

4.3.1 Application to 2001–05 Parliament

Figures 4.4–4.6 shows the results of the genetic optimisation moving over time for each session between 2001–04. We do not plot the results from the shorter 04/05 session, which only comprised of seven 'time' windows. The horizontal axis represents each time window of 200 EDMs with an overlap of 100 EDMs. It is labeled with the month corresponding to the middle EDM of that time window for reference (although the EDMs, and thus the months are not spread evenly over the session). The top axis gives the number of erroneously classified MPs for that window. The left vertical axis represents the 30 most popular issues. They are ordered by the sum of the weights of each issue, for the entire session. This sum of weights is given on the right axis. Each issue, for each time window, is plotted by a shaded box, which indicates the weight given to the issue. For clarity, we have multiplied each weight by 10 and have 4 levels of shading to represent increments of 2.5 along the new weight scale of 0 to 10 — the darker the shading, the higher the weight. This adds an extra

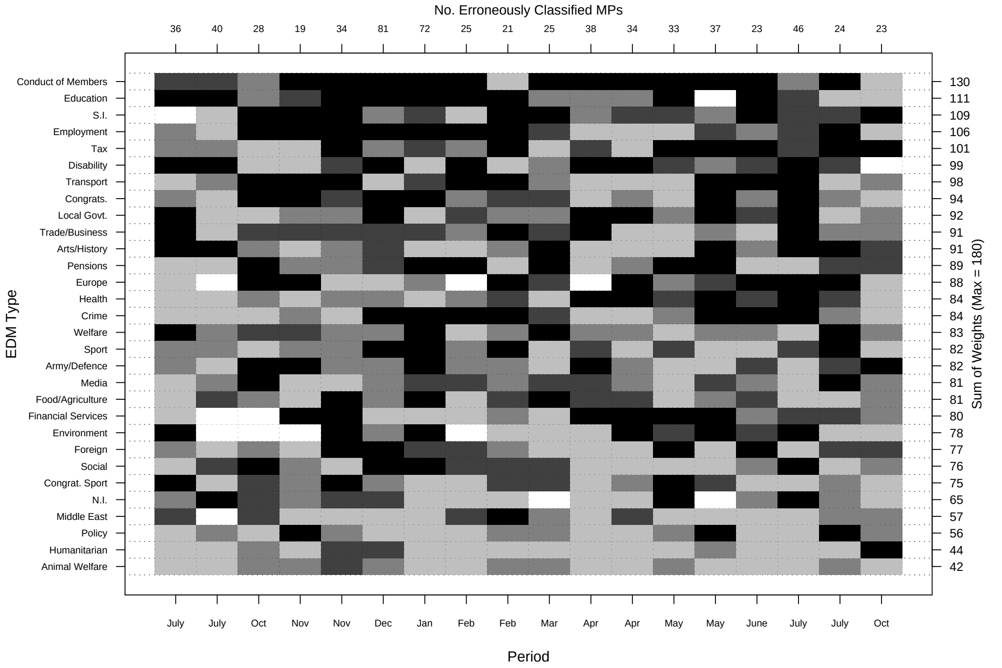


Figure 4.4: Issues separating main political parties 2001/02 session. Weight shading: white 0–2.5, light grey 2.5–5, dark grey 5–7.5, black 7.5–10.

dimension to the plot, with changing weights easily readable from the plot.

Although we present the results from each of the sessions, we believe they should be treated as *initial* results. They are limited due to uncertainty over the number of locally optimal solutions which exist and the likelihood of them being found using the current methods. The largest limiting factor was computational power. For the solutions reported in this chapter, it took up to 3 days to complete the genetic optimisation for a single time window (using computers with 2.2 GHz processors and 2Gb of RAM). Over our 10 repetitions of the procedure for each time window, the optimal solutions found differed considerably. This indicates the size and complexity of the problem and why the work presented here is treated as preliminary. We thus only summarise some of the findings and discuss future work in Section 4.4.

Results from the 2001/02 session in Figure 4.4 show some expected results. Conduct of Members, and Statutory Instruments (SI) are *within-party* EDMs and generally expected to get support from just one party. We see that they are amongst the most divisive of between-party issues. The highly divisive areas of education, employment and tax are also found

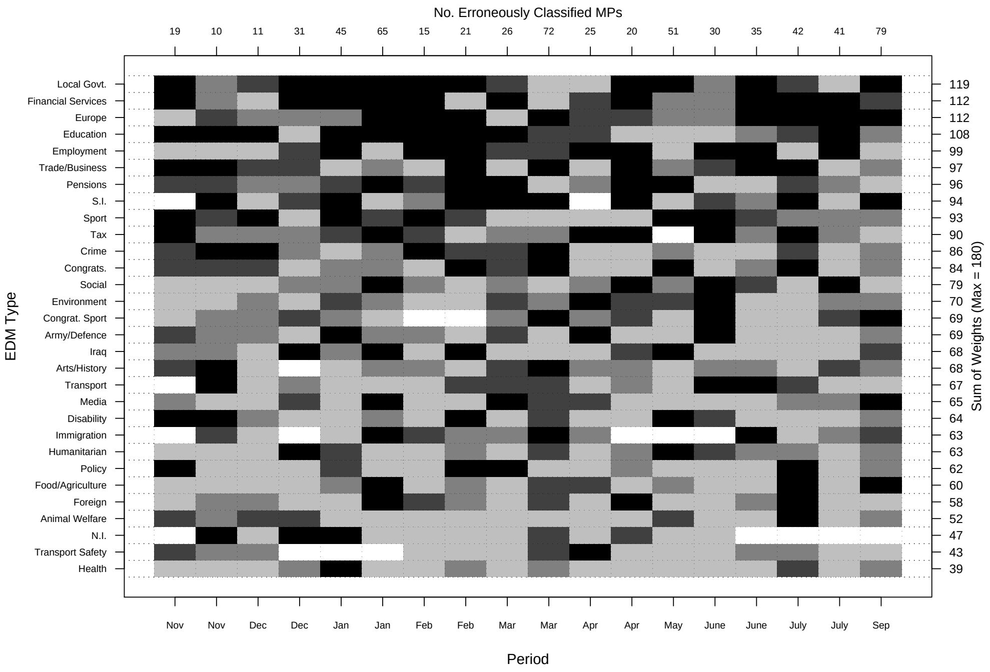


Figure 4.5: Issues separating main political parties 2002/03 session. Weight shading: white 0–2.5, light grey 2.5–5, dark grey 5–7.5, black 7.5–10.

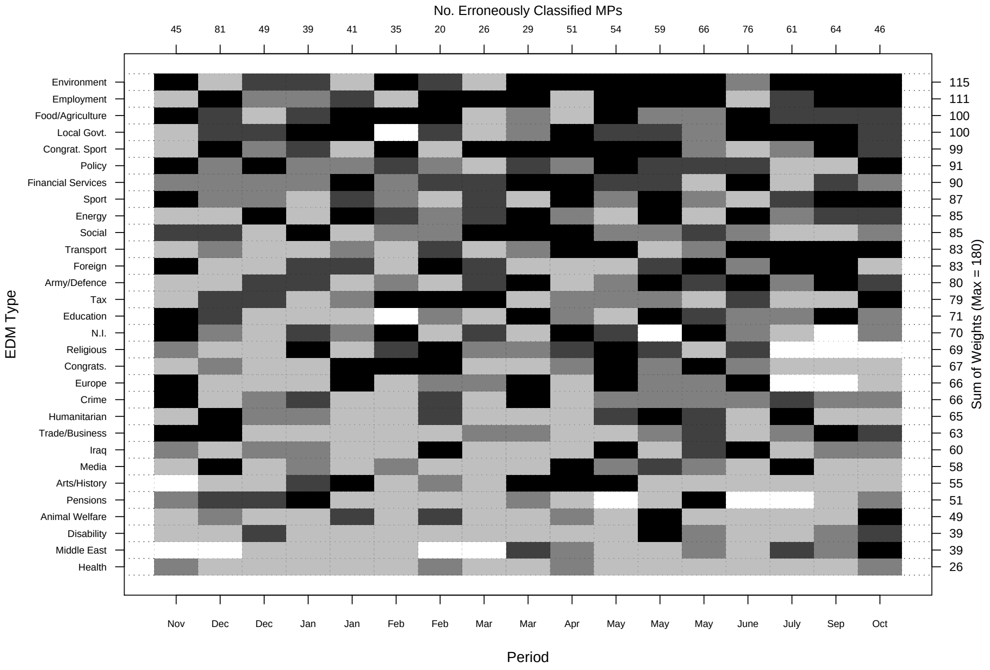


Figure 4.6: Issues separating main political parties 2003/04 session. Weight shading: white 0–2.5, light grey 2.5–5, dark grey 5–7.5, black 7.5–10.

to split the parties. Humanitarian issues, which are shown not to divide parties, could be considered *cross-party* EDMs, and are predictably given low weights. Animal welfare is also given a low weight and to a certain degree could be considered to unite the parties. This comes following a time of mass animal culling due to the foot and mouth outbreak (in early 2002 the UK was no longer considered ‘infected’, over a year since the initial outbreak).

For the 2002/03 session, plotted in Figure 4.5, similar issues to the previous session are towards the top and bottom of the scales. It is interesting to note that issues involving financial services came second and both Iraqi issues and army and defense issues were mid-table at a time when they were high on the political agenda. Further, health issues do not appear to be cause of much separation between the parties.

Dividing issues, which may be expected, are at the top of the results for the 2003/04 session in Figure 4.6. Congratulations in sport is perhaps unexpectedly high on the list of weights, with Environmental, along with Food and Agricultural issues also featuring highly. During a time which included the invasion of Iraq, issues with the Army and Defense were highly divisive although more specific issues concerning Iraq were not. Health is yet again low on the list.

The 2004/05 session (which we do not plot) was much smaller due to the general election in May 2005. Health is yet again low in the list of issues which separate the parties, whereas transport and army and defense issues feature highly.

As explained, these results should be treated as preliminary. We highlighted certain features but do not speculate on any reasons behind them at this stage. Nonetheless, results highlighted here show the wealth of information which is contained and is a motivation for future work.

We next digress somewhat to consider the extent to which propensity to sign EDMs affects the scaling solution and thus the results plotted in this chapter.

4.3.2 Scaling Solutions and Propensity to sign EDMs

We next briefly investigate the relationship between the number of EDMs an MP signs and the principal components (PC) of the scaling solution. This section aims to give a feel of the behaviour of the data by considing a single time window of EDMs. It is thus not an exhaustive investigation but highlights issues within the data.

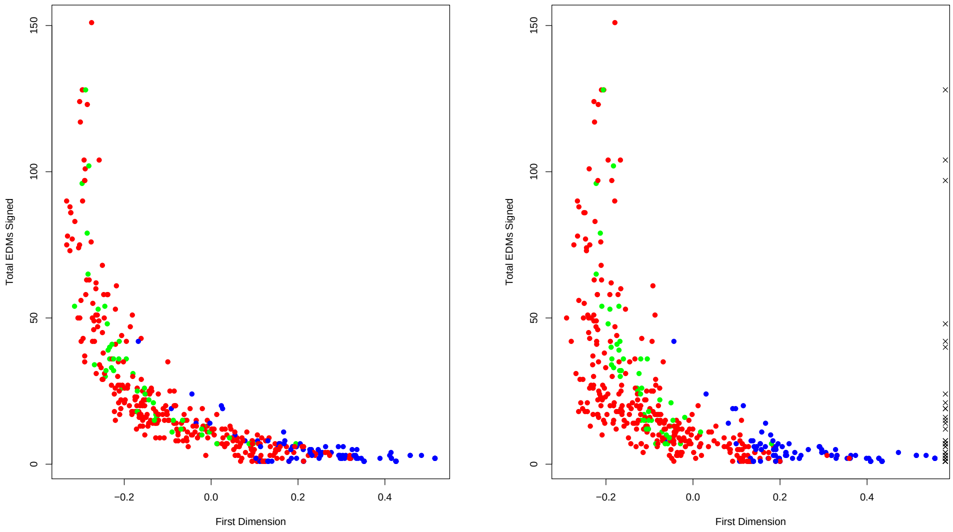
As discussed in Section 3.2, due to the asymmetric nature of the Jaccard coefficient of dissimilarity, the number of EDMs which an MP signs will have an impact on their dissimilarity. For the cohesion measure we accounted for this by simulating the signing of EDMs by MPs and using the cohesion of this data as a calibration level. We next show how the lower dimensions of the scaling solutions represent the MPs propensity to sign EDMs.

Figure 4.7(a) shows that the relationship between the first principal component and the total number of EDMs signed by each MP for the first time window of the 01/02 session. The number signed appears to be exponentially decreasing with the value of the principal component. This relationship is also observed, but to a lesser extent for the second PC. Higher dimensions, however, show no such trend.

The notion that propensity to sign is manifested within the scaling solution raises some interesting questions over the validity of the feature selection. Are the issues with higher weights also those which attract the most or least signatures? Are the MPs who are misclassified those who sign the most or least number of EDMs? Without considering these questions we do not know if the plots in Figures 4.4 – 4.6 are showing us anymore than just what issues are signed the most. We attempt to answer these questions by investigating the scaling solution and an MPs propensity to sign for a single time window.

Figure 4.7(b) again shows first dimension of the scaling solution against total number of EDMs signed (for each MP) using the optimal weighted scaling solution. The relationship is not as strong as the unweighted solution, although there is still a distinct decreasing pattern. Indicated on the far right of the plot are markers representing the number of EDMs each erroneously classified MP signed. We see from the plots that although there is a relationship between the two measures, it is not enough to be able to predict position of an MP, based on number of EDMs signed. Furthermore, there is nothing to suggest that MPs are misclassified based on the number of EDMs signed (this was also observed for the second dimension of the scaling solution).

Other investigations show very little evidence to suggest that an EDM type will be given a higher or lower weighting depending on how many signatures it receives.



(*a*) (*b*)

Figure 4.7: First principal component against total EDMs signed for first time window of 2001/02 session. (a) Unweighted, (b) Weighted. ● = Labour, ● = Conservative, ● = Liberal Democrat. X indicates number of EDMs signed by erroneously classified MPs.

4.4 Conclusions and Further Work

Much can be gained from the analysis of the feature selection. We see results which support known behaviour amongst the MPs and also issues which would not have been expected to divide the parties to such an extent. However, it is clear that computation constraints limit the results of this procedure. Although we take care to repeat the optimisation routine, there is still significant deviation between each set of results. Furthermore, the 'eleventh' repetition of the genetic optimisation, which has a starting population inclusive of the ten previous optimal solutions, rarely improves beyond those initial solutions. This either means that local optima are few and have been found by our procedure, or that there is much work to be done in this area. With 30 variables, the latter would be the obvious conclusion. Increasing the population size of the GA would increase the search area and could be coupled with a reduction in the number of generations to reduce computation. This would still be far from ideal and a reduction in the number of variables may be further beneficial, to reduce the current computational time.

In this chapter we used linear discriminant analysis as a quick and effective method for classification and therefore did not investigate how well suited it was to the scaling solution of the MPs. LDA has the assumption of normally distributed classes (or political parties in this case) and equal variance matrices in the groups. It may be that this method is not as suitable as others for the data, but as explained within the chapter, our aim was not to get the *best* (initial) classification of MPs, but to obtain *a* classification, on which we set to improve by assigning weights to the EDM types. LDA suited our needs for this initial investigation, being both simple and well understood. The development of other classification techniques are left as future work.

Although the results presented here are of a moving time window, both the size of the window and the *step size* are large. Any significant improvement in computational time, be it from reducing the complexity of the problem or an increase in computational resources could allow smaller step sizes and windows and give the analysis a more continuous feel.

We further investigated how the lower dimensions of the scaling solution are related to an MP's propensity to sign EDMs. We considered one time window of EDMs and found that such a relationship exists for the first two dimensions of the scaling solution. Higher dimensions did not show such a pattern. Furthermore, MPs do not appear to be misclassified based on the number of EDMs which they sign, and there was little evidence to suggest that EDMs with the most (or least) number of signatures were those picked by the feature selection

We believe that the results here are far from conclusive, but the methodology appears sound and worthy of further exploration. If efficiencies can be found in the methodology to increase accuracy (statistical, computational or otherwise) then many intriguing questions about the UK political parties may be answered.

Chapter 5

Literature Review II

5.1 Introduction

This chapter reviews literature which is related to the following chapters of this thesis. In contrast to previous chapters, the focus now turns to the transformation and variance stabilisation of data.

We first provide some theory of the discrete wavelet transform and some types of smoothing operators, suitable for both raw data and the wavelet transform of data. We then describe some models for time series count data before exploring some variance stabilising and Gaussianising transformations, which form the basis of our work in the remaining chapters.

5.2 Discrete Wavelet Transform

This section describes the discrete wavelet transformation (DWT) on a series of data $\mathbf{x} = x_1, \ldots, x_n$, where $n = 2^J$ and $J \in \mathbb{Z}$. For a more detailed introduction to wavelets, we refer the reader to Daubechies (1992) or Vidakovic (1999). For a more gentle introduction to wavelets, see Burrus *et al.* (1998). Details of the continuous wavelet transform can also be found within these references.

A wavelet family is generated by dilations and translations of a function ψ, called the mother wavelet. Wavelets have oscillating, wave-like characteristics, such that $\int \psi(x)dx =$

0, but have their ‘energy’ concentrated around a certain time point. Away from this point, wavelets have a fast decay.

A wavelet representation of a function $f(x)$ is generated from a *scaling function*, or *father wavelet* denoted by $\phi(x)$. This function belongs to a closed subspace of a multiresolution analysis in $L^2(\mathbb{R})$ (see Vidakovic (1999, page 51)). The set of $\phi(x)$ translated over $\mathbb{Z}$ form an orthonormal basis for the closed subspace within $L^2(\mathbb{R})$ and it can be shown that

$$\phi(x) = \sum_{k\in\mathbb{Z}} h_k \sqrt{2}\phi(2x-k),$$

which is known as the scaling equation. The coefficients $\{h_k\}_{k\in\mathbb{Z}}$ are the low-pass filter associated with ϕ. It can be further shown that

$$\langle \phi_{j-1,k}, \phi_{j,n} \rangle = h_{n-2k},$$

and that

$$\begin{aligned} \phi_{j-1,k}(x) &= \sum_{n\in\mathbb{Z}} \langle \phi_{j-1,k}, \phi_{j,n} \rangle \, \phi_{j,n}(x), \\ &= \sum_{n\in\mathbb{Z}} h_{n-2k}\phi_{j,n}(x), \end{aligned} \tag{5.2.1}$$

indicating that the scaling function at a certain scale can be expressed in terms of a translated scaling function from the next scale.

The mother wavelet can be represented in terms of the father wavelet as

$$\psi(x) = \sum_{k\in\mathbb{Z}} g_k \sqrt{2}\phi(2x-k).$$

The coefficients $\{g_k\}_{k\in\mathbb{Z}}$ are known as high pass filters associated with ψ. Similar to (5.2.1), it can be shown that

$$\psi_{j-1,k}(x) = \sum_{n\in\mathbb{Z}} g_{n-2k}\phi_{j,n}(x).$$

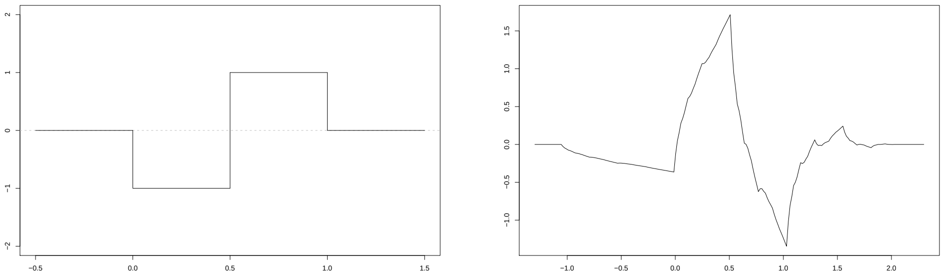


Figure 5.1: Left: Haar Wavelet; Right: Daubechies Extremal Phase wavelet with 2 vanishing moments.

The wavelet basis function, $\psi_{j,k}(x)$, is derived from the mother wavelet and defined by

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k), \qquad \text{for } j,k \in \mathbb{Z}.$$

Due to the construction of $\psi(x)$, the wavelet family $\{\psi_{j,k}(x)\}$ inherits the same orthonormal basis property as $\phi(x)$, but over the whole of $L^2(\mathbb{R})$.

The Haar wavelet is the simplest example of a wavelet. The Haar father wavelet can be defined by

$$\phi(x) = \begin{cases} 1, & x \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

From this, the Haar mother wavelet $\psi(x)$ can be derived, and shown to be:

$$\begin{aligned} \psi(x) &= \phi(2x) - \phi(2x-1) \\ &= \begin{cases} -1 & 0 \leq x \leq 1/2, \\ 1 & 1/2 \leq x \leq 1, \\ 0 & \text{otherwise}. \end{cases} \end{aligned} \tag{5.2.2}$$

Figure 5.1 shows two examples of wavelets; the Haar wavelet (left) and the Daubechies extremal phase wavelet with two vanishing moments (right). For more details of these, and other wavelets, see Daubechies (1992).

It can be shown that a function $f(x)$ in $L^2(\mathbb{R})$ can be represented as

$$f(x) = \sum_{k\in\mathbb{Z}} s_{j_0,k}\phi_{j_0,k}(x) + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(x), \tag{5.2.3}$$

where j_0 is known as the *primary resolution level* and $s_{j_0,k}$ and $d_{j,k}$ are the father and mother wavelet coefficients (also known as the $smooth$ and $detail$ coefficients). These coefficients are the inner products

$$\begin{aligned} s_{j_0,k} &= \langle f(x), \phi_{j_0,k}(x)\rangle \, , \\ d_{j,k} &= \langle f(x), \psi_{j,k}(x)\rangle \, . \end{aligned} \tag{5.2.4}$$

The representations of $f(x)$ in (5.2.3) and (5.2.4) highlight the recursive nature of the wavelet transform when computing the wavelet coefficients: an initial set of smooth and detail coefficients are formed, from which all other coefficients are then generated. These coefficients give information about the function $f(x)$ at a scale 2^j at the location, or time point $2^{-j}k$.

At the finest scale of the transform, the smooth coefficients, $s_{j_0,k}$, can be thought of as a smoothed representation of the original data whereas the detail coefficient, $d_{j,k}$ can be thought of as the information lost by this smoothing operator. For an efficient representation, we want sparse sets of wavelet coefficients, that is, many of the detail coefficients to be close to, or equal to zero. This is useful for compression and also for denoising, which we shall discuss later.

In the following chapters, we are primarily concerned with the application, rather than the theory, of the discrete wavelet transform. We thus outline here the methodology of obtaining the wavelet coefficients in an applied, rather than theoretical sense.

The DWT first forms smooth and detail vectors of the sequence of observations $\mathbf{x}$. These coefficients are known as the finest level of the transform. Coarser levels are found by recursively performing the decomposition on the smooth coefficients, until we are left with just a single smooth and detail coefficient. Computationally, the detail coefficients are stored at each level of the transformation, whereas the smooth coefficients are used to form the next set of coefficients.

As an example of this decomposition, we next define the wavelet transform (and its inverse) for the Haar mother wavelet from (5.2.2), which is used extensively in the following chapters. We follow the exposition (and notation) from Fryzlewicz *et al.* (2007).

Given an input vector $\mathbf{x} = (x)_{i=1}^{n}$ where $n = 2^J$, the Discrete Haar Transform (DHT) is performed as follows:

1. Let $s_i^J = x_i$, for $i = 1, \ldots, n$.
2. For each $j = J-1, J-2, \ldots, 0$, recursively define vectors $\mathbf{s}^j$ and $\mathbf{d}^j$:

   $$s_k^j = \frac{s_{2k-1}^{j+1} + s_{2k}^{j+1}}{2};\ d_k^j = \frac{s_{2k-1}^{j+1} - s_{2k}^{j+1}}{2}, \tag{5.2.5}$$

   for $k = 1, \ldots, 2^j$.
3. The operator H, where $H\mathbf{x} = (\mathbf{s}^0, \mathbf{d}^0, \ldots, \mathbf{d}^{J-1})$, defines the DHT.

The inverse DHT is performed as follows:

1. For each $j = 0, 1, \ldots, J-1$, recursively form s^{j+1}

   $$s_{2k-1}^{j+1} = s_k^j + d_k^j;\ s_{2k}^{j+1} = s_k^j - d_k^j \tag{5.2.6}$$

   for $k = 1, \ldots, 2^j$.
2. Set $x_i = s_i^J$, for $i = 1, \ldots, n$.

The elements of $\mathbf{s}^j$ (and $\mathbf{d}^j$) are the smooth (and detail) of the original vector $\mathbf{x}$ at scale 2^j.

Each recursive step of the wavelet transform produces half the number of smooth and detail coefficients as the previous level. This is known as *decimation* or *downsampling*. Most wavelet methods (and methods used within this thesis) use such ‘filters’ with a base equal to 2 so that the number of coefficients are halved. See Vidakovic (1999) for more details.

The exposition above can be easily adapted for wavelets other than Haar. The formulation of the smooth and detail coefficients involve using more coefficients from the previous level of the wavelet transform. Haar only uses 2 points, if say another wavelet used 3 points and we still moved over the data in pairs, two adjacent coefficients would depend on the same element from the previous level, although these will be given a different weight for the different coefficient locations. This overlap of coefficients adds to the smoothness of the wavelet.

Both the downsampling, and the number of elements used in the wavelet transform can cause boundary problems, where the number of elements needed in the formulation of the smooth and detail coefficients are greater than the number which exist. We direct the reader to (Nason, 2008, chap 2) for a discussion on boundary conditions for the discrete wavelet transform.

We make extensive use of the Haar wavelet transform within our variance stabilising procedures, which are described in Section 5.5.4.

As previously mentioned, the sparse nature of the wavelet decomposition makes the wavelet transform useful for denoising. Suppose we have an observed signal which is believed to be composed of ‘signal + noise’. If there were no noise present, we would expect the detail coefficients of the wavelet decomposition to be sparse, as the detail lost from the smoothing coefficients will be small (as the signal itself is smooth).

Noise within the signal would cause the detail coefficients to be non-zero as they detect the sudden jumps in the signal. Thresholding can be used to ascertain which of the non-zero coefficients are purely noise, and which represent signal information. A *thresholding* level can be set, below which the detail coefficients are believed to contain ‘noise’. These wavelet coefficients are then set to zero, and the inverse wavelet transform will result in a noise free signal.

There are many methods to calculate a thresholding level, and Jansen (2001) gives an overview of many popular schemes. A common thresholding method used in the remaining chapters is empirical Bayes thresholding, by Johnstone & Silverman (2004, 2005b), which we describe in Section 5.3.1.

5.3 Some Smoothing Methods

In this section we outline some smoothing methods which are commonly used in the following chapters. We begin with empirical Bayes thresholding. Although it can be applied generally to any sequence of data, we focus on its use in the thresholding of wavelet coefficients to remove noise. We then outline smoothing filters and kernel estimators, which can be applied to most datasets and are used in the following chapters. These methods are either currently used with certain data sets, or have been shown to perform well over a wide range of data and are used within the following chapters. Furthermore, all methods have been coded for use in the statistical program R on computers with a 2.2GHz AMD Opteron processors and 2Gb RAM.

5.3.1 Empirical Bayes Thresholding of Wavelets

We next describe empirical Bayes thresholding as described by Johnstone & Silverman (2004, 2005b). Its implementation in R is via the `EbayesThresh` package, detailed by Johnstone & Silverman (2005a). This technique is shown to perform well over a range of simulated and real data, and as such we use it often in the the following chapters.

Although the ideas can be applied to many data sequences, we focus on its use for *thresholding* the coefficients of wavelet decomposition. As mentioned in Section 5.2, an efficient wavelet representation of many classes of function has sparse wavelet coefficients and these can be thresholded to obtain an estimate of the underlying signal of the original data. The idea therefore assumes sparsity of the wavelet coefficients of the true signal and the empirical Bayes thresholding approach from Johnstone & Silverman (2004, 2005b) places a prior on the true wavelet coefficient of the form

$$d^*_{j,k} \sim (1 - w_j)\delta_0 + w_j\gamma, \tag{5.3.7}$$

for each level j. Here, w_j is the (prior) probability of the wavelet coefficient being non-zero, γ is the density of the wavelet coefficient, conditional on it being non-zero, and δ_0 is the density conditional on it being zero. The method uses a heavy-tailed distribution for γ, such as the Laplace distribution, or the 'quasi-Cauchy' density which is defined by Johnstone & Silverman (2005b).

If $d^*_{j,k}$ has the prior distribution (5.3.7) and the observed wavelet coefficient is such that $d_{j,k} \sim N(d^*_{j,k}, \sigma^2)$, we can find the posterior distribution of $d^*_{j,k}$ conditional on the observation $d_{j,k}$. The median of this posterior distribution can be used as an estimate for the 'true' wavelet coefficient $d^*_{j,k}$. This acts as a thresholding method, since for a fixed w, there will be a function $t(w)$ such that the median will equal zero if and only if $|d^*_{j,k}| \leq t(w)$ (see Johnstone & Silverman (2005b) for more detail). Other thresholding methods involving the (post) mean, and soft or hard thresholding of the function $t(w)$ are also considered by Johnstone & Silverman (2004, 2005b).

In statistics, wavelet coefficients are assumed to be typically sparse at finer resolution levels with the coarser levels having larger detail coefficients representing a lot of signal. Johnstone & Silverman (2004, 2005b) suggest applying empirical Bayes separately to each level of the wavelet transform. Further, they suggest finding the parameters of the prior distribution at each scale using marginal maximum likelihood estimators.

Isotonic Regression

At this points it is worth mentioning isotonic regression, which is used within empirical Bayes thresholding (and can be performed using the `isotone` function in the `EbayesThesh` package) as it is also used within the data-driven Haar-Fisz transform (see Section 5.5.4) to estimate a non-decreasing monotonic function for data with a particular mean-variance relationship.

Given our sequence of values x_i and a set of weights w_i, the least squares isotone regression finds the monotonic increasing sequence x^*_i for which

$$\sum w_i(x_i - x^*_i)^2, \tag{5.3.8}$$

is minimised. This is achieved using the pool-adjacent violators algorithm (see Friedman & Tibshirani (1984) for an overview), modified to allow for the weights w_i.

Briefly, the method identifies local maxima and minima in the sequence in order to locate decreasing subsequences. These subsequences in the data are replaced by a single value equaling the weighted values of the subsequence. The corresponding weights are replaced by the sum of the weights over the subsequence. This procedure is iterated until (5.3.8) is minimised.

Isotonic regression is used extensively in the remainder of this thesis, indirectly as part of empirical Bayes thresholding and directly in modifications to the data-driven Haar-Fisz transform.

5.3.2 Smoothing Filters

Filter smoothing of data smooths in a given window around each data point. The filter smoothing function, f at time t, taken over data, x is given by

$$f_t = b^{-1} \sum_{i=1}^{b} c_i x_{t+i-\lfloor b/2 \rfloor}, \tag{5.3.9}$$

where b is the number of observations used to form each local mean (the bandwidth), $\lfloor x \rfloor$ is the largest integer less than or equal to x and $\{c_i\}^b$ is the set of filter coefficients. When $c_i = 1$ for $i = 1, \ldots, b$, (5.3.9) is the *running means* estimator of x_t. Simply put, it is the mean of a window of b observations surrounding the data point x_t. We use the running means filter in Chapter 6.

A filter much used with meteorological data (and used in Chapter 7) is the binomial filter. For a filter size of b, and $m = 0, \ldots, b$, the coefficients c_i of the filter (5.3.9) take the following form

$$c_i = k_m / \sum_{m=0}^{b} k_m \qquad \text{where} \qquad k_m = \frac{b!}{m!(b-m)!}.$$

with a slight change in notation from (5.3.9) in that $i = 0, \ldots, b$. Panofsky & Brier (1968) give a numerical example of a binomial filter, whereas Aubury & Luk (1995) give a more thorough account of the theoretical properties.

For independent, identically distributed data from the Poisson distribution, where $x_t \sim \mathrm{Poi}(\lambda)$, it is clear from (5.3.9) that $\mathrm{var}(f_t)$ will comprise of (up to a constant) the sum of the variances of the Poisson variables, i.e.

$$\begin{aligned}
\mathrm{var}(f_t) &= \mathrm{var}(b^{-1}\textstyle\sum_{i=1}^{b} c_i x_{t+i-\lfloor b/2\rfloor}), \\
&= b^{-2}\textstyle\sum_{i=1}^{b} \mathrm{var}(c_i x_{t+i-\lfloor b/2\rfloor}), \\
&= b^{-2}\textstyle\sum_{i=1}^{b} c_i^2\,\mathrm{var}(x_{t+i-\lfloor b/2\rfloor}), \qquad \text{since } c_i \text{ does not depend on } t, \\
&= b^{-2}\textstyle\sum_{i=1}^{b} c_i^2\lambda, \\
&= (\lambda/b^2)\textstyle\sum_{i=1}^{b} c_i^2.
\end{aligned} \tag{5.3.10}$$

As $\mathrm{var}(x_t) = \mathbb{E}(x_t) = \lambda$, the variance is proportional to the mean, and the estimator will exhibit more variance when the signal itself has more variance (i.e. larger mean).

5.3.3 Kernel Smoothing

Kernel estimators smooth data by considering weighted data surrounding the point of estimation. The kernel regression estimator for data $x_1, \ldots, x_n$ is given by

$$Y_i = r(x_i) + \varepsilon_i,$$

for a regression function r where ε are independent and identically distributed with $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2$ for $i = 1, \ldots, n$. An estimator $\hat{r}$ for the regression curve can be derived from the kernel density estimator. One popular estimator is the *Nadaraya-Watson estimator* (Nadaraya (1964), Watson (1964)), defined as

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{b}) y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{b})},$$

for a kernel function K, bandwidth b and the estimated density is at point x. For a full introduction to kernel smoothing, see Wand & Jones (1995) or Simonoff (1996)

The kernel regression estimator much used in the remainder of this thesis is the local plugin bandwidth estimator, known as *lokern*, by Brockmann *et al.* (1993). Lokern is a non-parametric regression estimation technique using kernel estimators. The procedure automatically chooses a local plugin bandwidth, which puts special weight on the stability aspects of the bandwidth size so that the estimator is not too noisy (or too smooth). The optimal bandwidths are estimated by considering the asymptotic optimal mean square error of the bandwidths.

A common feature of local bandwidth estimators is that a gain in the mean integrated square error (MISE) is coupled with a larger variability in the estimator, particularly if the sample size is small. This is considered in case of the lokern estimator, by Brockmann *et al.* (1993) and the resulting estimator can adapt for different features within the data.

An equivalent method using a global bandwidth estimator is also described by Brockmann *et al.* (1993). Lokern is shown to improve on this, even with small sample sizes. Lokern can be implemented in R using the `lokern` add-on package.

5.4 Time Series Count Data Models

In this section we review some models of count data. We refer the reader to Jung *et al.* (2006) for a nice review and comparison of a whole series of techniques. A wide variety of models are shown in the books by MacDonald & Zucchini (1997), Winkelmann (2003) and Cameron & Trivedi (1998). This section is not exhaustive and we aim only to give a flavour of some of the models which are currently used. We discuss at the end of the section why we do not consider such models in the remainder of this thesis.

Count data models are broadly classified as being observation-driven or parameter-driven. In the former, the conditional distribution of a time series y_t is specified as a function of past observations $y_{t-1}, \ldots, y_1$ whereas in the parameter-driven model, autocorrelation is introduced through a latent process.

Jung *et al.* (2006) compare two such models in terms of their ability to account for dynamic and distributional properties of count data (the work also proposes a new method of estimation of the likelihood for the model). The well known parameter-driven Poisson model with stochastic autoregressive mean (SAM) by Zeger (1988) is used in the paper. This models a time series of observations y_t for time $t = 1, \ldots, T$, on a sequence of covariates x_t by

$$y_t|(x_t, u_t) \sim \mathrm{Poi}(\exp\{x_t'\phi\}u_t),$$

where u_t is a latent non-negative stochastic process and ϕ is a vector of regression parameters. The conditional distribution of $y_t|(x_t, u_t)$ is therefore assumed to be Poisson with mean $\mu_t = \exp(x_t'\phi)u_t$. The latent process u_t is introduced to account for possible over-dispersion and serial correlation within the data and it is often assumed that $\lambda_t = \ln(u_t)$ is a Gaussian first order autoregressive process, satisfying

$$\lambda_t = \delta\lambda_{t-1} + v\varepsilon_t, \qquad \varepsilon \sim \text{iid}\, N(0,1).$$

The parameters, ϕ, δ and v are to be estimated.

A recent observation-driven model is the autoregressive conditional Poisson (ARCP) model by Heinen (2003). For a time series of data y_t with all prior observations denoted by Y_{t-1}, y_t is modelled by

$$y_t|Y_{t-1} \sim \text{Poi}(\mu_t),$$

with an autoregressive conditional mean

$$\mathbb{E}(y_i|Y_{t-1}) \equiv \mu_t = \omega + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{j=1}^{q} \beta_j \mu_{t-j},$$

with α_j, β_j and ω being positive, to ensure μ_t is non-negative.

A further observation-driven model worth noting is the first order Poisson autoregressive (AR(1)) process by McKenzie (1988). This is considered by Al-Osh & Alzaid (1988) as a special case of their integer valued autoregressive model (INAR), first described in Al-Osh & Alzaid (1987).

A random variable is said to follow a first order INAR process with Poisson marginals (written $y \sim \text{INAR}(1)$) if

$$y_t = \alpha \circ y_{t-1} + \varepsilon_t.$$

Amongst many conditions on the variables (as summarised by Winkelmann (2003)), y_{t-1} and ε_t are independent Poisson variables.

The symbol $\circ$ represents a mixture operation and $\alpha \circ y_t$ denotes the number of elements out of $t-1$ that survive to period t. The probability of survival is given by α. For the AR(1) process by McKenzie (1988), $\circ$ is *binomial thinning*.

Winkelmann (2003) and MacDonald & Zucchini (1997) detail and discuss these models, as well as describing the Poisson moving average process of order one by McKenzie (1988), the similar integer-valued moving average (INMA) process from Al-Osh & Alzaid (1988) and the INAR(1) model for negative binomial marginals.

The models described above are just some of a range of count data models. All essentially involve the estimation of parameters, a task in itself which is often not straightforward. The models also concern data in which the counts are Poisson-like and are correlated. In the remainder of this thesis we use and develop techniques which do not require Poisson data (although there are assumptions on the mean-variance relationship) and are therefore more widely applicable. Furthermore, for the Iraq and central England temperature data sets, introduced in Chapters 6 and 7 respectively, we show that after mean correction and transformation, the time series are not correlated, raising further questions of the appropriateness of such count data models for our data.

5.5 Variance Stabilising Transforms

We now describe some variance-stabilising (and Gaussianising) transformations which are used in the remainder of this thesis.

5.5.1 Anscombe Square Root Transform

For Poisson data, where $r \sim \mathrm{Poi}(\lambda)$, Anscombe (1948) derived the transformation

$$y = \sqrt{r + c}, \tag{5.5.11}$$

for stabilising the variance of the variable r. The mean and variance of the transform can be calculated via a Taylor series expansion and it can be shown that for large λ, the transformed variable y has a 'most nearly constant variance' of $\frac{1}{4}$ when the constant, $c = \frac{3}{8}$. See Nason (2008, chap. 6) for further details and a derivation of the square root transform.

Zhang *et al.* (2006) use a similar transformation within their procedure for variance stabilisation of Poisson counts. First, the observed signal is filtered, and the resulting signal is variance stabilised using the Anscombe transform, but pre-multiplied by a constant. Although different 'filters' can be used, the method focuses on the use of wavelets to first transform the observed data. Methods are outlined to estimate the constants of the transformation set as well as a wavelet based denoising step.

5.5.2 Box-Cox

The primary objective of the Box-Cox transformation is the *Gaussianisation* of the observed signal (that is, to make the observed signal more Gaussian). A secondary effect of this is that the variance of the data is often stabilised (as discussed by Kendall *et al.* (1983, page 103)).

Due to its good Gaussianising performance, the transformation is used heavily in the following chapters as a comparison to our new methods. We also use the ideas of finding optimal transformation parameters via likelihood functions in Chapter 8, and as such, we give the transformation and derivation of parameter estimation in detail.

Box & Cox (1964) consider the parametric family of power transformations

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases} \tag{5.5.12}$$

For some unknown λ, the transformed observations are assumed to be independently Gaussian, with constant variance σ^2 and expectation

$$\mathbb{E}\{\mathbf{y}(\lambda)\} = \mu,$$

where $\mathbf{y}(\lambda)$ is the vector of transformed observations $y_1(\lambda), \ldots, y_n(\lambda)$.

The likelihood of the transformed variables in relation to the original observations is obtained by multiplying the Gaussian density by the Jacobian of the transformation. The likelihood $L(\lambda, \mu, \sigma^2)$, dependent on the transform parameter λ, mean μ and variance σ^2 is

$$L(\lambda, \mu, \sigma^2) = \prod_{i=1}^{n} ((2\pi\sigma^2)^{-1/2} \exp\left\{-(y_i(\lambda) - \mu)^2/2\sigma^2\right\} J, \tag{5.5.13}$$

where the Jacobian, J, is

$$J = \prod_{i=1}^{n} \left| \frac{\partial y_i(\lambda)}{\partial y_i} \right| . \tag{5.5.14}$$

The maximum likelihood estimates are found in two steps. First, for a fixed λ, (5.5.13) is, except for a constant factor, the likelihood of the least squares problem with response $y(\lambda)$ (as the Jacobian does not involve μ or σ). Hence the maximum-likelihood estimate of μ, denoted for a fixed λ by $\widehat{\mu}(\lambda)$, is

$$\widehat{\mu}(\lambda) = \overline{\mathbf{y}(\lambda)},$$

the mean of the transformed observations. The estimate of σ^2 for a given λ, $\widehat{\sigma}^2(\lambda)$, is

$$\widehat{\sigma}^2(\lambda) = \sum_{i=1}^{n} (y_i(\lambda) - \overline{\mathbf{y}(\lambda)})/n = S(\lambda)/n,$$

where $S(\lambda)$ is the residual sum of squares of the $y_i(\lambda)$.

Thus, for a fixed λ, the partially maximised log-likelihood is, up to proportionality,

$$l_{\text{max}}(\lambda) = -(n/2)\log\hat{\sigma}^2(\lambda) + \log J, \tag{5.5.15}$$

and is therefore a function of λ which depends both on the residual sum of squares $S(\lambda)$ and on the Jacobian J (where $J = J(\lambda)$).

A simpler form of $l_{\text{max}}(\lambda)$ can be obtained by working with the normalised transformation

$$z(\lambda) = y(\lambda)J^{1/n}. \tag{5.5.16}$$

For the transformation in equation (5.5.12), we have

$$\partial y_i(\lambda)/\partial y_i = y_i^{\lambda-1},$$

which gives

$$\log J = (\lambda - 1)\sum \log y_i.$$

The normalised power transform can then be written as,

$$z(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}, & \text{if } \lambda \neq 0, \\ \dot{y} \log y, & \text{if } \lambda = 0, \end{cases} \tag{5.5.17}$$

where $\dot{y}$ is the geometric mean of the observations.

The partially maximised log-likelihood of the observations can then be written, apart from constant, as

$$l_{\text{max}}(\lambda) = -(n/2)\log(R(\lambda)/n), \tag{5.5.18}$$

where

$$R(\lambda) = \widehat{\sigma_z^2}(\lambda)n, \tag{5.5.19}$$

is the residual sum of squares of $z(\lambda)$ and $\widehat{\sigma_z^2}$ is the variance of the transformed observations.

The maximum likelihood estimate $\hat{\lambda}$ is the value of the transformation parameter for which $l_{\text{max}}(\lambda)$ is a maximum. Equivalently, it is the value for which the residual sum of squares, $R(\lambda)$ is minimised. A common way to find $\hat{\lambda}$ is to plot $l_{\text{max}}(\lambda)$ (or $R(\lambda)$) for various values of λ.

An extended form of the transformation in (5.5.12) which takes two parameter values was also proposed by Box & Cox (1964) and is defined as:

$$y(\lambda) = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda_1}, & \text{if } \lambda_1 \neq 0; \\ \log(y+\lambda_2), & \text{if } \lambda_1 = 0. \end{cases} \tag{5.5.20}$$

The additional parameter allows for a constant to be added (or subtracted) from the data before transformation as with the one parameter model. An example of its use is in survival time experiments, where the origin of the response is not a true lower limit and thus a constant is subtracted from the data. It can also be used to estimate an optimal constant to add to negative data to ensure positivity.

Estimation of the second parameter can be incorporated into the likelihood equation of (5.5.13) and continued as with the one parameter transformation. The Gaussianised form, equivalent to (5.5.17), is given by

$$z(\lambda_1, \lambda_2) = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda \left\{G(y+\lambda_2)\right\}^{\lambda_1 - 1}} = \dfrac{q^{\lambda_1} - 1}{\lambda \dot{q}^{\lambda_1 - 1}}, & \text{if } \lambda_1 \neq 0, \\ \log(y+\lambda_2) G(y+\lambda_2) = \dot{q} \log q, & \text{if } \lambda_1 = 0, \end{cases} \tag{5.5.21}$$

where $\dot{q} = G(y + \lambda_2)$ is the geometric mean of the observations after addition of the parameter λ_2.

Analogously to (5.5.19), the partially minimised sum of squares for the two parameter model, $R(\lambda_1, \lambda_2)$ is defined as

$$R(\lambda_1, \lambda_2) = \widehat{\sigma_z^2}(\lambda_1, \lambda_2) n. \tag{5.5.22}$$

where $\widehat{\sigma_z^2}(\lambda_1, \lambda_2)$ is the variance of the transformed observations. For the one parameter model, plots of $R(\lambda)$ against λ are close to a parabola. As described by Atkinson (1987), contour plots of (5.5.22) fall into two broad classes. In some examples there is a local minimum of the sums of squares surface, in the region of which the contours are approximately elliptical. For the estimates producing these local minima, approximate Gaussianity holds. In other examples, the residual sums of squares declines steadily to zero as λ_2 approaches $-y_{min}$. There may or may not be a local minima, but there will always be a region of parameter space in which $R(\lambda_1, \lambda_2)$ can be made arbitrarily small. This can be shown as follows. If $0 < \lambda_1 < 1$, both λ_1 and $1 - \lambda_1$ are greater than zero. Then from (5.5.21), $z(\lambda_1, \lambda_2)$ can be written as

$$z(\lambda_1, \lambda_2) = \dot{q}^{1-\lambda_1}(q^{\lambda_1} - 1)/\lambda_1.$$

As $\lambda_2 \rightarrow -y_{min}$, at least one value of q becomes very small and thus $\dot{q}$ will also become small. It follows that $\dot{q}^{1-\lambda_1}$ becomes small and the residual sum of squares decreases to zero. It will therefore always be possible to make $R(\lambda_1, \lambda_2)$ arbitrarily small, but the resulting parameters may not result in Gaussianity of the transformed observations. We direct the reader ahead to Figure 7.4 for a plot of the residual sums of squares, which displays some of the problems outlined here.

5.5.3 Some Data-Driven Variance Stabilising Transformations

When the underlying noise distribution of a signal is unknown, it is often desirable to estimate the variance stabilising transform from the data. Procedures using a data-driven method include the ACE (alternating conditional expectation) procedure by Breiman & Friedman (1985) and AVAS (additivity and variance stabilization), by Tibshirani (1988).

Given random variables X and Y, the ACE procedure looks to find the transformations $\theta(Y)$ and $\phi(X)$ that maximise the correlation between the transformed variables, cor($\theta(Y)$, $\phi(X)$), subject to $\mathrm{var}(\phi(X)) = 1$. The transformations have the added property that they minimise $\mathbb{E}(\theta(Y) - \phi(X))^2$, subject to $\mathrm{var}(\phi(Y)) = 1$.

The procedure is iterative and alternates between the two conditional expectations

$$\theta(X) = \mathbb{E}(\phi(Y)|X) \qquad \text{and} \qquad \phi(Y) = \frac{\mathbb{E}(\theta(X)|Y)}{[\mathrm{var}(\mathbb{E}(\theta(X)|Y))]^{1/2}}, \tag{5.5.23}$$

using the previous expectation of one function to get an update of the other until $\mathbb{E}(\theta(Y) - \phi(X))^2$ no longer decreases.

When the distribution of the data is unknown, scatterplot smoothers are used to replace the conditional expectation in (5.5.23).

Tibshirani (1988) points out several drawbacks of the ACE algorithm and suggests that it is better suited for correlation analysis rather than regression. The AVAS algorithm, which is a modification of the ACE procedure has several advantages as it is designed specifically for regression (advantages stated by Tibshirani (1988) include being able to reproduce model transformations and sensitivity to the marginal distribution of the predictors, X).

A further transform is by Linton *et al.* (1997) who detail an algorithm for transforming additive nonparametric regression models and derives asymptotic distributions of the estimators.

5.5.4 Haar-Fisz Transform

The Haar-Fisz (HF) transform, proposed by Fryzlewicz & Nason (2004), uses the Haar wavelet transform to decompose the input vector into smooth and detail coefficients and then stabilises their variance at all levels. The HF uses the mean-variance relationship of the data to smooth the Haar coefficients.

We first define the discrete Haar transform (DHT), before describing modifications to it resulting in the Haar-Fisz transform (HFT) for Poisson data and the data-driven Haar Fisz transform (DDHFT), from Fryzlewicz *et al.* (2007), for data where the exact distribution is unknown. We follow the exposition from Fryzlewicz *et al.* (2007) (and also use the same notation for wavelet coefficients with subscript denoting level, superscript denoting location, in the rest of this thesis).

Let $\mathbf{X} = (X_i)_{i=1}^n$ denote an input vector to the HF transform. The following list specifies the generic distributional properties that X must possess.

1. The length, n, of $\mathbf{X}$ must be a power of two. We denote $J = \log_2(n)$.

2. $(X)_{i=1}^n$ must be a sequence of independent, nonnegative random variables with finite positive means $\mu_i = \mathbb{E}(X_i) > 0$ and finite variances $\sigma_i^2 = \mathrm{var}(X_i) > 0$).

3. The variance of σ_i^2 must be a non-decreasing function of the mean μ_i:

$$\sigma_i^2 = h(\mu_i), \tag{5.5.24}$$

   where the function h is independent of i.

For Poisson data, $X_i \sim \mathrm{Poi}(\lambda_i)$, we have $\mu_i = \lambda_i$ and $\sigma_i^2 = \lambda_i$, which gives $h(\mu) = \mu$.

The Haar-Fisz transform (HFT) decomposes data $\mathbf{X} = (X_i)_{i=1}^n$, where $n = 2^J$ using the Discrete Haar Transform (DHT) described in Section 5.2. The detail coefficients, d_k^j are then modified with the aim of stabilising their variance (and making them closer to Gaussian). The inverse DHT is then applied to these modified coefficients to bring the sequence back to the original data domain, where the transformed data has a stabilised variance and is also closer to Gaussian. We now describe the variance stabilising and Gaussianising of the d_k^j.

Consider first $d_1^{J-1} = (X_1 - X_2)/2$. Assume that the Poisson distributions of X_1 and X_2 are identical (which is likely if the underlying mean is piecewise constant). This implies that the distribution of d_1^{J-1} is symmetric around zero. We want to stabilise the variance of d_1^{J-1} around $2^{(J-1)-J} = 1/2$. So, we divide d_1^{J-1} by $2^{1/2}$ times its own standard deviation.

We have

$$\mathrm{var}(d_1^{J-1}) = 1/4(\mathrm{var}(X_1) + \mathrm{var}(X_2)) = \sigma_1^2/2,$$

which gives $2^{1/2}(\mathrm{var}(d_1^{J-1}))^{1/2} = \sigma_1 = h^{1/2}(\mu_1)$. In practice μ_1 is unknown and we estimate it locally by $\mu_1 = (X_1 + X_2)/2 = s_1^{J-1}$. The approximate variance-stabilised coefficient f_1^{J-1} is given by:

$$f_1^{J-1} = \frac{d_1^{J-1}}{h^{1/2}(s_1^{J-1})}.$$

Fryzlewicz *et al.* (2007) continue this example to find a value for f_1^{J-2} (and subsequent levels), which are of a similar form to f_1^{J-1}. The coefficients f_k^j are called the *Fisz coefficients* of **X** (as Fisz (1955) studied properties of variables of a similar form to f_k^j).

We now give the general algorithm for the HFT when the function h is known.

1. Let $s_i^J = X_i$, for $i = 1, \ldots, n$.

2. For each $j = J-1, J-2, \ldots, 0$, recursively form the vectors $\mathbf{s}^j$ and $\mathbf{f}^j$:

$$s_k^j = \frac{s_{2k-1}^{j+1} + s_{2k}^{j+1}}{2}; \; f_k^j = \frac{s_{2k-1}^{j+1} - s_{2k}^{j+1}}{2h^{1/2}(s_k^j)}, \tag{5.5.25}$$

   for $k = 1, \ldots, 2^j$.

3. For each $j = 0, 1, \ldots, J-1$, recursively modify s^{j+1}:

$$s_{2k-1}^{j+1} = s_k^j + f_k^j; \; s_{2k}^{j+1} = s_k^j - f_k^j,$$

   for $k = 1, \ldots, 2^j$.

4. Set $\mathbf{Y} = \mathbf{s}^J$.

The relation $\mathbf{Y} = \mathcal{F}_h\mathbf{X}$ defines a nonlinear, invertible operator $\mathcal{F}_h$ which is called the Haar-Fisz transform (of **X**) with variance function h.

5.5.5 Data-Driven Haar-Fisz Transform

When h is unknown it must be estimated from the data. Since $\sigma_i^2 = h(\mu_i)$, we estimate the mean and variance of $X_1, X_2, \ldots$ and use these values to estimate h. Let the empirical estimates of the mean and variance of X_i be

$$\hat{\sigma}_i^2 = \frac{(X_i - X_{i+1})^2}{2},$$

and

$$\hat{\mu}_i = \frac{X_i + X_{i+1}}{2},$$

respectively.

The regression model

$$\hat{\sigma}_i^2 = h(\mu_i) + \varepsilon_i$$

is used, where $\varepsilon_i = \hat{\sigma}_i^2 - \sigma_i^2 = (X_i - X_{i+1})^2/2 - \sigma_i^2$ and “in most cases” $\mathbb{E}(\varepsilon_i) \approx 0$, to estimate h.

For each $k = 1, \dots, 2^{J-1}$, we have $\hat{\mu}_{2k-1} = s_k^{J-1}$ and $\hat{\sigma}_{2k-1}^2 = 2(d_k^{J-1})^2$, which leads us to our final regression model

$$2(d_k^{J-1})^2 = h(s_k^{J-1})^2 + \varepsilon_k. \tag{5.5.26}$$

In other words, we estimate h from the finest-scale Haar smooth and detail coefficients of $(X_i)_{i=1}^n$, where the smooth coefficients serve as pre-estimates of μ_i, and the squared detail coefficients serve as pre-estimates of σ_i^2.

The unknown h is restricted to be a non-decreasing function of μ and is estimated from the regression problem (5.5.26) via least-squares isotone regression, using the ‘pool-adjacent violators’ algorithm described in detail in Johnstone & Silverman (2005a). The resulting estimate, denoted here by $\hat{h}$, is a non-decreasing, piecewise constant function of μ.

The DDHFT is performed as with the HFT above, except that $\hat{h}$ replaces h.

Applications using the DDHFT can be seen in Motakis *et al.* (2006) where the DDHFT is used on microarray data and in Chapter 6 where the number of coalition casualties in Iraq is estimated. Modifications to allow negative data for both HFT and DDHFT are shown in Chapter 7, whereas in Chapter 8, we investigate a maximum likelihood approach to the Haar-Fisz transforms.

A further application of the HFT is described by Fryzlewicz *et al.* (2006), where the variance of a time series is estimated and used within a non-stationary model. Forecasting using the model is also discussed and empirical results show mixed performance, compared to other well known models.

5.5.6 Recent Work Generalising Fisz Variance Stabilising Transforms

Recently, Jansen (2006) has extended the ideas of the Haar-Fisz transformation for smoothing Poisson data. The transform can firstly be seen as an extension to the Haar-Fisz transform, modified so that in the variance stabilising step, any family of wavelet transforms can be used (with the notion that the coefficients are being *Gaussianised* instead of variance stabilised). Incorporated into the transform is a thresholding step: a level-dependent threshold is applied to the coefficients before being multiplied by the Gaussianising coefficient to return noise-free wavelet coefficients. The inverse-transformed coefficients are then considered an estimate of the Poisson intensities. A new Bayesian thresholding scheme, specifically for Poisson data is incorporated into the relevant step of the procedure.

In simulations, the transform was shown to have good performance, in terms of the mean square error compared to the Haar-Fisz and Anscombe transformations. The transformation, however, requires the data to be Poisson and hence can be viewed as a generalisation of the Haar-Fisz transformation. In the Haar sense it is similar to performing the HFT but after the wavelet coefficient are variance stabilised, they are thresholded to remove noise. The inverse stage of the transform thus produces denoised intensity estimates. The thresholding step could be replaced with a different method.

Fryzlewicz (2007) defines a ‘wavelet-Fisz’ transformation for Poisson data and a data-driven wavelet-Fisz transform for when the distribution of the data is unknown. These transforms either use the known mean-variance relationship, or an estimation via a Nadaraya-Watson estimator. A thresholding step is also suggested within the methodology which uses the estimated variance and local means of the data to set thresholding levels (for the decomposed wavelet coefficients).

Similar to Jansen (2006) the transform can be adapted for any family of wavelet, and when the distribution is known (and Poisson) and the Haar wavelet is used, the transformation reduces to the HFT. In contrast to Jansen (2006), the aim of this transformation is variance stabilisation, and the data-driven version of the transform allows a greater degree of flexibility in assumption for real data.

The transformation is shown to perform well at estimating the underlying intensity of signals drawn from the Blocks and Bumps signals *without knowing* the original noise distribution.

Chapter 6

Estimating the Intensity of Conflict in Iraq

6.1 Introduction

6.1.1 Background

This chapter addresses the question of estimating the true intensity of coalition conflict, in terms of coalition deaths, in Iraq since the current conflict began on 20th March 2003. The chapter is based on the paper by Nason & Bailey (2008).

Generally, a large proportion of statistical work is concerned with both accurately quantifying mortality and also the reasons and causes for such mortality, for example, epidemiological studies. The work in this chapter focuses deliberately on estimation of the true intensity but does *not* consider the rights and wrongs, or causes, of the conflict itself.

Our primary data set consists of the number of deaths of coalition personnel. The existence of such data raises extremely important questions for a variety of concerned parties including the military, the respective governments, the people of Iraq and people from coalition countries. For example: is the 'true' intensity increasing, decreasing or did it stay flat? Or, did the true intensity increase or decrease during certain different periods of political instability? As the Iraq Body Count website (`www.iraqbodycount.net`) points out: "Knowledge of war deaths must be available to all".

Why consider this problem? Since the conflict started in 2003 several websites have appeared with the laudable aim of tracking the number of deaths in the conflict (for example, `www.icasualties.org`, `www.iraqbodycount.net`). Some of these sites provide graphs showing the raw data but also estimates of the 'underlying mean'. Unfortunately, most of these estimates are not very good, primarily because they do not take into account the distributional properties of the data. In particular, these estimates do not take account of the fact that the variance of the data depends strongly on the mean. We later show that the number of deaths exhibits a clear non-decreasing *mean-variance relationship*. We have obtained our data from `www.icasualties.org`.

Our primary concern is to get good estimates of the underlying death intensity. We concentrate on the recorded number of deaths of coalition service personnel which are accurately recorded by the military and, hence, not subject to measurement error (although the record does not include those 'missing in action' but these numbers are extremely small: four unaccounted for up to 2nd July 2007, CNN Website (2007)). Also, although *every* death is one too many, the actual number of recorded coalition deaths per day is small in *statistical* terms. Hence, as with any low intensity count data, it is a statistical challenge to estimate the underlying intensity.

A related problem, which is of great concern and importance, is the number of non-coalition deaths stemming directly or indirectly from the conflict. We do not analyze these here because:

1. Controversially, non-coalition deaths are not officially recorded by coalition forces, see Roberts *et al.* (2004). So, at best, the number of such deaths are themselves estimated by external agencies, typically through media reports. These are subject to measurement error which requires a whole set of new techniques.

2. As also highlighted by Roberts *et al.* (2004) the number of non-coalition deaths is much higher than the number of coalition deaths. With high intensity count data a 'central limit theorem' behaviour sets in and these cases tend to be 'more Gaussian' and more standard mean estimates, as currently used by the media and websites, work reasonably well.

6.1.2 Methodology

The problem that prevents us from using simple methods to estimate the mean intensity of conflict is that these kinds of small count time series data often exhibit a non-trivial mean-variance relationship as described next.

Suppose the number of deaths during week t is denoted by $X_t \geq 0$. Let μ_t and σ_t^2 denote the (marginal) mean (or intensity) and variance of X_t. We claim here that the variance is some (non-decreasing) function of the mean. Mathematically, we write $\sigma_t^2 = h(\mu_t)$ for some h.

The classical example of such a setup also arose in a military context. von Bortkiewicz (1898) described data which counted the number of cavalrymen killed by horse or mule kicks in 13 corps of the Prussian army. This data is presented in modern form in Andrews & Herzberg (1985) and can be obtained online in the `vcd` package for the R statistical system. For the Prussian data the classical analysis assumes that the deaths, X_t, are distributed as Poisson random variables. In this case it is known that the mean equals the variance, $\sigma_t^2 = \mu_t$, and hence the non-decreasing function h turns out to be the identity function $h(\mu) = \mu$. This distributional form appears not to be the case for the Iraq data as later sections will demonstrate.

An effective approach for this kind of data is that of variable transformation. Let us denote the number of deaths per day of coalition forces (from all causes) by A_t. The idea is to find a transformation function of A_t which creates a new variable which has a relatively constant variance (that does not depend on the new variable's mean) and also with a marginal distribution closer to Gaussian. A popular and quick choice in this instance could be the $\log$ transformation, or maybe the square root transformation if one suspected Poisson data. A better choice might be the famous Box-Cox transformation due to Box & Cox (1964). This is described in detail in Section 5.5.2. Recall that the Box-Cox transform of variable X is given by

$$\{(X + \lambda_2)^{\lambda_1} - 1\}/\lambda_1 \tag{6.1.1}$$

for $\lambda_1 \neq 0$ and $\log(X + \lambda_2)$ if $\lambda_1 = 0$. The parameters λ_1, λ_2 can sometimes be selected by maximum likelihood methods. However, as we shall see in Section 6.2.3, the two-parameter

Box-Cox method does not always work well (in terms of bias, variance stabilisation and Gaussianisation), and sometimes it cannot even be calculated at all.

The main transformation method considered here is the *Data-Driven Haar-Fisz* transform (DDHFT) recently introduced by Fryzlewicz *et al.* (2007) and described in Section 5.5.4. This adopts a multiscale approach that has proven to be extremely effective. Recall that the method works by estimating the mean-variance relationship and then stabilizing the relevant time series at all scales and locations simultaneously. The operation of the DDHF transform is denoted by $\mathcal{F}_{\hat{h}}$ where the subscript denotes that the mean-variance function has been estimated by $\hat{h}$.

There is a large literature for the analysis of time series count data, for example, Winkelmann (2003) or see Jung *et al.* (2006) for a nice review and comparison of a whole series of techniques. We review several of these models in Section 5.4. Most of these methods address the separate issue of parameter estimation in models which often involve exogenous variates and/or other time-constant parameters (e.g. analogues of the constant parameter autoregressive processes). Our goal here is different in that we estimate the mean intensity which is inherently time-varying. The other point to recall from Section 5.4 is that much of the literature is concerned with Poisson-like response data (maybe with over- or under-dispersion) which exhibits serial correlation whereas our modelling demands fewer distributional assumptions and hence could be applied more widely. Additionally, as shall be demonstrated in Section 6.2.5, after appropriate mean correction the residual time series are not autocorrelated. Hence there is a real question over the appropriateness of many of the models for count data time series in the literature for *this* data set.

6.2 Analysis of Deaths from All Causes

6.2.1 The Data

The number of deaths per day (from all causes), A_t, $t = 1, \dots 1024$ from 12th June 2003 until 31st March 2006 is depicted in Figure 6.1 and a number of features are apparent. Overall, the number of deaths per day is usually less than or equal to 5. In fact, approximately 91% of days have 5 deaths or less. However, it can also be observed that there are periods

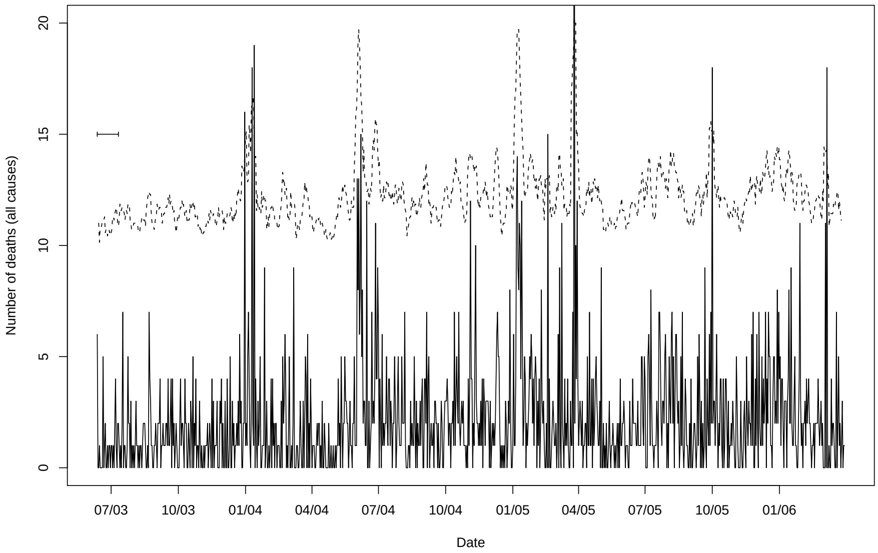


Figure 6.1: Number of deaths per day from all causes from 12th June 2003 until 31st March 2006 (solid line). A marker indicating the width of 30 days is plotted on the far left of the plot at level 15. Value of 'off-scale' observation is 37. The dashed line is a 7-day running mean (translated upwards by 10 so that it is is not obscured by the data).

where the number of deaths per day are higher, although there might not be a large number on each and every day during such periods. For example, the numbers of deaths on days 570 to 590 (1st Jan 2005 to 21st Jan 2005) were:

1, 6, 1, 1, 4, 12, 14, 9, 8, 11, 10, 4, 12, 6, 1, 2, 5, 4, 0, 2, 1

During this period the number of deaths per day was much higher than usual. However, the *variation* is also larger than during "quieter" periods. In other words the variation of the data is related to the mean level: the higher the mean the higher the variance or a non-decreasing mean-variance relationship. This phenomena can also be observed directly in Figure 6.1.

We assume that the mean intensity (and hence the variance) changes over the period of the conflict. However, we also implicitly assume that the change in mean intensity is not too fast. Although we admit sudden changes we do not assume continual rapid change (that is the mean could change suddenly, but *not* change rapidly day after day for a prolonged period). We believe these assumptions are realistic but we have not tested them in any formal statistical sense.

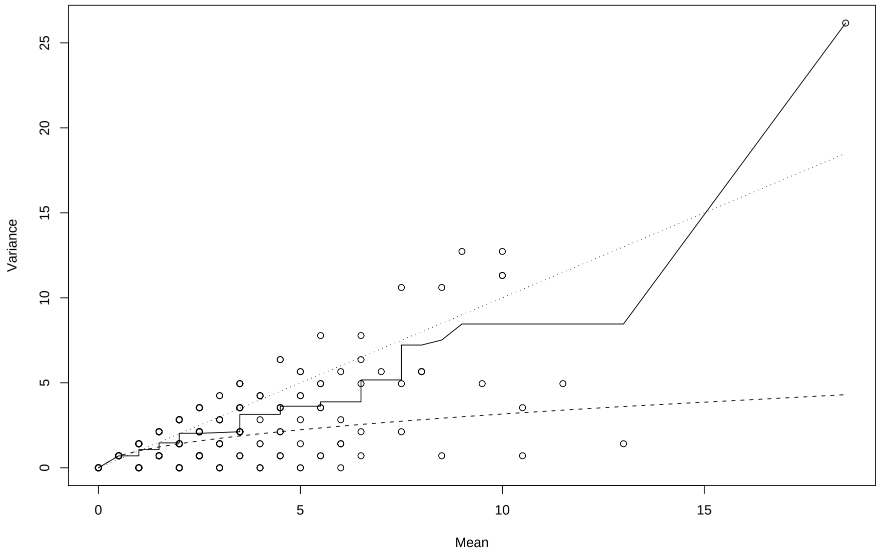


Figure 6.2: All causes. Small circles: plot of estimated local variance $\hat{\sigma}_t^2$ versus local mean $\hat{\mu}_t$. Solid line : estimated mean-variance relationship function $\hat{h}$ estimated using isotonic regression on the small circles. Dashed/dotted shows lines $y = \sqrt{x}$ and $y = x$ respectively.

6.2.2 Estimating the Mean-Variance Relationship with DDHFT

The first step of the DDHF transform estimates the mean-variance relationship, $\hat{h}$. Figure 6.2 shows the local standard deviations, $\hat{\sigma}_t$ plotted against estimated local means, $\hat{\mu}_t$ (as estimated by DDHFT) and also the best non-decreasing fit (isotonic regression) is plotted as a solid line.

It appears that the best non-decreasing fit lies mostly between the $y = \sqrt{x}$ (Poisson-like) and $y = x$ (χ^2-like) lines (i.e. $\mu_t \leq \hat{h}(\mu_t) \leq \mu_t^2$). Although the best-fit line does not coincide with $y = x$, it is much closer to it than the $y = \sqrt{x}$ line.

The prime objective of our method is to stabilise the variance of the transformed series (confirming that this is performed successfully is described in the next sections, particularly model-checking in Section 6.2.5). The exact nature of $\hat{h}$ is not of great importance here but it is an interesting by-product which gives a general idea of the mean-variance relationship. It would be interesting to study further the properties of $\hat{h}$, we discuss this in Section 6.6.

6.2.3 Estimating the True Intensity

Using DDHFT. The DDHFT takes A_t into a new series by applying the operator $\mathcal{F}_{\hat{h}}$ to obtain $a_t = \mathcal{F}_{\hat{h}} A_t$ and a_t is assumed to be well-modelled by

$$a_t = f_t + \epsilon_t, \tag{6.2.2}$$

where f_t is the transformed signal and ϵ_t is distributed as iid $N(0, \sigma_a^2)$. It turns out, as we shall verify later, this model for a_t is a very good one for our data.

Our primary aim is to obtain good intensity estimates. So rather than apply a single smoothing method to the DDHF-transformed data we applied several, three of which are listed in the Appendix B.1 (two wavelet shrinkage ones labelled S1 and S3, and one local kernel regression one labelled S2). After smoothing we apply the inverse DDHF transform to obtain an estimate in the original data domain. Figure 6.3 shows our 3 smoothing methods as applied to A_t. Roughly, all estimators show more or less the same, although there are some differences. From July 2003 until about January 2005 there has been a slow rise from about 1.8 deaths per day to 2.8 deaths per day and since then a decline and plateau at 2.5 deaths per day. The estimators are flexible enough to detect some sharp rises in deaths during January 2004 (large protests for direct elections), late June 2004 (power transferred from coalition to Iraqis) and smaller peaks centred around late Jan 2005 (Iraqi election 30th January 2005, also this period has single deadliest day for coalition since the war began), late Mar 2005 (Iraqi assembly meets for the first time), early Aug 2005 (Iraqi constitution drafted), early Oct 2005 (Iraqi voting on constitution) and late Dec 2005 (Parliamentary elections held).

The reader can make what they will of the intensity estimates. However, our *opinion* is that, although the mean intensity for all causes of death does oscillate, there is a trend upwards from the beginning of the series until about January 2005 and then the intensity levels off and then a slight decrease to another plateau at about June 2005.

Using Box-Cox. The one-parameter Box-Cox transformation, (6.1.1) with $\lambda_2 = 0$ cannot be used as, obviously, the number of deaths on a given day can be zero and $\log(0)$ is not defined. A popular recommendation in this case is to apply Box-Cox to, e.g. $1 + A_t$ but then

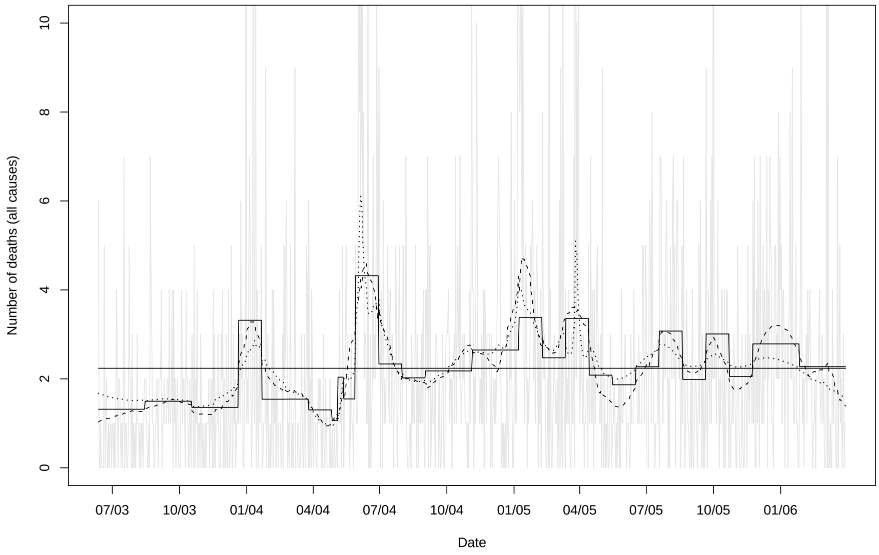


Figure 6.3: Several estimators of the mean of A_t. Method S1 (solid line), S2 (dotted), and S3 (dashed). Original A_t sequence is shown in grey. The mean of the whole A_t sequence is shown as a horizontal solid line.

one must ask the question why 1? This then leads onto use of the two-parameter Box-Cox transform (6.1.1). Unfortunately, for our data, and also in situations of this kind it is well-known that the likelihood is/can be unbounded and sensible parameter estimates are difficult to obtain, see (Atkinson, 1987, 9.3). So generally, we do not use Box-Cox here. The popular choice of $1 + A_t$ *was* tried but resulted in poorer variance stabilisation and Gaussianisation properties than the DDHFT as judged by Breusch-Pagan tests and Kolmogorov-Smirnov tests respectively. Additionally, one often pays a bias penalty when using transformation methods. Both Box-Cox and DDHFT methods incur a penalty but the bias associated with the DDHFT is, overall, dramatically less than with Box-Cox. See Appendix B.2 for some empirical bias calculations that demonstrate this good performance. Theorem 3 from Fryzlewicz (2007) shows that the DDHFT procedure, using a Nadaraya-Watson estimate of h, is asymptotically unbiased.

Running Means. Several websites use running means to generate estimates of true intensity. We described running means in Section 5.3.2. Recall that mathematically, the running mean at time t is given by:

$$r_t = b^{-1} \sum_{i=1}^{b} A_{t+i-\lfloor b/2 \rfloor}, \tag{6.2.3}$$

where b is the number of observations used to form each local mean, the bandwidth, and $\lfloor x \rfloor$ is the largest integer less than or equal to x. The $\lfloor b/2 \rfloor$ term in (6.2.3) causes the running mean r_t to be computed on a 'window' of observations centred on A_t of length b.

The dashed line in Figure 6.1 shows a 7-day running mean for the A_t time series. It is extremely variable compared to the estimates in Figure 6.3.

One problem with running means is deciding on how to choose the window width b. Most websites choose b too small which results in an extremely variable estimate and certainly of little use in estimating the underlying intensity. Another problem is that the b parameter is global and does not adapt to local signal characteristics. If we chose the window width well for one part of the series it would almost certainly be wrong for another part. In this respect the wavelet shrinkage and local kernel smoothing that we use are superior. A third, and serious problem for this kind of data which exhibits non-constant variance, is that the estimate itself is more variable in areas of high variability — this can be clearly seen in the dashed line in Figure 6.1 (and is also demonstrated in Section 5.3.2).

6.2.4 A Bootstrap Test for Variance Stabilisation

Before we proceed with model checking, we suggest a new bootstrap test for variance stabilisation. However, as will be shown, this test does not perform well compared to the Breusch & Pagan (1979) test for heteroskedascity.

Briefly, our test operates in the following way.

1. Select a sample of consecutive points from the signal A_t, centred on some random point t, with random length $L > 20$, L even.

2. Split the sample from 1. about its centre into two equally sized parts. Perform Mood's two-sample test for a difference in scale parameters (`mood.test` in R, see Conover (1971)), and record the p-value for this test.

3. Repeat 1. and 2. $n_{\text{search}} = 250$ times and take the median of the p-values. This is the

test statistic.

4. Repeat 3. Bsims $= 99$ times but each time on a different random permutation of the initial series, A_t.

5. Compare the test statistic in 3. with the bootstrap simulations in 4. to obtain an overall p-value for the test.

Essentially this test works as follows. If the variance of A_t, denoted by σ_t^2 is constant. then permuting the values will have no effect on the distribution of the test statistic calculated in 3. If the variance σ_t^2 changes over time then the p-values from Mood's test in 2. will be small. Likewise, the median p-value as computed in 3. will also be small. Thus, if the test statistic is large compared to the bootstrap simulation, we conclude that A_t has a non-constant variance.

Note that we assumed earlier that the mean intensity would not be subject to prolonged periods of rapid change. If we did not assume this then long periods of rapid intensity change could not be detected by our test.

Comparison with Breusch-Pagan Test

We generate data to test both our bootstrap test and the Breusch-Pagan test for heterogeneity. We first generate data from the Gaussian distribution of length 512 with zero mean and unit standard deviation. We then replace the last k points of the data with new values, again drawn from the Gaussian distribution, but with standard deviation σ^2. We vary k to take values 16, 32, 64, 128 and 256 and σ ranges from 1 to 3 in increments of 0.1. For each of these sequences, which are known to have a non-constant variance, we measure the power of the tests, that is, the number of times they correctly identify the variance as non-constant. For each value of k and σ we generate 100 sequences and take the mean of the size.

For each value k, we plot the power against the value of σ and compare the effectiveness of the two tests. Figure 6.4 shows the power for the tests for values $k = 32$ and $k = 64$ respectively. For smaller k, the bootstrap test has a poorer performance and for larger k (i.e. ≥ 64) the bootstrap compares more favourably. Further, increasing the variables within the test, such as the sample size L, the number of iterations N_{search} or replications

Bsims, are likely to improve the performance for smaller k. However, this will increase the computational time for the test. We therefore choose not to use the bootstrap test to analyse the effectiveness of the DDHF transform for the Iraq data and instead use the Breusch & Pagan (1979) test.

6.2.5 Model Checking

We now consider the statistical properties of A_t and the DDHF-transformed version a_t. Our hypothesis is that A_t is some uncorrelated sequence with a marginal distribution possessing the mean-variance relationship as estimated in the previous section.

Autocorrelation. Figure 6.5 shows several autocorrelation (acf) plots. The first, plot (a.) shows the acf of the original sequence A_t and plot (b.) shows the same for the DDHF-transformed sequence. There is some indication that the sequences might be autocorrelated but one must remember that we believe that the mean of each series is not constant (as it is this that we are trying to estimate). Autocorrelation figure 6.5.c shows the acf after subtracting the mean estimate S3 from a_t. It can be seen that after the varying mean has been taken into account the acf more or less disappears. Hence, we have some justification for assuming that $\{\epsilon_t\}$, in the model for the transformed data (6.2.2), is uncorrelated. Figure d. shows the acfs of the equivalent of c. but in the original data domain. The acf has almost entirely disappeared. Hence, once the local mean has been successfully estimated we have evidence that the sequence A_t is uncorrelated.

Constant variance. We tested the constant variance assumption using the Breusch & Pagan (1979) test. For S1, S2, and S3 the p-values are 0.9, 0.76 and 0.63. Hence there is no evidence for non-constant variance.

Gaussianity. We applied the Kolmogorov-Smirnov test to the residuals from each of the fits shown in Figure 6.3. The p-values of the residuals from S1, S2 and S3 are 0.009, 0.066 and 0.044 (here H_0 is the usual hypothesis that the samples are Gaussian with a given mean and variance, estimated here by their sample values). So for method S2 there is (formally) no evidence against Gaussianity. For S3 which is significant at the 5% level, but not the 1% level there may be weak evidence against Gaussianity. For S1 there is evidence of non-Gaussianity. All of these tests are sensitive to the mean removal. It must also be

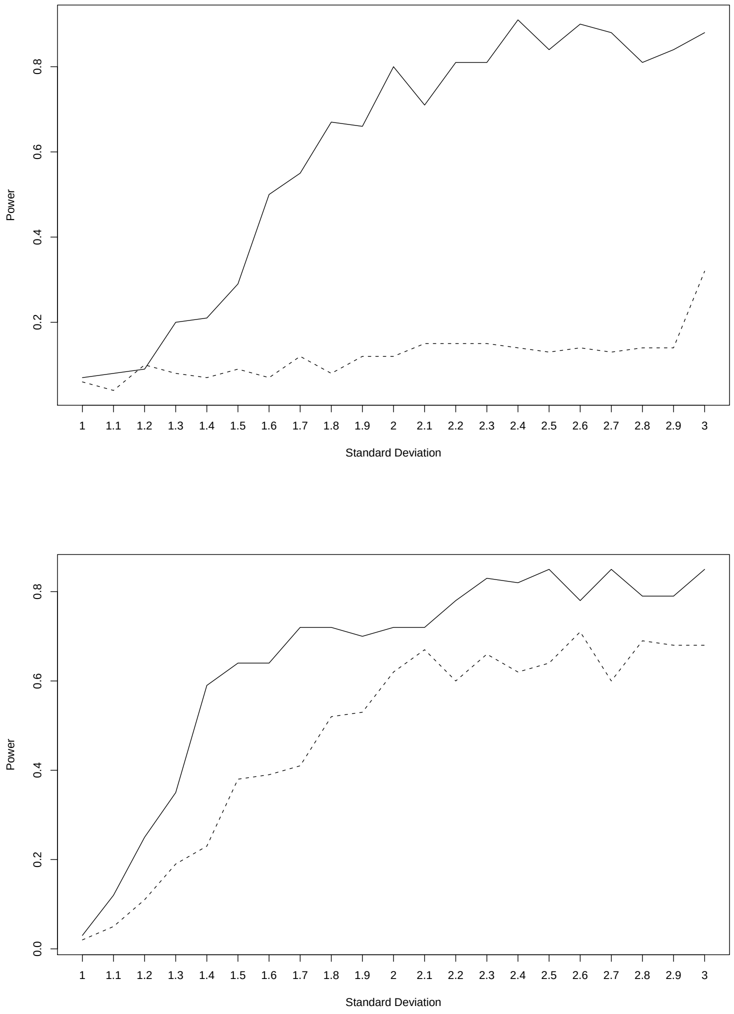


Figure 6.4: Power of the Breusch-Pagan (solid line) and the bootstrap tests (dashed line) for Gaussian data, mean 0 and standard deviation 1, with modified points of standard deviation σ. Top: 32 modified points. Bottom: 64 modified points

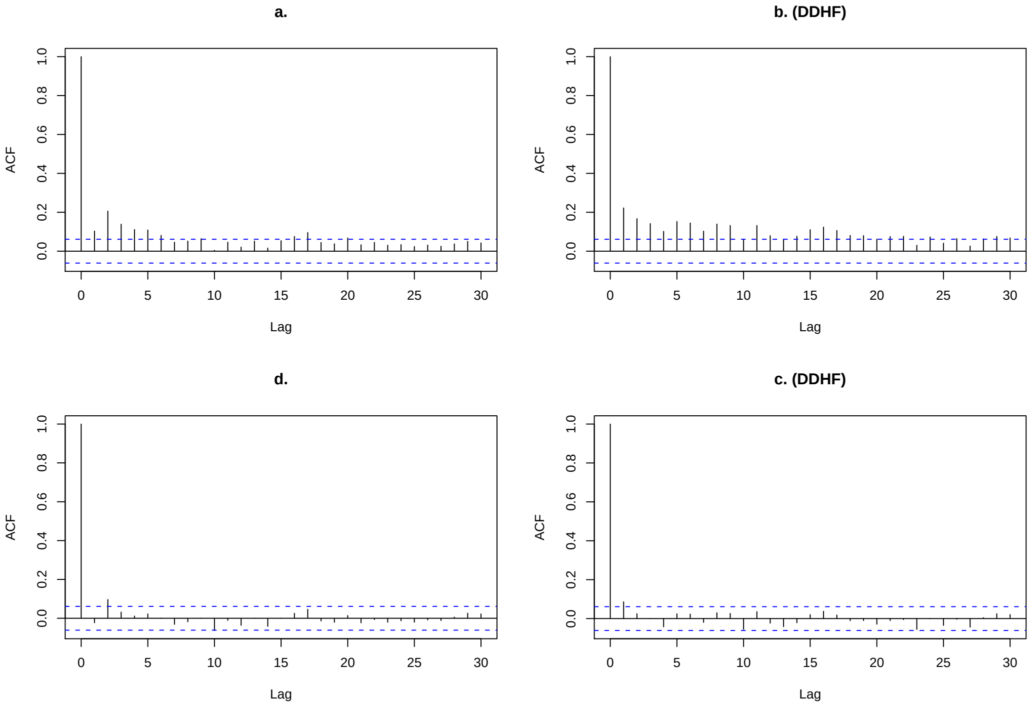


Figure 6.5: Clockwise from top left: autocorrelation functions of (a.) the number of deaths due to all causes, A_t; (b.) a_t, the DDHF transform of A_t; (c.) a_t minus signal estimate S3 (below); (d.) A_t minus signal estimate inverse DDHF-transformed S3.

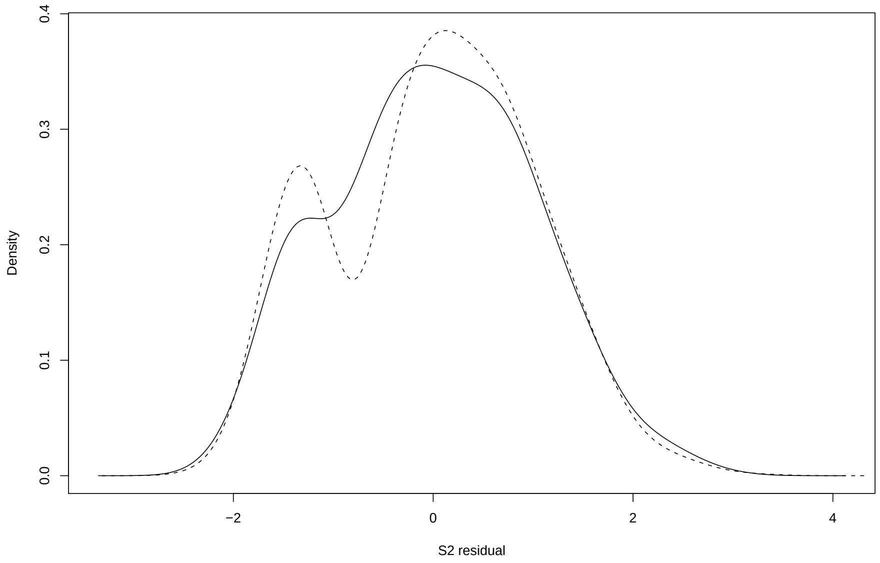


Figure 6.6: Density estimate of residuals from S1 fit. Solid: residuals using DDHFT. Dashed: residuals using Box-Cox.

remembered that the prime objective of variance stabilisation is to make variance constant and Gaussianisation is only a secondary effect (that is why the constant variance p-values above are so much better).

Having said that, the DDHF-transformed variates are much more Gaussian than those produced by the Box-Cox transforms that we tried. For example, see Figure 6.6, which shows the density estimates of residuals from the S1 fits using both DDHFT and Box-Cox transformation methods. The Box-Cox residuals are clearly bimodal. The DDHFT residuals have a ‘shoulder’ at about -1.5 but the density’s symmetry is better (so, roughly speaking, more Gaussian looking). Similar pictures were observed from the residuals of S2 and S3.

6.3 Analysis of Deaths from Hostile Actions

Figure 6.7 shows the number of deaths from hostile actions, which we denote by H_t, for the same date range as for the deaths due to all causes.

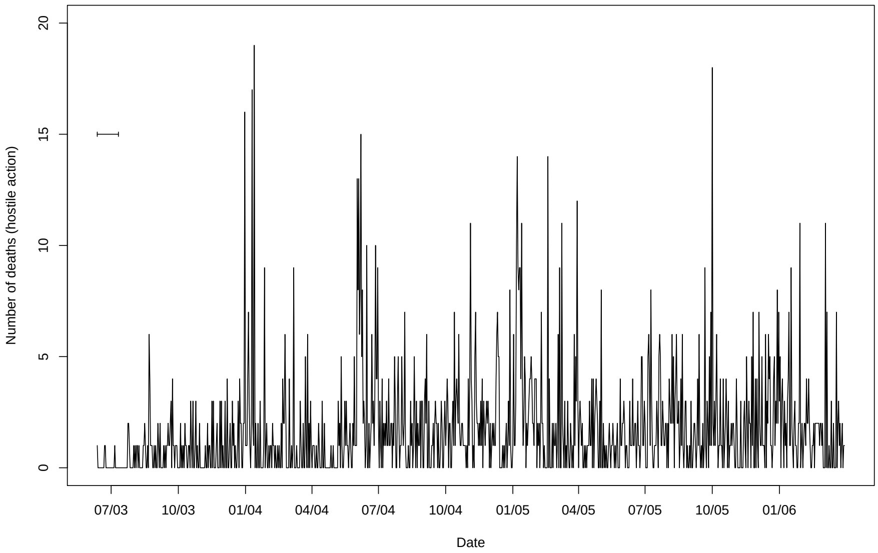


Figure 6.7: Number of deaths per day resulting from hostile action from 12th June 2003 until 31st March 2006. A marker indicating the width of 30 days is plotted on the far left of the plot at level 15.

6.3.1 Estimating the Mean-Variance Relationship

As in Section 6.2.2 the first stage of the DDHFT algorithm is to estimate the mean-variance relationship. Figure 6.8 again shows the estimated local standard deviations, $\hat{\sigma}_t$ plotted against estimated local means, $\hat{\mu}_t$ and also the best non-decreasing fit (isotonic regression) is plotted as a solid line. Once more the best non-decreasing fit is closer to the $y = x$ line.

6.3.2 Estimating the True Intensity

We again use the DDHFT and transform H_t to a sequence h_t. After smoothing using methods S1, S2 and S3 we obtain estimates as shown in Figure 6.9. The differences between Figures 6.3 and 6.9 show that the deaths in Feb/Mar 05 were largely due to non-hostile actions as the second peak around that time is missing from Figure 6.9. Referring back to the original records confirms that many non-hostile action deaths occurred around that time. Further analysis of the differences between non-hostile and hostile deaths is presented in Section 6.4.

The estimate of hostile death intensity shows a decline from Feb 2005 to July 2005

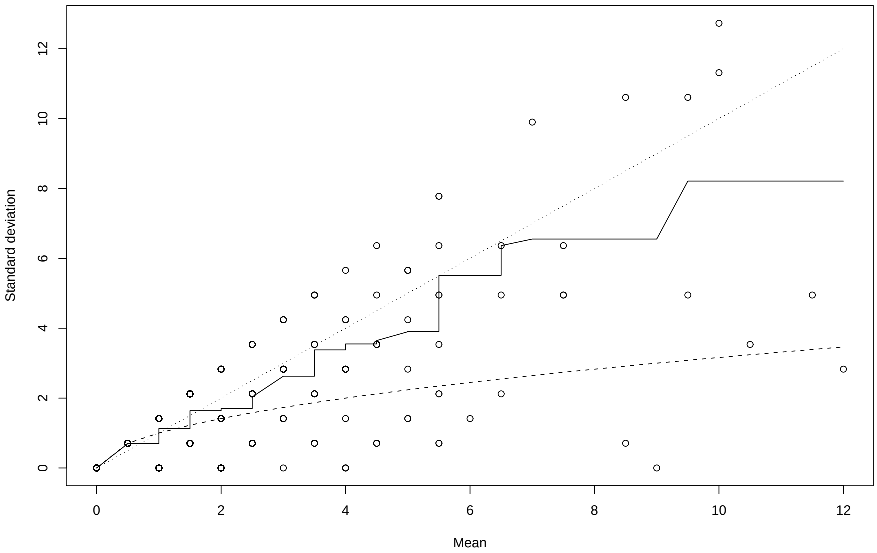


Figure 6.8: Hostile causes. Small circles: plot of estimated local standard deviation, $\hat{\sigma}_t$, versus local mean, $\hat{\mu}_t$. Solid line: estimated mean-variance relationship function $\hat{h}$ estimated using isotonic regression on small circles. Dashed/dotted shows lines $y = \sqrt{x}$ and $y = x$ respectively.

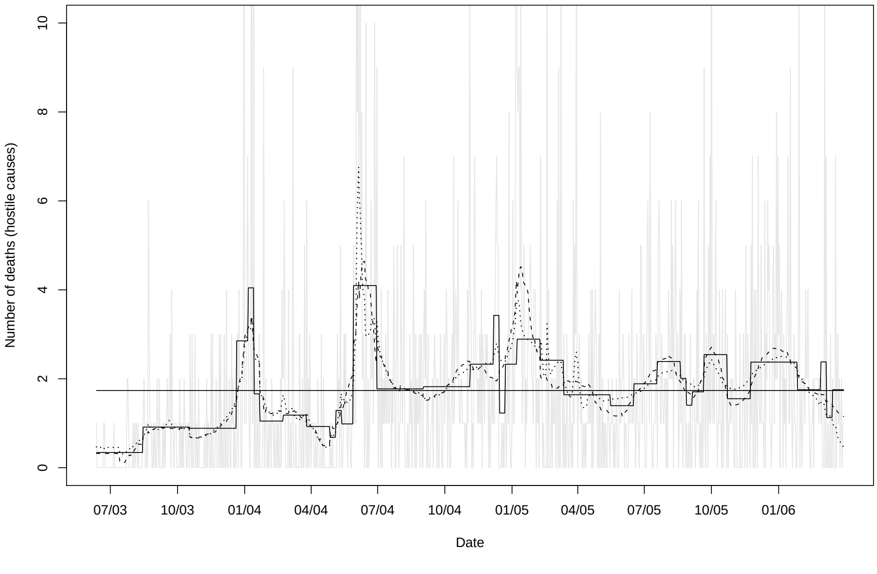


Figure 6.9: Several estimators of the mean of H_t. Methods S1 (solid line), S2 (dotted), and S3 (dashed). Original H_t sequence is shown in grey. The mean of the whole H_t sequence is shown as a horizontal solid line.

before another increase begins (apart from the extra peak, the “all deaths” series also shows a decline). During this period there was a increase in the number of terrorist attacks upon Iraqi citizens (Jan 2005: Iraqi elections held; Feb 28th 2005 saw the largest number of Iraqi deaths in a single incident; Apr 2005 saw selection of Iraqi President and Prime Minister “amid escalating violence”, BBC Website (2007)).

6.3.3 Model Checking

Similar model checking activities were performed for the deaths from hostile causes time series. The acf plots were not noticeably different from the patterns seen for “all deaths” shown in Figure 6.5.

For constancy of variance the Breusch-Pagan test indicated strong non-constancy of variance. On further examination we believe that this is due to the almost zero count at the very beginning of the series (around July 2003 in Figure 6.7). If we omit the first 30 observations in the Breusch-Pagan test then the p-values indicate no evidence for non-constancy of variance. (The p-values for S1, S2 and S3 are 0.08, 0.14 and 0.1 respectively).

For checking the Gaussian nature of residuals the Kolmogorov-Smirnov p-values for the residuals for S1, S2 and S3 were 0.004, 0.04 and 0.01, again, strictly non-Gaussian at the “5% level” but, for S2 and S3 at least not *too* non-Gaussian!

6.4 Differences due to Hostile and Non-Hostile Events

In this section we make use of the S2 kernel estimates for the intensity of deaths due to “all causes” and hostile causes (the results were similar when we used the S1 and S3 methods). Let us denote the mean intensity that we estimated in Section 6.2 for all causes by μ_t^A and the mean intensity that we estimated in Section 6.3 for hostile deaths by μ_t^H (these were plotted as lines S2 in both Figures 6.3 and 6.9 respectively). Figure 6.10 shows both estimates plotted on the same plot. Overall, on most occasions, a large “all causes” intensity is associated with a large “hostile” intensity. However, there are occasions when the “all causes” exceeds the “hostile” intensity. In particular, there is a big hump during March 2005 which appears to be due to entirely non-hostile causes.

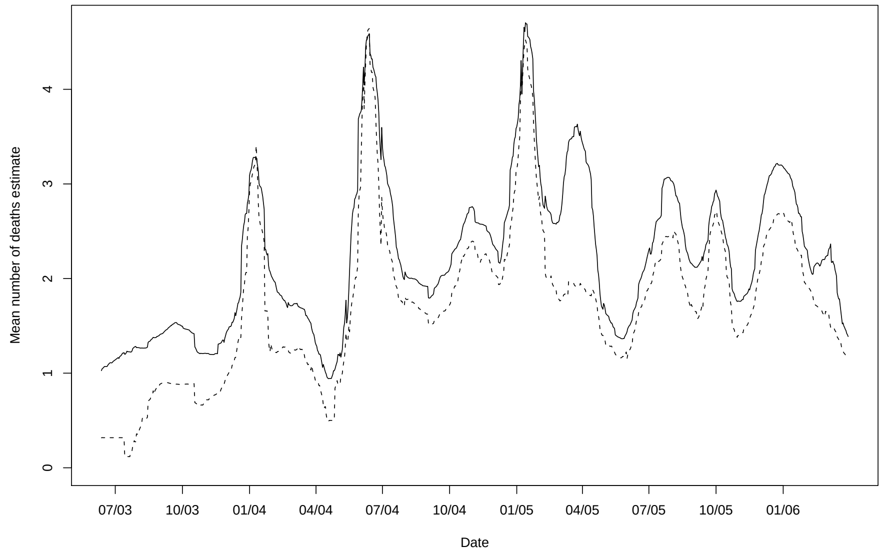


Figure 6.10: Time series plot of $\hat{\mu}_t^A$ (solid line) and $\hat{\mu}_t^H$ (dashed).

Figure 6.10 raises the following question: is the intensity of non-hostile deaths related to the intensity of hostile deaths? For example, one might think of several potential hypotheses: more ‘accidental’ deaths occur during period of increased hostile stress? Or more ‘accidental’ deaths occur when there is less concern about the hostile threat. Or some other relationship might hold.

A plot of the numbers of deaths (the *data*), $A_t - H_t$ versus H_t does not reveal much due to the noise in these processes. Figure 6.11 shows a plot of $\hat{\mu}_t^A - \hat{\mu}_t^H$ versus $\hat{\mu}_t^H$ for each time point. There is evidence of a slight negative correlation (in fact, numerically the correlation is -0.24). The dots in the bottom right hand of the plot are due to three separate periods in time and the spikes to the top and extreme top-left are both individual and separate periods. Hence, the tentative conclusion is that fewer accidental deaths occur when the hostile threat is greater. Further, it must be the case that the smaller number of accidental deaths is not just because larger *numbers* of coalition forces are involved in battle as the numbers involved in these skirmishes are relatively small. We propose that some other less direct mechanism is at work. For example, it could be that in times of known higher hostile threat that people are more vigilant and less subject to non-hostile action deaths.

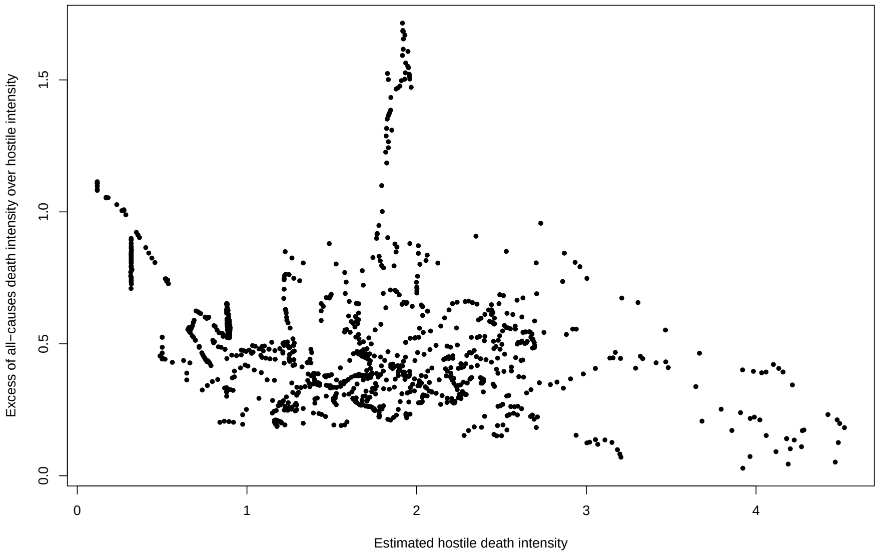


Figure 6.11: Plot of $\hat{\mu}_t^A - \hat{\mu}_t^H$ versus $\hat{\mu}_t^H$. (Note that six values where $\hat{\mu}_t^A - \hat{\mu}_t^H$ are slightly negative are omitted. The negative values to two decimal places are $-0.12, -0.06, -0.06, -0.05, -0.05, -0.03$.)

6.5 Recent Work

Recently, Spirling (2007) suggested using reversible jump Markov chain Monte Carol (RJMCMC) techniques to investigate possible 'jumps' in the number civilian casualties, where the overall rates of attack appear to change. This technique has the advantage that estimation of the actual number of deaths is not required so they are not faced with the problems associated with such a measure. The paper thus considers the frequency of attacks, rather than their actual size and uses data from `www.iraqbodycount.org` to obtain a *minimum* possible death toll (and avoid any possibility of over-counting).

The work attempts to answer the following questions: How many change points occurred? If this number was known, *when* did they occur? And finally, if change points and dates were known, what were the effects? The number of points is assumed to be four and this is used to discover when the jumps occurred. A comparison with events in Iraq, as well the deaths per day before and after each 'jump' are used to suggest possible effects of these changes.

The author gives a 90% highest posterior density (HPD) of the range of dates at which the jumps occur. This range for the first two jumps cover a period of over 9 months each. Also presented is a plot of the change in rates of casualty incidence over time. The large range in dates when the jumps occured, plus a fairly flat rate of change in casualty incidents suggests the notion of four jumps is perhaps unwise. This first jump is stated as occuring on 26 January 2004. During this time, our data shows a large peak at a time of protest for direct election. It is feasible that the two data sets are linked and we suggest that their ‘jump’ point during this time is caused by a *peak* in the intensity, rather than an overall increase.

Regardless of the interpretations of their results, the work presents an interesting alternative to estimating the actual number of deaths. Similar to work in this chapter, trends and real life events are used to analyse results. We briefly suggest extensions of their methods, in relation to coalition deaths, in the next section.

6.6 Some Interpretations and Next Steps

In this chapter we have proposed an analysis of the number of deaths of coalition personnel due to both “all causes” and hostile action during the current Iraq conflict. Our main aim was to supply good estimates of the mean intensity for both of these time series and to improve on the highly variable estimates presented on various websites computed using simple running means. As described in the text above, although oscillatory, the mean level of the conflict intensity increased until about January 2005, then leveled off until about June 2005 and then underwent a slight decrease and a further leveling off until the end of the series.

We also showed that, for both these data sets, the marginal variance of the series is approximately equal to the square of the mean of the series. This is in contrast to the classical ‘Poisson’ military example due to von Bortkiewicz (1898) and exhibits a greater degree of variability at higher intensity levels.

Another, more tentative, conclusion is that the intensity of non-hostile deaths is inversely related to the intensity of hostile deaths. However, this conclusion should be subjected to further scrutiny.

Further technical observations are that the DDHFT method stabilises (and Gaussianises) the data well and better than the Box-Cox method with a ‘popular’ parameter of $\lambda_2 = 1$ for data that contain zeroes. It was not possible to find good parameter values for the two-parameter Box-Cox transformation as the likelihood was unbounded. We also tried AVAS due to Tibshirani (1988) but did not get good results and so we do not report them here (using the `avas()` function from the package `acepack()` in R).

There are many avenues for further investigation in this area. It would be interesting to identify and study the theoretical properties of $\hat{h}$ both to enable the construction of confidence intervals and also to understand the robustness of the procedure and implications for the subsequent intensity estimation problem. This article makes use of isotonic regression which itself confers a degree of robustness and localisation when compared to, e.g. parametric regressions. It is also important to note that for discrete data the number of repeated points at a given μ_t location is usually relatively high so outlier identification is often easier when compared to the common situation of one observation at each μ_t.

Another possibility would be to obtain forecasts of the future behaviour of the time series, both of the future mean intensity and also its first derivative (to discover whether the conflict was improving or deteriorating).

One might also wonder whether it would be more worthwhile to study death rates rather than the absolute numbers. In particular, areas with fewer troops, more insurgents, or different operational policies might influence death rates significantly. One problem is that acquiring such data is extremely difficult.

The RJMCMC method of Spirling (2007) could be applied to the coalition death toll to pick out points at which the intensity increases. Although estimation of the numbers is not an issue with our analysis, the results would be of interest and similar areas of jumps within both datasets could provide links between coalition and civilian deaths.

Chapter 7

Haar-Fisz Transforms for Negative Data

7.1 Introduction

Chapter 6 showed how the data-driven Haar-Fisz transform can be used to stabilise variance and Gaussianise count data, where the mean-variance relationship is estimated from the data. The Box-Cox transform was also considered for the task, but not used due to problems of choosing the transformation parameters and often (as was the case for the Iraq data), the maximum likelihood can be unbounded.

In this chapter we introduce the central England temperature anomaly (CET) data set; an annual temperature record of various locations in the UK. Its records date back as far as 1772 and is considered to be an accurate measurement of annual temperature. As such, it is used to analyse the trend in temperature change by estimating an underlying intensity for the data. We demonstrate how existing methods to estimate trend using a running means estimator is not suitable as it does not take into account any underlying mean-variance relationship within the data.

We investigate this mean-variance relationship and propose using a variance stabilising transform before intensity estimation. We highlight the unbounded nature of the likelihood equations when using the Box-Cox transform to stabilise the data, as well as the limitations of the data-driven Haar-Fisz transform (DDHFT).

We then outline modifications to the DDHFT so that it can be used to stabilise the variance of the CET data. We create different transforms for when the mean-variance relationship is known and unknown, and apply the latter to the CET data to stabilise its variance. We can then use estimation techniques more suitable for the variance stabilised data to obtain new estimates of the underlying trend in temperature change.

7.2 Central England Temperature Data 1772–2006

The central England temperature (CET) data set is the longest instrumental record in the world and consists of temperature measurements taken from a roughly triangular area of England extending from the Lancashire plains in the north, to London in the southeast and Herefordshire in the southwest. It is described in detail by Parker *et al.* (1992) (which is based on work by Manley (1953)). The CET data is taken from a succession of observing sites and has been adjusted to remove heterogeneities between data sets, caused by changes in exact location and methods over time. The final data set, which is commonly scaled to be relative to 1961–1990, is referred to as 'anomalies'.

The CET data has been used in many climatological studies. Regularly updated plots of the data are published online as part of the Met Office Hadley Centre observation datasets at `http://hadobs.metoffice.com/hadset/`. The raw data, which includes daily, monthly and seasonal observation values can also be downloaded from the website. Figure 7.1 shows the annual mean 'anomalies' from the CET data set from 1772–2006, plotted in grey.

A much used smoothing technique for meteorological data is binomial filtering, as described in Aubury & Luk (1995) (and detailed in Section 5.3.2). The 21-point binomial filter has been used to smooth the CET data and is published on both the online plots and in various other publications. This binomial filter, applied to the CET data can be seen as the black line in Figure 7.1. The filter takes a weighted mean of a 'window' of data which surrounds each point. Questions arise over the values of the weights and the size of the window to use, and certain choices of the global parameters may result in poor performance over some areas of the data. A further problem, as with the Iraq data of Chapter 6, is that

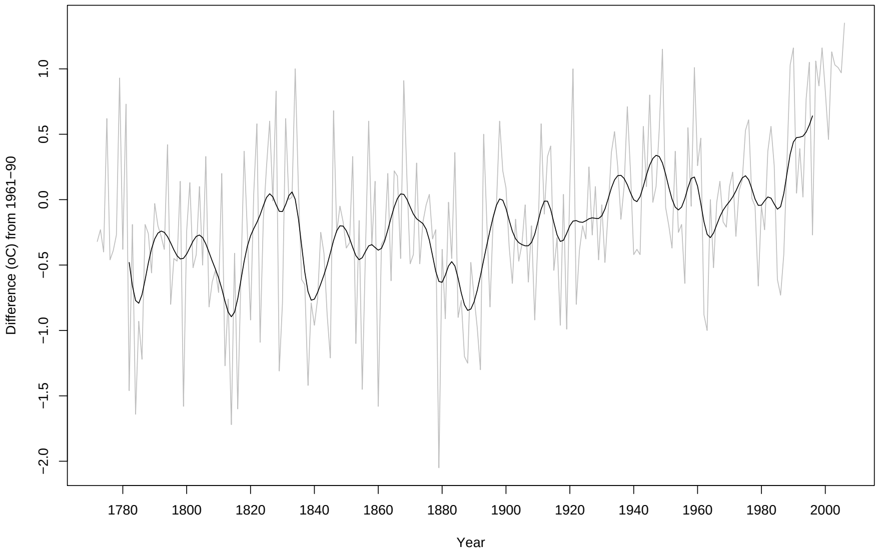


Figure 7.1: Central England Temperature 1772-2006 (grey), with 21-point binomial filter (black).

the estimate itself is more variable in areas where the data has high variability, as mentioned in Section 5.3.3.

Many smoothing techniques assume a level of Gaussianity within the data and thus make assumptions about the independence of a mean-variance relationship of the data. We next investigate a putative mean-variance relationship for the CET data more detail.

7.2.1 Analysis of Mean-Variance Relationship

We applied the technique of Section 5.5.5 to estimate the unknown mean-variance relationship (h) of the CET data (using finest level Haar wavelet coefficients to produce 'pilot' estimates of the mean and variance). The estimates are plotted as small circles in Figure 7.2. For negative means, the correlation between mean and variance is -0.062. For positive means this value is 0.40 and when the largest 5 mean values are omitted, this rises to 0.54.

We further investigate the complexity of the mean-variance relationship found using Haar coefficients as estimates of the mean and variance. We use a smoothing spline (using the function `smooth.spline` in R) as a basis of this further investigation. For the mean, we smooth our data using a cubic spline and take the smoothed value to be a local estimate

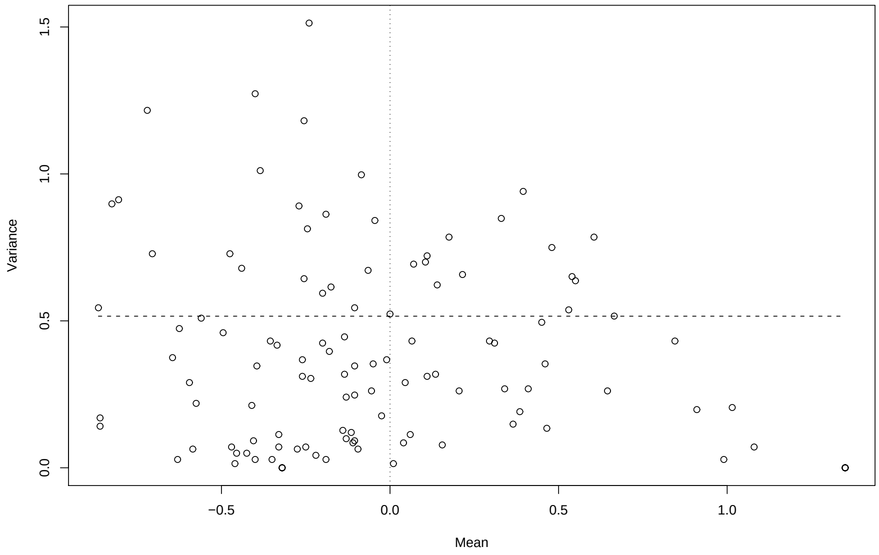


Figure 7.2: Central England Temperature data. Plot of pilot variance, $\hat{\sigma}_i^2$, versus pilot mean, $\hat{\mu_i}$. Dashed line: estimation of mean-variance function using DDHFT. Dotted line: mean = 0

of the mean. An estimate of the local variance is achieved by first taking the logarithm of the squares of the data. These are then smoothed using splines and the logarithm inverted to give the local variance approximation.

The plot of estimated mean and variance is shown in Figure 7.3. The spline estimation clearly shows a decreasing mean-variance relationship for negative means (with a -0.67 correlation). For positive means, (in particular those greater then 0.5), there is a much smaller amount of data and there appears to be a positive mean-variance relationship (with a 0.86 correlation for means greater then zero). This, however, may be caused by a boundary effect of the spline estimate. The plot also suggests that any change from a negative to a non-negative mean-variance relationship might indeed occur at a point greater than zero. As the data is limited, we can not ascertain from either estimates displayed in figures 7.2 and 7.3 the nature of the mean-variance relationship for positive means. We leave and additional investigation of this as future work and assume the transition occurs at mean zero.

Before smoothing the data, we wish to first transform the data in order to stabilise the variance which we have shown seems to be dependent on the mean. We attempt this with both the Box-Cox transform, and the data-driven Haar-Fisz transform to highlight problems

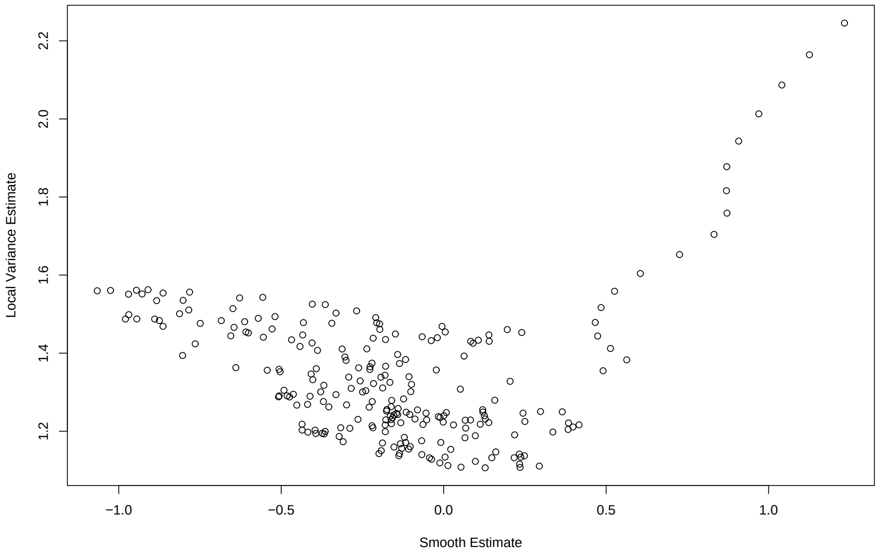


Figure 7.3: Spline estimate of local mean and local variance.

which arise in their application and to motivate modifications of the DDHFT.

7.3 Existing Variance Stabilising Transformations of the CET Data

Box-Cox Transform. To highlight problems with the issue of choosing suitable parameters for the Box-Cox transform, we show results from applying maximum likelihood estimation Box-Cox techniques to the CET data set. As we have negative data, we have to use the two parameter Box-Cox transform. As detailed in Section 5.5.2, we simplify the calculations by Gaussianising the transformation. Maximising the likelihood function becomes equivalent to minimising the the residual sum of squares $R(\lambda_1, \lambda_2)$ as defined by (5.5.22).

Atkinson (1987) describes how plots of $R(\lambda_1, \lambda_2)$ are not sensitive to the behaviour as $\lambda_2 \rightarrow$ -y_{min}, so we instead work in the scale defined by

$$\lambda_2 = -y_{\text{min}}(1 - 10^{\varepsilon}). \tag{7.3.1}$$

When $\varepsilon = 0$, $10^{\varepsilon} = 1$ and $\lambda_2 = 0$. For $\varepsilon > 0$, $\lambda_2 > 0$ and $\varepsilon < 0$, $\lambda_2 < 0$. Furthermore, as

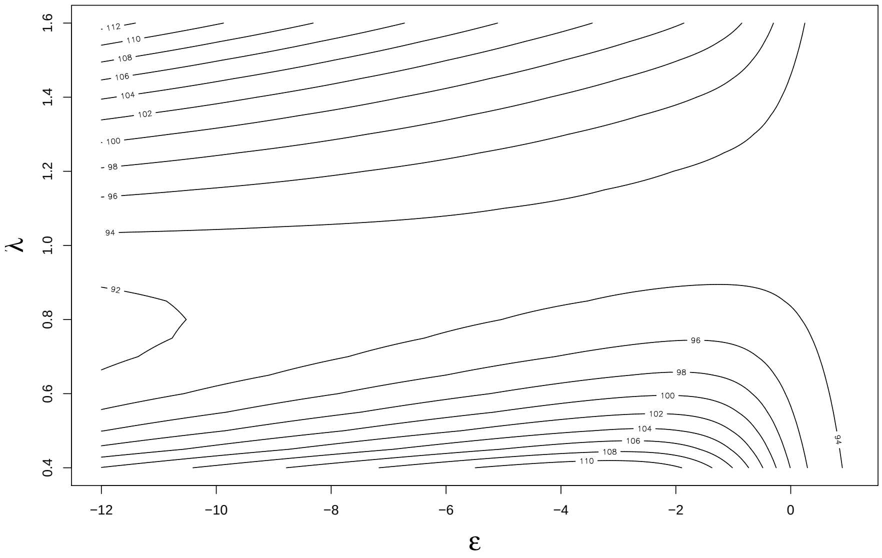


Figure 7.4: Residual sums of squares, $R(\lambda_1, \varepsilon)$ for the CET data. ε range is equivalent to values of λ_2 between -2047.95 and 2.05.

$\varepsilon \to -\infty$, $\lambda_2 \to$ -y_{min}.

Note that ε can take any real value and that for ε greater than zero, positive values of λ_2 are obtained. Also, .

Optimal parameters are thus found by minimising the reparameterised (Gaussianised) residual sum of squares $R(\lambda_1, \varepsilon)$ of the transformed CET data. The contour plot of $R(\lambda_1, \varepsilon)$ is shown in Figure 7.4. For the results plotted, $\varepsilon \in (-12, 2)$ which corresponds to values of λ_2 between -2047.95 and 2.05 ($-y_{\text{min}}$).

The contour plot shows the unbounded behaviour of $R(\lambda_1, \lambda_2)$. It indicates that $R(\lambda_1, \lambda_2)$ approaches a minimum as $\lambda_2 \to -y_{\text{min}}$. We showed at the end of Section 5.5.2 how $R(\lambda_1, \lambda_2)$ will get smaller as $\lambda_2 \to -\text{y}_{\text{min}}$ and thus for the CET data, no clear transformation parameters can be found.

Data-Driven Haar-Fisz Transform. We next use the data-driven Haar-Fisz transform (with h unknown), as described in Section 5.5.4, to transform the data and attempt to stabilise the variance. Figure 7.2 showed the pilot mean-variance estimates of the CET data from the DDHF transform. The estimate of the mean-variance function $\hat{h}$, using isotone regression (see Johnstone & Silverman (2005b)), is shown as a dashed line and does not fit the data well as the basic DDHFT assumes that the mean-variance relationship is strictly non-decreasing.

Negative mean values and an apparent decreasing mean-variance relationship cause problems in our current estimation of $\hat{h}$. Scaling the data by -1 could ensure a positive mean-variance relationship (if indeed the relationship is decreasing throughout the entire data) and a constant could be added to the data to ensure positivity. These would both add further parameters and questions could be raised as to how these are appropriately selected. Regardless of these issues, if the mean-variance relationship is not monotonic, as is the case with the CET data, the DDHFT as it stands is not suitable.

To cope with the mean-variance behaviour exhibited by the CET data, we next propose modifications to the DDHFT.

7.4 Modifications to Haar-Fisz Transforms for Real Data

The Haar-Fisz transform ($h()$ known) and the data-driven Haar-Fisz transform ($h()$ unknown) are described in Section 5.5.4. Both require positive data and assume the variance σ_i^2 to be a non-decreasing function of the mean μ_i. We next detail modifications to both transformations so that they are suitable for use with both positive and negative data.

7.4.1 The Negative Haar-Fisz Transform for Poisson Data

Here we outline our methodology which is a modification to the Haar-Fisz (HF) algorithm, for Poisson data, as described in Fryzlewicz & Nason (2004). We model our data y_i to be such that

$$P(Y=y)=\begin{cases}\dfrac{\lambda^{|y|}e^{-\lambda}}{2|y|!}, & \text{if } y\in\mathbb{Z}\setminus\{0\},\\ e^{-\lambda}, & \text{if } y=0\end{cases} \tag{7.4.2}$$

for $\lambda \in 0, 1, \ldots$. Therefore the probability of y_i taking a negative value, say $-y$, is equal to the probability of it taking the positive value y.

The Haar-Fisz transform for such real-valued (potentially negative) data denoted by the vector $\mathbf{v} = (v_0, v_1 \ldots, v_{N-1})$ for $N = 2^J$ where $v_i \in \mathbb{R}$ for all i, is defined as follows.

1. Let $s_i^J = v_i$ for $i = 1, \ldots, n$.

2. For each $j = J-1, J-2, \ldots, 0$, recursively form the vectors $\mathbf{s}^j$, $\mathbf{d}^j$ and $\mathbf{t}^j$:

$$s_k^j = \frac{s_{2k-1}^{j+1} + s_{2k}^{j+1}}{2};\ d_k^j = \frac{s_{2k-1}^{j+1} - s_{2k}^{j+1}}{2};\ t_k^j = \frac{|s_{2k-1}^{j+1}| + |s_{2k}^{j+1}|}{2}, \qquad (7.4.3)$$

   and immediately define $\mathbf{f}^j$ by:

$$f_k^j = \begin{cases} 0 & \text{if } t_k^j = 0, \\ d_k^j / \sqrt{t_k^j} & \text{otherwise}, \end{cases} \qquad (7.4.4)$$

   for $k = 1, \ldots, 2^j$, noting that if $s_{2k-1}^{j+1}, s_{2k}^{j+1} \geq 0$, then $t_k^j = s_k^j$ in (7.4.3).

3. For each $j = 0, 1, \ldots, J-1$, recursively modify s^{j+1}:

$$s_{2k-1}^{j+1} = s_k^j + f_k^j;\ s_{2k}^{j+1} = s_k^j - f_k^j, \qquad (7.4.5)$$

   for $k = 1, \ldots, 2^j$, and store the vector $\mathbf{t}^j$ for use in the inverse of the transformation.

4. Set $\mathbf{u} = \mathbf{s}^J$.

This procedure differs from the existing HF methods by defining the coefficients $\mathbf{t}^j$. This is the local mean of the magnitudes of our data, which we use to stabilise the variance. For strictly positive data, $\mathbf{t}^j$ is the same as $\mathbf{s}^j$ (which is then also equal to the *magnitude* of the mean).

Taking absolute values removes information about the sign of the data — information which is needed when inverting the transform (specifically when inverting (5.5.25)). We next give a numerical example to illustrate these problems.

7.4.2 Numerical Exposition

For clarification of the above procedure, we consider a simplified example with only two data points. These are general points from a vector of data and could be from any level of the recursive transformation.

Let the values of these two data points be

$$s^{j+1}_{2k-1} = -2 \qquad \text{and} \qquad s^{j+1}_{2k} = 5.$$

Then from (7.4.3) we have $s^j_k = 1.5$, $d^j_k = -3.5$ and $t^j_k = 3.5$. From 7.4.4 we get $f^j_k = -\sqrt{7/2}$.

We then form the Haar-Fisz transformed variables using (7.4.5)to get

$$s^{j+1}_{2k-1} = \frac{3}{2} - \sqrt{\frac{7}{2}} \approx 0.40, \qquad s^{j+1}_{2k} = \frac{3}{2} + \sqrt{\frac{7}{2}} \approx 3.37.$$

In the context of a larger data set, we would expect the data to have a stabilised variance and to be more ‘Gaussian’.

Common practice would be to smooth this transformed data and then to invert the smoothed values to get back to the original data domain. Just to ‘see what happens’, we intuitively follow a similar method to the inverse Haar-Fisz transform to invert the data.

Say, for example, we used a smoothing technique to obtain estimates of the underlying intensity of our transformed data as

$$\hat{s}^{j+1}_{2k-1} = 0.5 \qquad \text{and} \qquad \hat{s}^{j+1}_{2k} = 3.5. \tag{7.4.6}$$

We then wish to invert the data back to the original data domain so we retrace our steps backwards using the values in (7.4.6) to first produce

$$\hat{s}^j_k = \frac{\hat{s}^{j+1}_{2k-1} + \hat{s}^{j+1}_{2k}}{2} = 2, \qquad \hat{f}^j_k = \frac{\hat{s}^{j+1}_{2k-1} - \hat{s}^{j+1}_{2k}}{2} = -1.5,$$

which is undoing the effect of (7.4.5).

We next undo the effect of (7.4.4) (the ‘Fisz’ step), to obtain the detail coefficients, defined by

$$\hat{d}^j_k = \hat{f}^j_k \sqrt{t^j_k}. \tag{7.4.7}$$

This, however, raises the interesting question of which value of t^j_k to use. In the above, we imply that the original value (from the ‘forward’ transform in (7.4.3)) is used. This is intuitive as we wish to get back to the original data domain, which this value is associated with. However, since smoothing the data, this value is no longer specific to the values of $\hat{s}^{j+1}_{2k-1}$ and $\hat{s}^{j+1}_{2k}$ which we are using. Should we then use the updated value, $\hat{t}^j_k$ defined by

$$\hat{t}^j_k = \frac{|\hat{s}^{j+1}_{2k-1}| + |\hat{s}^{j+1}_{2k}|}{2} = 2?$$

Using either of these values for t would not guarantee that this step is invertible. We return to the numerical example to explain this point further.

Using the original value of t^j_k, which we now denote using the further subscript $t^j_{k_1}$, we obtain

$$\hat{d}^j_k = -1.5\sqrt{3.5},$$

from (7.4.7). We then undo the operation in 7.4.3 to obtain our smoothed estimate of the original data:

$$\begin{array}{rclclcr} \hat{s}^{j+1}_{2k-1} & = & \hat{s}^j_k + \hat{d}^j_k & = & 2 - 1.5\sqrt{3.5} & \approx & -0.806, \\ \hat{s}^{j+1}_{2k} & = & \hat{s}^j_k - \hat{d}^j_k & = & 2 + 1.5\sqrt{3.5} & \approx & 4.806. \end{array} \tag{7.4.8}$$

If we were to produce a new value of t from this data, denoted by $t^j_{k_2}$, we would find that $t^j_{k_2} \approx 2.806$. We would like these values of $t^j_{k_l}$ (for $l = 1, 2, \dots$) to be the same, so that we can 'undo' any steps taken. We thus put our value of $t^j_{k_2}$ back into (7.4.7) and calculate subsequent updates, $t^j_{k_3}, t^j_{k_4}, \dots$.

This process can be repeated until $t^j_{k_l} = t^j_{k_{l-1}}$, for some l, where l is the number of update iterations. It can be shown that using this iterative update procedure, $t^j_{k_l}$ converges (as $l \to \infty$) and takes the values

$$t^j_k = \begin{cases} |\hat{s}^j_k|, & \text{if } (\hat{f}^j_k)^2 \leq |\hat{s}^j_k|, \\ |\hat{f}^j_k \sqrt{t^j_k}|, & \text{if } (\hat{f}^j_k)^2 \geq |\hat{s}^j_k|. \end{cases} \tag{7.4.9}$$

Proof of this convergence is given in Appendix C. It is interesting to note that this result is a generalisation of the Haar-Fisz transform for positive data, as we always have that $(\hat{f}^j_k)^2 \leq |\hat{s}^j_k|$.

7.4.3 Inverse Transformation

Our inverse of the transformation can thus be summarised as follows:

1. Apply the Haar DWT to $\hat{\mathbf{u}}$ to produce $(\hat{\mathbf{s}}^J, \hat{\mathbf{f}}^J, \hat{\mathbf{f}}^{J-1}, \dots, \hat{\mathbf{f}}^1)$.

2. Define:
$$\hat{t}^j_k = \begin{cases} |\hat{s}^j_k|, & \text{if } (\hat{f}^j_k)^2 \leq |\hat{s}^j_k|, \\ |\hat{f}^j_k \sqrt{t^j_k}|, & \text{if } (\hat{f}^j_k)^2 \geq |\hat{s}^j_k|. \end{cases}$$

3. Apply the inverse Haar DWT to $\hat{\mathbf{s}}^j$ and $\hat{\mathbf{f}}^j$, undoing the effect of (7.4.4) as each scale is produced to give:
$$\hat{s}^{j+1}_{2k-1} = \hat{s}^j_k + \hat{f}^j_k \sqrt{t^j_k}, \tag{7.4.10}$$
and
$$\hat{s}^{j+1}_{2k} = \hat{s}^j_k - \hat{f}^j_k \sqrt{t^j_k}. \tag{7.4.11}$$

4. Set $\mathbf{v} = \mathbf{s}^J$.

7.4.4 Negative Data-Driven Haar-Fisz Transformation for h Unknown

Recall that when the mean-variance relationship is unknown, a function estimate can be obtained by fitting isotonic regression to local estimates of the mean and variance to obtain the function $\hat{h}$, as described by Fryzlewicz *et al.* (2007). Recently, Fryzlewicz (2007) proposed using a Nadaraya-Watson estimator for the mean-variance function. Here, we extend methods using isotonic regression.

We wish to fit a curve which takes account of the different behaviour of the function for positive and negative means. That is, we wish h to be strictly non-increasing for negative means and strictly non-decreasing for positive means.

Assuming this behaviour is correct, we further classify the data into two different situations. Firstly, we may believe that the mean-variance relation is different for positive and negative means and hence we estimate two separate curves, each based solely on either the positive or negative local estimates of the mean and variance.

Alternatively, we might assume that the mean-variance relationship for positive and negative means is equal in value, but opposite in sign. We thus estimate the *absolute* mean-variance relationship and translate it for the negative means. We discuss both of these possibilities next.

DDHFT2. In our first modification to the estimation step, we wish to estimate the mean-variance function $\hat{h}$ such that it is non-increasing for negative μ_i and non-decreasing for positive μ_i. We consider these as two separate functions, called $\hat{h}^-$ and $\hat{h}^+$ respectively, which are calculated using isotone regression separately on both positive and negative μ_i. Our estimate of $\hat{h}$ is then defined as

$$\hat{h} = \begin{cases} \hat{h}^-, & \text{if } \mu_i < 0, \\ \hat{h}^+, & \text{if } \mu_i \geq 0. \end{cases} \tag{7.4.12}$$

This estimate is used as in the original DDHFT. We refer to this modified version as DDHFT2 in the remainder of this chapter.

DDHFT3. Our second method of modifying the mean-variance estimation procedure is as follows. We assume that the mean-variance function we are trying to estimate is such that

$$\sigma_i^2 = h(|\mu_i|). \tag{7.4.13}$$

We thus look at the relationship between the absolute value of the mean and the variance. This is achieved by 'flipping' the negative estimates of the mean to the positive domain, and estimating a non-decreasing curve h as with the original DDHFT. The function h is then translated back to the negative domain for the corresponding negative means. The function $\hat{h}$ is therefore an even function of the mean, μ_i and is defined by:

$$\hat{h} = \begin{cases} h, & \text{if } \mu_i \geq 0, \\ -h, & \text{if } \mu_i < 0. \end{cases} \tag{7.4.14}$$

Where h is the mean-variance estimate from (7.4.13). We refer to this method as DDHFT3 in the remainder of this chapter.

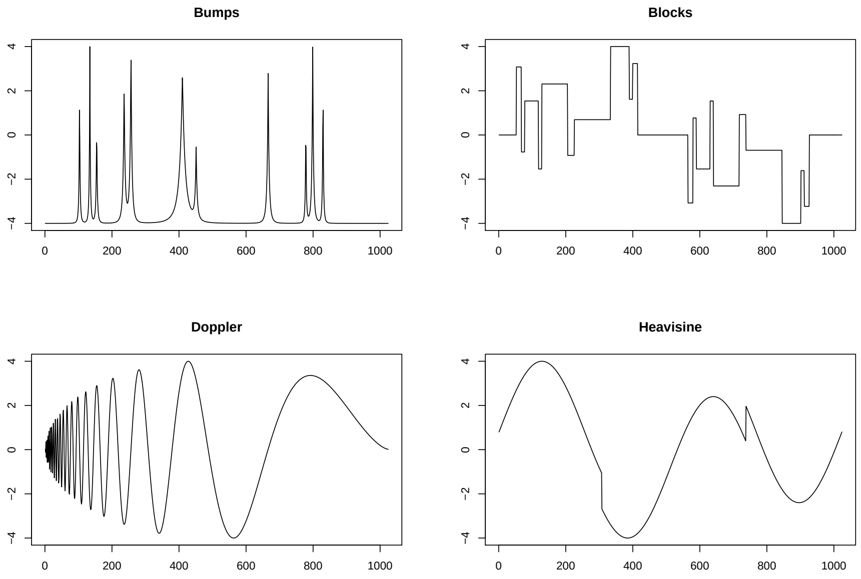


Figure 7.5: Donoho and Johnstone intensity functions translated to have (min, max) of (-4,4).

7.5 Simulated Comparisons

We compare our modifications to the DDHFT with h unknown to the one and two parameter Box-Cox transform. We use the test functions as described in Donoho & Johnstone (1994) as underlying signal intensities which we corrupt with noise and test our methods by comparing how well the true underlying intensity is detected.

The test functions used are the Bumps, Blocks, Heavisine and Doppler signals, of length $n = 1024$, which are linearly shifted and scaled to achieve (min, max) intensities of (-4, 4). We also modified the blocks signal to be two consecutively joined blocks signal, in which the second had been scaled by -1. Each signal we used can be seen in Figure 7.5.

We use the test signals above to act as different underlying Poisson intensities λ_i, $i = 1, \ldots, n$. Each of the signals is corrupted with noise to produce our observed sequence $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ with $n = 1024$. As we wish the v_i to have both positive and negative values, we define the sequence of variables $\mathbf{v}$ such that:

$$v_i \sim \text{Poisson}(|\lambda_i|)\text{sgn}(\lambda_i). \tag{7.5.15}$$

For each of the 4 test signals above, we carry out the following:

1. Create a sequence of variables, v_i, as defined by (7.5.15).
2. Transform the data using both the 1 and 2 parameter Box-Cox transform and our two versions of the DDHFT.
3. Take the discrete wavelet transform of the data, for each of Daubechies extremal phase wavelets with 1 to 10 vanishing moments.
4. Use wavelet thresholding to smooth the transformed data, using EbayesThresh from Johnstone & Silverman (2005a) (as described in Section 5.3.1).
5. Take the inverse wavelet transform of the thresholded sequence.
6. Take the inverse of the method used in step 2.

For each signal, we then have a sequence of known intensities λ_i along with our corresponding estimate, which we denote by $\hat{\lambda}_n$. The mean squared error,

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(\lambda_n - \hat{\lambda}_n)^2, \tag{7.5.16}$$

is used to compare our estimated intensities with the known intensities, the smaller the MSE, the closer the estimate is to the 'truth'.

7.5.1 Simulation Results

The results reported in Tables (7.1)–(7.4) are the mean and standard error (SE) of the MSE (to three decimal places) for 100 replications of the above procedure. For the one parameter Box-Cox transform, the data are arbitrarily shifted to have minimum value of 1. Optimal (or near optimal) values for the transformation parameters are found using the functions `boxcox.fit` and `box.cox` from the R packages `geoR` and `car` respectively. Our modifications to the DDHFT, as well as our threshold smoothing methods use code and functions from `EbayesThresh` and `DDHFm` packages in R.

Overall, apart from the Haar wavelet, there is no significant difference between the performance with different wavelets. For all but the Bumps data, the two methods of the

Bumps								
Wavelet Family	1 Param. B-C		2 Param. B-C		DDHFT 2		DDHFT 3	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	0.609	0.051	0.599	0.053	0.450	0.050	$\boxed{0.439}$	0.056
2	0.516	0.052	0.505	0.050	0.456	0.064	$\boxed{0.429}$	0.054
3	0.502	0.060	0.490	0.058	0.494	0.073	$\boxed{0.466}$	0.058
4	0.537	0.062	0.527	0.060	0.517	0.069	$\boxed{0.502}$	0.063
5	0.528	0.058	0.517	0.058	0.500	0.068	$\boxed{0.490}$	0.065
6	0.511	0.055	0.502	0.054	0.489	0.055	$\boxed{0.486}$	0.058
7	0.505	0.064	$\boxed{0.498}$	0.061	0.532	0.064	0.524	0.064
8	0.553	0.065	$\boxed{0.544}$	0.064	0.575	0.064	0.576	0.062
9	0.574	0.068	$\boxed{0.563}$	0.067	0.567	0.065	$\boxed{0.563}$	0.064
10	0.553	0.060	$\boxed{0.544}$	0.059	0.564	0.061	0.558	0.062

Table 7.1: Mean and standard error (SE) MSE values for Bumps signal. The best mean MSE for each family of wavelet is surrounded by a box. The overall optimal value is found using the DDHFT3 and the wavelet with 2 vanishing moments.

Blocks								
Wavelet Family	1 Param. B-C		2 Param. B-C		DDHFT 2		DDHFT 3	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	$\boxed{0.555}$	0.162	0.562	0.161	0.660	0.207	0.701	0.202
2	0.844	0.090	0.842	0.090	$\boxed{0.638}$	0.096	0.642	0.108
3	0.863	0.081	0.860	0.081	$\boxed{0.650}$	0.097	0.651	0.093
4	0.859	0.085	0.856	0.085	0.717	0.101	$\boxed{0.716}$	0.100
5	0.865	0.090	0.862	0.091	0.733	0.110	$\boxed{0.722}$	0.102
6	0.871	0.085	0.869	0.085	$\boxed{0.713}$	0.092	0.717	0.088
7	0.866	0.084	0.863	0.084	$\boxed{0.692}$	0.093	0.693	0.095
8	0.854	0.090	0.852	0.089	$\boxed{0.673}$	0.093	0.681	0.090
9	0.848	0.087	0.844	0.087	$\boxed{0.663}$	0.102	0.666	0.098
10	0.846	0.085	0.844	0.086	0.675	0.104	$\boxed{0.663}$	0.097

Table 7.2: Mean and standard error (SE) MSE values for Blocks signal. The best mean MSE value for each family of wavelet is surrounded by a box. The overall optimal value is found using the DDHFT2 and the wavelet with 2 vanishing moments.

Heavisine								
Wavelet Family	1 Param. B-C		2 Param. B-C		DDHFT 2		DDHFT 3	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	0.634	0.88	0.615	0.088	$\boxed{0.144}$	0.023	0.147	0.038
2	0.345	0.101	0.345	0.101	0.114	0.026	$\boxed{0.109}$	0.024
3	0.322	0.099	0.318	0.099	0.121	0.030	$\boxed{0.113}$	0.033
4	0.342	0.116	0.340	0.116	0.111	0.032	$\boxed{0.108}$	0.032
5	0.334	0.102	0.334	0.102	0.106	0.031	$\boxed{0.103}$	0.032
6	0.332	0.108	0.329	0.109	0.107	0.034	$\boxed{0.105}$	0.034
7	0.329	0.107	0.329	0.105	0.115	0.034	$\boxed{0.105}$	0.034
8	0.345	0.112	0.343	0.109	0.117	0.030	$\boxed{0.112}$	0.033
9	0.353	0.112	0.351	0.111	0.114	0.030	$\boxed{0.113}$	0.034
10	0.351	0.116	0.350	0.115	0.109	0.031	$\boxed{0.108}$	0.034

Table 7.3: Mean and standard error (SE) MSE values for Heavisine signal. The best mean MSE value for each family of wavelet is surrounded by a box. The overall optimal value is found using the DDHFT3 and the wavelet with 5 vanishing moments.

Doppler								
Wavelet Family	1 Param. B-C		2 Param. B-C		DDHFT 2		DDHFT 3	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	0.799	0.104	0.785	0.103	$\boxed{0.349}$	0.042	0.355	0.052
2	0.530	0.117	0.537	0.112	0.372	0.050	$\boxed{0.359}$	0.048
3	0.482	0.111	0.487	0.110	0.363	0.054	$\boxed{0.339}$	0.049
4	0.469	0.112	0.475	0.109	0.315	0.051	$\boxed{0.292}$	0.044
5	0.450	0.106	0.456	0.105	0.289	0.045	$\boxed{0.278}$	0.043
6	0.435	0.113	0.441	0.110	0.311	0.048	$\boxed{0.297}$	0.043
7	0.453	0.112	0.461	0.108	0.335	0.054	$\boxed{0.313}$	0.050
8	0.447	0.105	0.454	0.103	0.333	0.053	$\boxed{0.312}$	0.049
9	0.456	0.113	0.466	0.112	0.305	0.049	$\boxed{0.289}$	0.040
10	0.442	0.108	0.449	0.108	0.294	0.054	$\boxed{0.276}$	0.045

Table 7.4: Mean and standard error (SE) MSE values for Doppler signal. The best mean MSE value for each family of wavelet is surrounded by a box. The overall optimal value is found using the DDHFT3 and the wavelet with 10 vanishing moments.

DDHFT out perform the Box-Cox transform across wavelets 2–10 (although Bumps is good for wavelets 1–6). It should be noted that the Box-Cox does give better results for Blocks data, when the Haar wavelet is used in the wavelet transform and this is the best overall. The Bumps data produces very similar results for the different methods, with the minimum value of the MSE resulting from one of the Haar-Fisz transforms 6 times out of 10, and once with the same minimal value as a Box-Cox transform.

When comparing the two proposed methods, the Bumps signal is the only data where DDHFT3 has a lower MSE that DDHFT2 over all wavelets, although this difference is quite small. For both the Heavisine and Doppler signals, only once does DDHFT2 outperform DDHFT3. Again, the difference in actual MSE values are minimal, both having a 2% smaller MSE. The MSE for the blocks signal is again very similar, with the minimum values showing more for DDHFT2.

As this is simulated data, it would be expected that both DDHFT2 and DDHFT3 would perform very similarly, as the noise added is symmetrical around zero. Further, given the models, it would be expected that DDHFT3 would out perform DDHFT2. This indeed is the case, with the DDHFT2 having a smaller MSE that DDHFT3 only a quarter of the time. An investigation of the mean-variance relationship is key in deciding which would be best for real-life data.

7.6 Choice of Modification Method

We have produced two modifications of the DDHFT which both assume certain behaviour of the mean-variance function h. When such behaviour is not known, it may be desirable to test h to assess its ‘goodness of fit’. We devise a bootstrap test for choice of transformation which, under the null hypothesis, tests whether the values of h are the same for positive and negative values of μ_i. If they are, we choose to use DDHFT3, otherwise we assume the relationship is independent and use DDHFT2.

7.6.1 Bootstrap Test

The test works as follows: we test the null hypothesis, H_0: positive and negative mean-variance the same (DDHFT3) against the alternative, H_A: the mean-variances are different (DDHFT2), i.e., H_0: model (7.4.14) is true, verses H_A: model (7.4.12) is true.

We denote the estimate of the mean-variance function as $\hat{h}$ for DDHFT3, when the negative data has been 'flipped' and $\widetilde{h}$ for DDHFT2, when the estimate is created by considering positive and negative means separately. We further define the positive and negative components of $\widetilde{h}$ as $\widetilde{h}^+$ and $\widetilde{h}^-$ from (7.4.12).

Given data $\mathbf{X} = X_1, X_2, \ldots, X_n$, with $n = 2^J$, $J = 1, 2, \ldots$, under H_0 we carry out the following:

1. Calculate the test statistic, $TS(\mathbf{X})$ on data $\mathbf{X}$, as described in Section (7.6.2).

2. Perform a bootstrap simulation of the data conditioned on $\hat{h}$ (as described in Section 7.6.3) to obtain a simulated local variance estimate, σ_i^{*2} for each of the local mean estimates, μ_i.

3. Repeat step 2., *Bsim* = 1000 times, calculating $\widetilde{h}(\mu_i)$ for the mean and simulated variance estimates. Calculate the test statistic using $\widetilde{h}(\mu_i)$, from (7.6.17) below, denoting the value by TS_b for $b = 1, \ldots, Bsim$.

4. Compare the test statistic in step 1. with the bootstrap simulations in step 3., to obtain an overall *p*-value for the test.

We next describe our test statistic and bootstrap methods.

7.6.2 Test Statistic

Given data $\mathbf{X} = X_1, X_2, \ldots, X_n$ (or local mean and variance estimates, μ_i and σ_i^{*2}), estimate $\widetilde{h}(\mu_i)$, using DDHFT2. In doing so, we form two separate estimates $\widetilde{h}^-$ and $\widetilde{h}^+$ for negative and positive μ_i respectively. Under the null hypothesis, we assume that the function $\widetilde{h}$ is an even function. That is, if we were to 'flip' its negative part $\hat{h}^-$ so it took values of positive μ_i, we should find that $\widetilde{h}^- = \widetilde{h}^+$. Thus, if the null hypothesis is true, the difference between these values will be small.

Our test statistic is defined as,

$$TS(\mathbf{X}) = \sum_{i=1}^{n} \left(\widetilde{h}^{+}(\mu_i) - \widetilde{h}^{-}(\mu_i) \right)^2, \tag{7.6.17}$$

which is the sum of the squares of the difference of the two estimates, at each of the local mean estimates $|\mu_i|$. Larger values indicate a significant difference between $\widetilde{h}^{-}$ and $\widetilde{h}^{+}$ and the need for separate estimates.

For real life data, $\widetilde{h}^{-}$ and $\widetilde{h}^{+}$ may not be of the same length due to a differing number of positive and negative μ_i values. In these cases we interpolate the respective $\widetilde{h}$ so they have equal length.

7.6.3 Bootstrap Simulations

We refer the reader to Davison & Hinkley (1997) for further details and examples of bootstrap tests. Our bootstrap simulations work as follows. For a given data set $\mathbf{X} = X_1, X_2, \ldots, X_n$, perform the first step of either DDHF transform (which is the discrete Haar wavelet transform), to obtain local estimates of the mean, μ_i and variance σ_i^2 Under the null hypothesis, we estimate the mean-variance relationship $\hat{h}(\mu_i) = \sigma_i^2$, for each local estimate $i = 1, \ldots, n/2$. For each μ_i, we now have a known variance, σ_i^2 and a fitted variance, $\hat{h}(\mu_i)$. We can thus calculate the fitted residuals as

$$r_i = \sigma_i^2 - \hat{h}(\mu_i), \tag{7.6.18}$$

for $i = 1, \ldots, n/2$. For each value of μ_i we create a simulated variance, denoted by σ_i^{*2} and defined by

$$\sigma_i^{*2} = \hat{h}(\mu_i) + r_j, \tag{7.6.19}$$

with j randomly sampled from $1, \ldots, n/2$, with replacement. We thus have a new set of local variance estimates, σ_i^{*2} for the local means, μ_i.

It is possible to invert the initial stage of the DDHF transformation to obtain ‘simulated data’, but in practice we use these mean and variance estimated directly in our calculation of the test statistic.

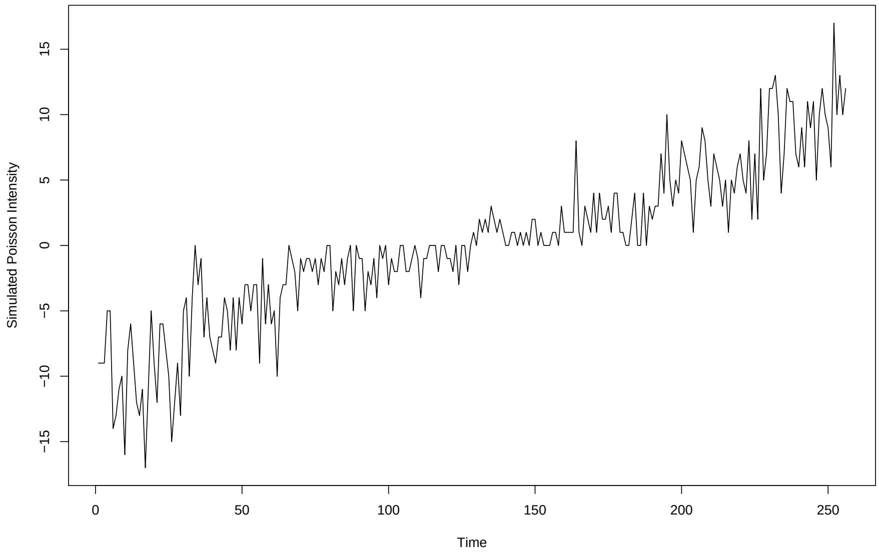


Figure 7.6: Example of simulated Poisson data.

Note that if the fitted variance is close to the original variance then the r_i in (7.6.18) will be small and the re-assignment step in (7.6.19) will cause little change in the variance (and thus the test statistic will be small).

7.6.4 Bootstrap Test Assessment: Test Size

We wish to assess the efficiency of our bootstrap test at identifying whether or not the underlying mean-variance relationship is symmetrical. For this we use simulated data sets for which the truth is known.

We first calculate the size of the test, that is, given that the distribution is known to have a symmetrical mean-variance relationship, we wish to calculate the number of times the tests reject the null hypothesis (i.e. choose DDHFT2 over DDHFT3).

Our simulated data is drawn from a Poisson distribution such that $X_i \sim \text{Poi}(\lambda_i)$ for $i = 1, \ldots, 256$. Our λ_i consist of 8 'blocks' of data of length 32, each with equal intensity λ, taking the values $10, 5, 2, 1, 1, 2, 5, 10$. The first half of the data is then scaled by -1 to create negative data. An example of such data can be seen in Figure (7.6). We create 500 such data sets and test the mean-variance relationship using our bootstrap test.

The proportion of times the test wrongly rejected the null in favour of the alternative was 0.18. This figure is fairly high considering the desired size of a test would be around 0.05. The length of the data and more specifically the *size* of the *blocks* can results in apparent differences between two (identically distributed) sets of generated data. Over a larger sample, two independent, identically distributed random samples should appear more similar and our test would expected to have a smaller ‘size’.

To test this, we calculate the size of the test on the same sequence of λ_i, but with ‘block’ size of 64 and 128 (giving sequence length 512 and 1024 respectively). The size corresponding to these sequences are 0.09 and 0.02. As expected, the tests improve with more data.

7.6.5 Bootstrap Test Assessment: Test Power

We next look at assessing the power of the bootstrap test. Our methodology involves using estimates of the mean-variance function taken from the data. We then change this estimate to produce a set-up close to the original data but with an unknown mean-variance relationship. Starting with data with a known symmetrical mean-variance relationship, we wish to alter the data such that this relationship gradually becomes increasingly different for positive and negative mean values. We do this for two different methods as follows.

We first alter the data by adding a constant to the values for which the mean is positive. This has the effect of shifting the positive part of the mean-variance function to the right. Figure (7.7) shows the mean-variance relationship of the simulated data from Figure (7.6). The solid line is the estimated mean-variance function of the simulated data, whereas the dashed line shows how this line alters for positive means, when a constant of 3 is added to the positive parts of the data.

Our second method of finding the power of the test consists of altering the data in such a way to cause some of the data to become over-dispersed. That is, for the mean-variance function corresponding to positive means, we wish to alter the data in such a way that for a given value of mean, the corresponding variance is larger. We do this by altering the detail coefficients of the HWT of the data, which act as a local estimate of the variance. Once altered, the transformation is inverted to obtain a new sequence which is the same for negative values, but is now over-dispersed for positive values. The detail coefficients

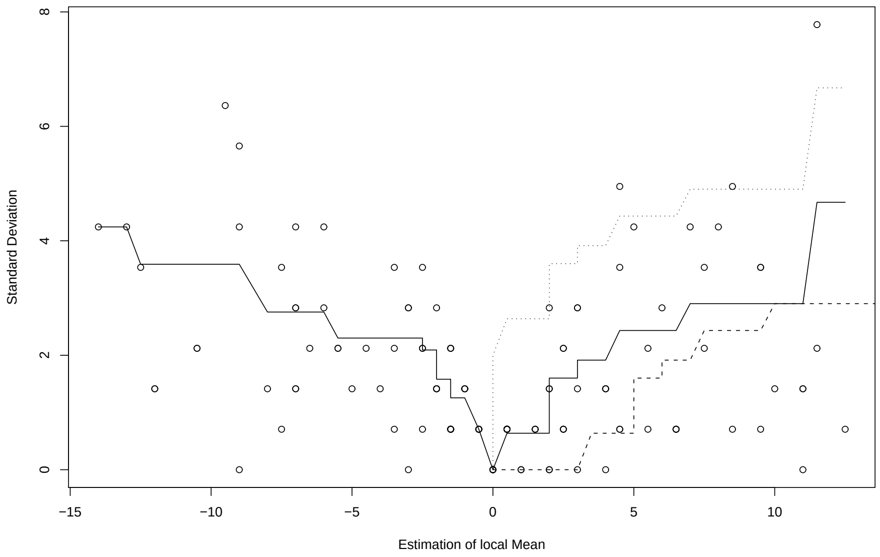


Figure 7.7: Mean-variance estimates from simulated data. Solid line: estimates using DDHFT2. Dashed line: estimate with constant 3 added to the positive mean. Dotted line: estimate with constant 2 added to the standard deviation (corresponding to positive mean values).

corresponding to positive means are transformed such that:

$$d^*_{1i} = d_{1i} + c,$$

where c is a chosen constant. The dotted line in Figure 7.7 shows an alteration with constant 2. Note that we add the constants to the estimates from DDHFT2 (and not DDHFT3) as we wish for the changes to be independent of the values of the negative μ_i.

For both methods, we add a constant to the mean or the variance which increases from 0 to 1 in step sizes of 0.1. For each, the bootstrap test is carried out on 250 generated data sets to detect this change and the power is calculated as the proportion of times the test successfully identifies a change in the mean-variance relationship at the 5% level. The mean size of each of these signals is 0.17 which is expected as we are not adding any constant to either mean or variance (so in essence performing the same calculation as in Section 7.6.4).

The power of the test for when the mean is altered can be seen in Figure 7.8 as a solid line. The dashed line represents the power for the over-dispersed data.

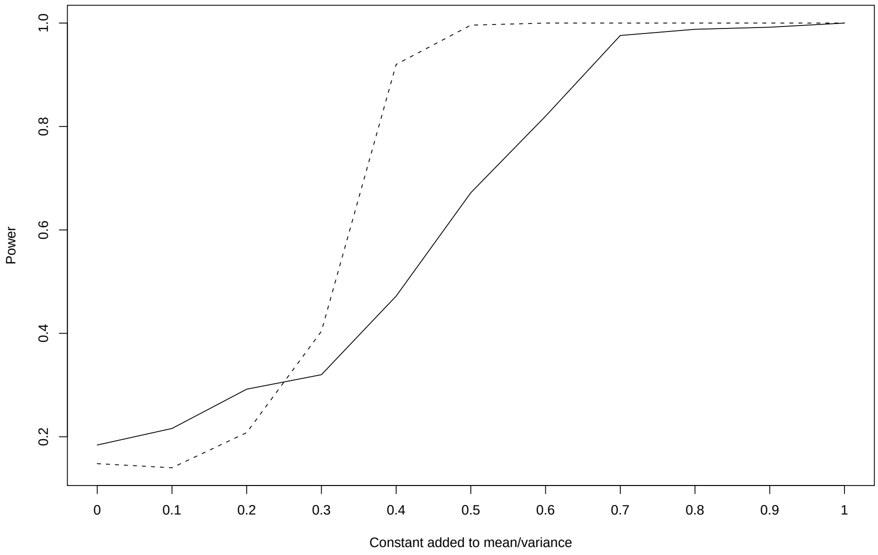


Figure 7.8: Power of bootstrap test. Solid line: changing mean. Dashed line: changing variance (over-dispersion).

7.7 Applications to the Central England Temperature Series

Figure 7.9 displays both estimates $\hat{h}$ and $\widetilde{h}$ for the local mean μ_i and standard deviation σ_i estimates of the CET data. The bootstrap test statistic will be influenced more by the large number of data points close to mean zero, and less for the more sparse data with larger mean magnitude. The *p*-value from the bootstrap test for the CET data was 0.14, so we can't reject the null hypothesis and therefore the use of DDHFT3. From our simulated investigation into the power of this bootstrap test on a similar sized data set, we would expect to accept the null hypothesis at the 95% level if either the difference in mean of the two estimates were greater than 0.7 or the difference in variance were greater than 0.4

In an applied sense, as the temperature deviates from the 'base' rate the range of values it takes will also increase. The symmetry of the mean-variance relationship suggests that the rate of this variance is to an equal level, whether the temperature is getting higher or lower (than the base rate).

Figure 7.10 shows the CET data set together with smoothed estimates. The solid line uses DDHFT3 and kernel regression smoothing using a local plugin bandwidth (using the `lokern` package in R. See Brockmann *et al.* (1993)). The corresponding values using

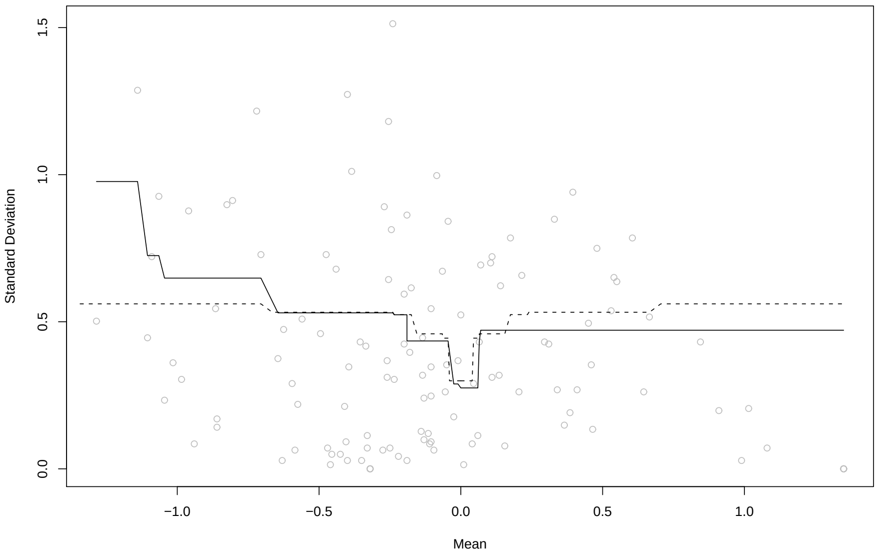


Figure 7.9: Central England Temperature data. Small circles: plot of local standard deviation, $\hat{\sigma}_i^2$, verses local mean, $\hat{\mu_i}$. Solid line: estimated mean-variance relationship function $\hat{h}$ using DDHFT2 method of isotonic regression. Dashed line: estimated mean-variance relationship using DDHFT3 (symmetric).

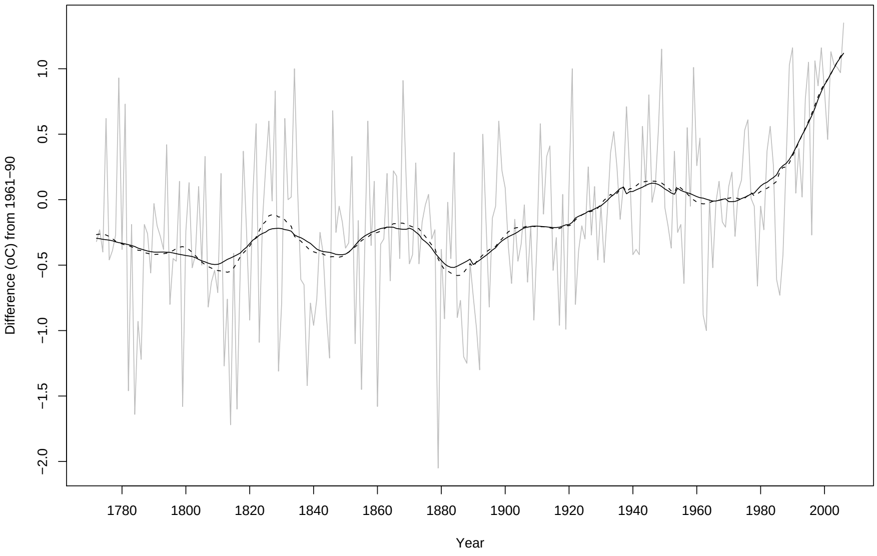


Figure 7.10: Smoothed CET data, using DDHFT modifications and kernel regression smoothing. Solid line: using DDHF2. Dashed line: using DDHF3.

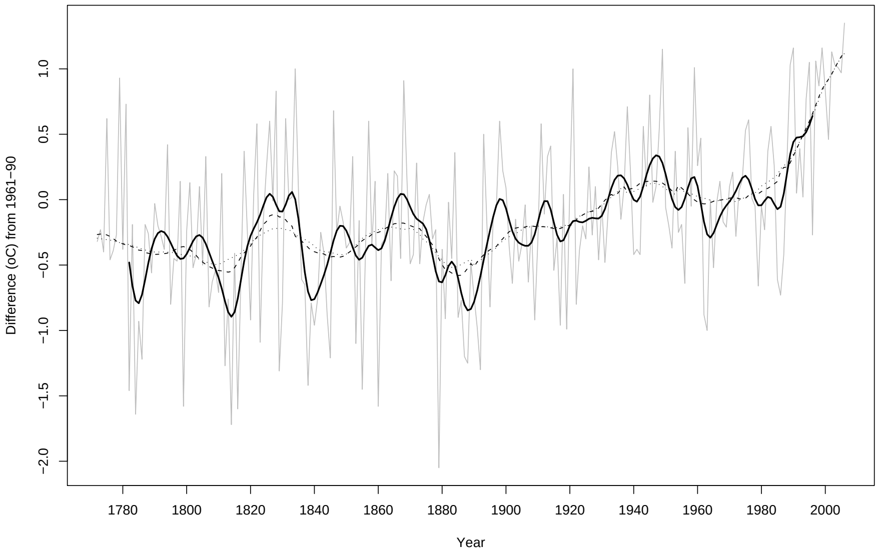


Figure 7.11: Smoothed CET data, using DDHFT modifications (with kernel regression smoothing) and a 21-point binomial filter. Solid line, Binomial filter. Dotted line: DDHFT2. Dashed line: using DDHFT3.

DDHFT2 are shown as a dashed line. For comparison, we plot these two estimates again in Figure 7.11 as a dotted and dashed line respectively, but with the 21-point binomial filter estimate as a solid line.

Both estimates are similar, with the DDHFT2 estimate varying more around 1820, and overall being slightly more variable than the DDHFT3. Our bootstrap test suggested that the DDHFT3 was more suitable for the data. The lack of variability means that peaks can be more accurately assessed and identified. Based on the DDHFT3 estimate, our intensity estimation appears less variable than that of the binomial filter shown in Figure 7.11. For the first half of the data, the temperature exhibits peaks in temperature but continually returns to a base rate of around -0.5. From the end of the 19th century, the temperature increases steadily. From 1970 the temperature rises at a much faster pace and continues up to the end of the data. Similar conclusions are drawn when comparing the DDHFT2 estimate.

7.7.1 Model Checking

It is worth considering the statistical properties of A_t and the DDHFT3 version a_t. Figure 7.12 shows several autocorrelation (acf) plots. The first, (a.) shows the acf of the original

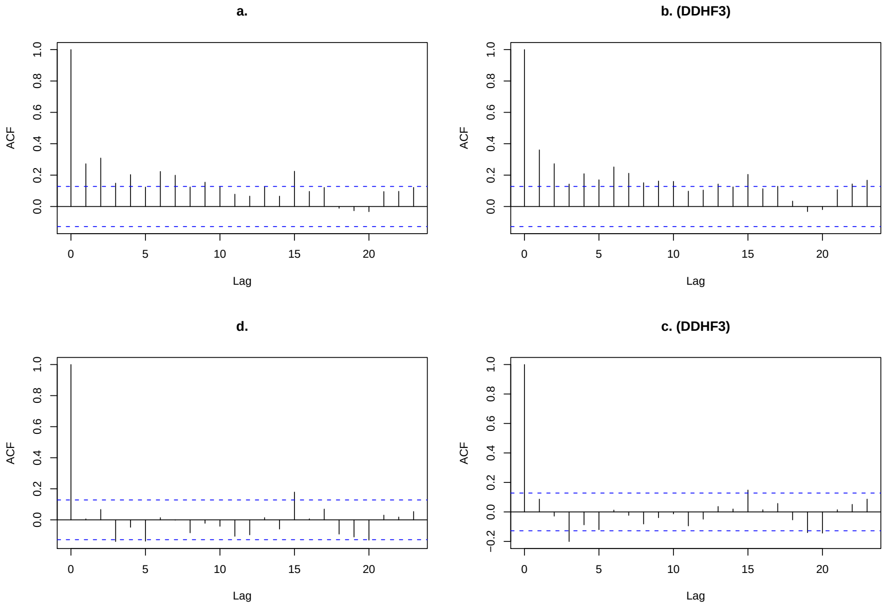


Figure 7.12: Clockwise from top left: autocorrelation functions of (a.) CET data, A_t; (b.) a_t, the DDHFT3 of A_t; (c.) a_t minus kernel regression smoothing estimate; (d.) A_t minus signal estimate inverse DDHFT3.

sequence, which we denote by A_t and plot (b.) shows the same for the DDHFT3 of A_t, denoted by a_t. There is some indication that the sequences might be autocorrelated, but we also believe that the mean of each sequence is not constant (as this is what we are trying to estimate). Figure 7.12c. shows the acf after subtracting the mean estimate using kernel regression smoothing from a_t. After the mean has been taken into account, the acf virtually disappears. Plot (d.) shows the acf of (c.) but in the original data domain. Again, the acf has almost entirely disappeared. Hence, once the mean has been estimated we have evidence that the sequence A_t is uncorrelated.

We test the constant variance assumption of the transformed residuals using the Breusch & Pagan (1979) test. For the kernel regression smoothing, the p-value is 0.16 and hence no (formal) evidence for non-constant variance.

We use the Kolmogorov-Smirnov test on the same residuals and find a p-value of 0.23 and hence no evidence against Gaussianity.

7.8 Conclusions and Future Work

This chapter has proposed modifications to the data-driven Haar-Fisz transformations for both known and unknown mean-variance functions. The modifications allow transformation of negative data with variance related to the absolute value of the mean. Comparisons using simulated data shows that our methods outperform the traditional Box-Cox transform over a variety of noise corrupted intensity signals. It should be noted though that for the underlying Bumps signal, the performance was less emphatic and the MSE results were similar over most wavelets. Also, for the Blocks signal, the Box-Cox transform out-performed the DDHF transforms only once, but the MSE for this single instance was lower than over all other replications.

This point shows that the tests were perhaps misleading. The wavelet used was varied to test the methods over a range of estimators but in fact what we are doing is testing how well the wavelets perform, given a certain transformation. Further simulated studies should only use the best performing wavelet for each signal, and compare it to different smoothing methods rather than different wavelets. Using other Gaussianising or variance stabilising transformations (such as Anscombe (1948) or the negative Haar-Fisz transform proposed within this chapter) would also benefit the study.

Our modification to the data-driven Haar-Fisz transform depends on whether the mean-variance relationship for negative means is the same as for positive means, i.e., if the function h is even. We proposed using bootstrap resampling to assess the significance of a test for symmetry. For simulated data, our test appears to perform well, although is less accurate for smaller data sets. For the central England temperature data, our bootstrap test does not reject the null hypothesis that the mean-variance function is symmetric around zero. We displayed smoothed values of the CET data using both methods, and concluded that from the end of the 19th century the temperature increases, and that this increase is much more rapid after the 1970s.

Although these methods appear sound, it is still questionable, however, how suitable these transforms are for the actual data. The transforms assume a ‘turning point’ in the mean-variance relationship where the estimated function changes from being non-increasing to non-decreasing. The modifications proposed in this chapter both assume this point to be at zero, which, from figures 7.2 and 7.3 may not be the case for the CET data. Furthermore, the behaviour of the mean-variance relationship for positive means remains unclear. This may not be the case for other data sets and generalising the modifications to select the turning point is left as future work.

Chapter 8

Gaussianisation using Haar-Fisz Transforms

8.1 Introduction

Gaussianity of a signal, or rather residual noise of a signal is a common requirement in may applications. For example, in signal estimation it is often assumed that the observed signal g_t, is such that

$$g_t = f_t + \varepsilon_t,$$

where f_t is the 'true' underlying signal and ε_t the noise, is distributed as iid $N(0, \sigma^2)$.

Applications in Chapters 6 and 7 focused primarily on the variance-stabilising properties of the HFT and the DDHFT. Although we never formally used them for Gaussianisation, we briefly mentioned testing the transformed signals to see how well they coped with this task. It was often found that they performed well, considering that this was not their main task. In this chapter we consider Haar-Fisz based transforms for the primary purpose of Gaussianisation, so that we may transform non-Gaussian signals for possible use within many other procedures with Gaussian constraints.

In order to use the Haar-Fisz transforms for Gaussianisation, we will introduce a parameter to the 'Fisz' step of the procedure (where the detail coefficients are 'stabilised'). By using a similar maximum likelihood parameter estimation method as the Box-Cox transform, we select our parameter to best Gaussianise the data.

This chapter is a computational study and the underlying theory is left as future work. We first look at the case where the distribution of the data is known (and Poisson) and then consider possible generalisations to make the process of mean-variance estimation data-driven.

8.2 General Haar-Fisz Transform

Recall from Section 5.5.4 that the Haar-Fisz transform (HFT) decomposes an input vector $\mathbf{v} = (v)_{i=1}^{N}$ where $N = 2^J$, using the discrete Haar transform to form smooth and detail vectors $\mathbf{s}^j$ and $\mathbf{d}^j$ of the original vector at scale j. The coefficients d_k^j are then variance stabilised by division of a function of the variance to produce a vector of *Fisz coefficients* $\mathbf{f}^j$, defined by

$$f_k^j = \frac{d_k^j}{h^{1/2}(s_k^j)}, \tag{8.2.1}$$

for $k = 1, \dots, 2^j$.

Kendall *et al.* (1983, page 103) make the comment that variance-stabilising transformations commonly Gaussianise as a by-product, although they do not tend to produce optimum Gaussianisation. Thus, the distribution of the transformed sequence in (8.2.1) would, to a certain degree be expected to be closer to Gaussian than the original sequence. For Poisson data, where $v_i \sim$ Pois(λ_i), we have $\mu_i = \lambda_i$ and $\sigma_i^2 = \lambda_i$ which gives $h(x) = x$, the identity function. The *Fisz coefficients* in (8.2.1) thus become

$$f_k^j = \frac{d_k^j}{s_k^{j\,1/2}}. \tag{8.2.2}$$

As detailed by Fryzlewicz & Nason (2004, Proposition 2), this transform asymptotically brings vectors of Poisson counts to Gaussianity with variance one, as the mean of the Poisson counts, and length of the data both tend to infinity. We next consider the HFT for the purpose of Gaussianisation and apply maximum likelihood techniques to a general form of (8.2.2) to select a transformation parameter which best Gaussianises the data.

We generalise (8.2.2) by replacing the square root function by a transformation parameter, which we call α. Furthermore, we define different values of α for each scale of the wavelet transform. Thus $\boldsymbol{\alpha} = (\alpha_J, \alpha_{J-1}, \ldots, \alpha_1)$ allows the transform to be 'local' for each of the J levels. The general *Fisz coefficient* of the kth element on the jth 'level' can then be defined as

$$f_k^j = \frac{d_k^j}{{s_k^j}^{\alpha^j}}. \tag{8.2.3}$$

Our aim is to select the parameters $\boldsymbol{\alpha}$ *for each scale* such that the transformed data is as Gaussian as possible. We select $\boldsymbol{\alpha}$ using maximum likelihood techniques. We denote the general Haar-Fisz transform by the operator $\mathcal{F}_{\boldsymbol{\alpha}}$ and its operation on the vector $\mathbf{v} = (v_1, v_2, \ldots, v_N)$, for $N = 2^J$ by $\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v}$. We wish to use a linear model with Gaussian errors to represent the transformation on a set of regression variables $\mathbf{X}$. As with Box & Cox (1964), we do not directly assume that the transformation can be written in the form $\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v} = \mu + \epsilon$ but instead assume that for some $\boldsymbol{\alpha}$,

$$\mathbb{E}(\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v}) = \mu,$$

where $\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v}$ is the vector of transformed observations of length N and μ is unknown. This is a similar set-up to the theory of choosing parameters for the Box-Cox transform (see Atkinson (1987), Chapter 6 and the review in Section 5.5.2), and as such variance stability is a secondary goal. To compare different values of $\boldsymbol{\alpha}$ it is necessary to compare the likelihood to that of the original observations $\mathbf{v}$ which is

$$\prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} \exp\{-(\mathcal{F}_{\boldsymbol{\alpha}} v_i - \mu)^2 / 2\sigma^2\} \mathcal{J}, \tag{8.2.4}$$

where the Jacobian is given by

$$\mathcal{J} = \prod_{i=1}^{N} \left| \frac{\partial \mathcal{F}_{\boldsymbol{\alpha}} v_i}{\partial v_i} \right|. \tag{8.2.5}$$

Note the slight change in notation for the Jacobian, $\mathcal{J}$, from that in Section 5.5.2. This is to avoid confusion with the number of levels of the wavelet transform, J. The Jacobian allows the transformed variables to be on the same scale for each value of $\boldsymbol{\alpha}$. For fixed $\boldsymbol{\alpha}$, (8.2.4) is the likelihood for the least squares problem with response $\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v}$ (as the Jacobian does not depend on either β or σ). Once again, using the same notation as Atkinson we denote the maximum likelihood estimates of β for a given vector $\boldsymbol{\alpha}$ by $\hat{\beta}(\boldsymbol{\alpha})$. The least squares estimates are therefore given by

$$\hat{\mu}(\boldsymbol{\alpha}) = \overline{\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v}},$$

the mean of the transformed observations. The residual sum of squares of the $\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v}$ is

$$S(\boldsymbol{\alpha}) \quad = \quad \sum_{i=1}^{n}(\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v} - \overline{\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v}})^2. \tag{8.2.6}$$

As with the Box-Cox transform, division of this by N yields the maximum likelihood estimate of σ^2 as

$$\hat{\sigma}^2(\boldsymbol{\alpha}) = S(\boldsymbol{\alpha})/N. \tag{8.2.7}$$

For fixed $\boldsymbol{\alpha}$ we maximise the log-likelihood over both μ and σ^2 by substituting the expressions for $S(\boldsymbol{\alpha})$ and $\hat{\sigma}^2(\boldsymbol{\alpha})$ from (8.2.6) and (8.2.7) respectively into the logarithm of the likelihood given by (8.2.4). This gives

$$l_{\text{max}}(\boldsymbol{\alpha}) = -(N/2)\log\hat{\sigma}^2(\boldsymbol{\alpha}) + \log\mathcal{J}. \tag{8.2.8}$$

This is the partially maximised log-likelihood and is a function of $\boldsymbol{\alpha}$ in terms of both the residual sum of squares and the Jacobian.

For the Box-Cox transformation, detailed in Section 5.5.2, Gaussianising (8.2.8) by division of $\mathcal{J}^{1/n}$ simplifies the likelihood equations by removing the dependency upon the Jacobian. There is no such obvious simplification for the HFT so we use the form of $l_{\text{max}}(\boldsymbol{\alpha})$ in (8.2.8). Furthermore, an algebraic representation of the Jacobian, $\mathcal{J}$, is not simple so we opt instead to use a numerical approximation of the Jacobian, adapted from the algorithm given by Press *et al.* (1992, page 388), which estimates the partial derivatives matrix in

8.2.5.

We now give the algorithm for the general HFT for data $\mathbf{v} = (v_1, v_2, \ldots, v_N)$, for $N = 2^J$, when the function h is known.

1. Let $s_i^J = v_i$, for $i = 1, \ldots, n$.

2. For each $j = J, J-1, \ldots, 1$, recursively form the vectors $\mathbf{s}^j$ and $\mathbf{f}^j$:

$$s_k^j = \frac{s_{2k-1}^{j+1} + s_{2k}^{j+1}}{2};\ f_k^j = \frac{s_{2k-1}^{j+1} - s_{2k}^{j+1}}{2h^{\alpha^{j+1}}(s_k^j)}, \tag{8.2.9}$$

   for $k = 1, \ldots, 2^j$ (where $h(s_k^j) = s_k^j$ when $\mathbf{v}$ is Poisson).

3. For each $j = 1, 2, \ldots, J$, recursively modify s^{j+1}:

$$s_{2k-1}^{j+1} = s_k^j + f_k^j;\ s_{2k}^{j+1} = s_k^j - f_k^j,$$

   for $k = 1, \ldots, 2^j$.

4. Set $\mathbf{Y} = \mathbf{s}^J$

We will refer to this transform as the general Haar-Fisz transform, and denote it by HFT$\boldsymbol{\alpha}$. We find the optimal α_j by maximising $l_{\text{max}}(\boldsymbol{\alpha})$ in (8.2.8) numerically, in R using the `optim` function over the α_j. The general HFT can then be inverted in the same way as the regular HFT, as described in Section (5.5.4), but using the values of $\boldsymbol{\alpha}$ to reverse the operator f_k^j in (8.2.9).

We define another version of this generalised Haar-Fisz transform by imposing the extra constraint that $\alpha_j = \alpha$ for all j in $J, J-1, \ldots, 1$. This single parameter (or constrained) model assumes that the α_j are the same for all levels of the Haar transform and we denote it by HFT$\underline{\alpha}$ in the remainder of this chapter. With a single parameter, the optimisation routines are considerably faster than the multi-parameter general transform HFT$\boldsymbol{\alpha}$, so its performance is of interest. Furthermore, a single parameter allows us to graphically explore the outcome of the optimisation routines on l_{max}.

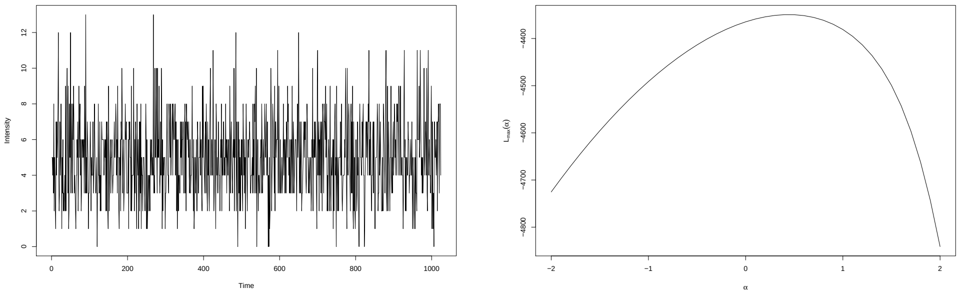


Figure 8.1: Left: Poisson signal of mean 5. Right: $l_{\max}(\boldsymbol{\alpha})$ over different values of α

8.2.1 Examples

As an example of our transformation, we generate a sequence of Poisson variables of length 1024 and mean 5, as shown in the left plot in Figure 8.1. We use our constrained version, HFT$\underline{\alpha}$ and calculate $l_{\max}(\underline{\alpha})$ for different $\underline{\alpha}$. The plot of $l_{\max}(\underline{\alpha})$ can be seen in the right plot in Figure 8.1. The maximum occurs at $\underline{\alpha} = 0.44$. Note that this is close to, but not the same as, the (variance stabilising) optimal value of the HFT of 0.5. Therefore, if the data is known to be Poisson, one would obtain near-Gaussianisation (but non-optimal), by applying the variance stabilising value of $\underline{\alpha} = 0.5$ instead of searching over all possible values.

It should be noted that the estimation of the Jacobian $\mathcal{J}$ can cause numerical difficulties in finding the maximum likelihood: although the logarithm of the Jacobian is finite (as with the case of the example), the actual Jacobian can be very large. This occurs to such an extent that a computational representation can become infinite and induce numerical instability. This is dependent on factors such as the length or intensity of the data, and we have observed these problems mostly for negative values of α smaller than -2, and very large positive values of α. In most circumstances we have found that a maximum likelihood estimate of α still exists (and has commonly been observed between the range of 0 and 2). We discuss this further in Section 8.4.

We test our code by generating a signal of length 512 from a Gaussian distribution with mean 5 and variance 1. The idea is to invert this sequence with known values of $\boldsymbol{\alpha}$, so that we have a sequence of data for which we know the parameters which will return the data

Known α_j	Mean Estimated Alpha									
	$\underline{\alpha}$	α_9	α_8	α_7	α_6	α_5	α_4	α_3	α_2	α_1
0	0.10	0.17	0.17	0.17	0.16	0.16	0.16	0.10	-0.04	-0.15
0.5	0.55	0.55	0.55	0.55	0.54	0.54	0.52	0.47	0.33	0.11
1	1.16	1.16	1.19	1.21	1.22	1.22	1.19	1.14	1.04	0.80
0.1–0.9	–	0.12	0.22	0.32	0.42	0.51	0.59	0.69	0.80	0.76

Table 8.1: Values of α_j for constant Gaussian signal (mean 5, variance 1), given known α_j

back to Gaussian. We can then compare these to those found using the maximum likelihood techniques.

We invert our Gaussian data using values of α fixed over each level so that $\boldsymbol{\alpha} = \underline{\alpha}$. We use arbitrary values of 0, 0.5 and 1. We optimise $l_{\text{max}}(\alpha)$ for both the general and constrained models to obtain our estimates of α. For the HFT$\boldsymbol{\alpha}$ we also invert our initial Gaussian sequence with α_j ranging from 0.1 – 0.9 in increments of 0.1 (as j decreases). For all of our values of α_j, we repeat over 100 random Gaussian sequences and the mean values of the estimates of α_j are given in Table 8.1.

We find $\underline{\alpha}$ values of 0.10, 0.55 and 1.16 respectively for each value of our α. For the multi-parameter HFT$\boldsymbol{\alpha}$, our methodology mostly finds values of α_j which are close to the known value, although less so for the initial value of $\boldsymbol{\alpha} = 1$. As our initial sequence was random, some fluctuation from the known true value of α is to be expected. When we vary α_j for each level and use the HFT$\boldsymbol{\alpha}$, our optimal values of α_j are very close to the known true values. This suggests a certain degree of uniqueness within the parameters. We do not, however, rule out the possibility that other sets of parameters may give equally good Gaussianisation.

In our initial investigations we find that the value of $l_{\max}(\alpha)$ appears less dependent upon α_j for coarser scales (smaller j), and a range of values of the α_j produce very little change in $l_{\max}(\alpha)$ compared to changing α_j at the smoother scales (larger j). The finer scales of the detail coefficients represent noise within the data, whereas the coarser scale coefficients contain mostly signal (as the noise can be thought of as having been 'smoothed' out over the finer scales). Transforming the signal, compared to the noise will have less of a Gaussianising effect. This is a possible explanation for why the likelihood appears less dependent on coarser α_j, although we leave any detailed investigation into the sensitivity of

$l_{\max}(\boldsymbol{\alpha})$ with j as future work.

We also note here the computational time for our transformations. Performing our calculations in R on a 2.2 GHz AMD Opteron with 2Gb RAM, we found that for data of length 1024, one simulation of our constrained model took between 2 and 3 minutes to run and the general transformation took between 40 and 60 minutes. This long computational time was caused by the estimation of the partial derivatives for each transformation parameter over each scale.

8.2.2 Gaussianisation Simulations

We compare the Gaussianisation of both the HFT$\underline{\alpha}$ and the HFT$\boldsymbol{\alpha}$ with that of the one parameter Box-Cox and the identity transformation on simulated data sets.

We generate 4 underlying intensity signals of length 1024: a constant signal of 4, and the Donoho & Johnstone (1994) Blocks, Doppler and Heavisine signals, which we transform linearly to have (minimum, maximum) of (1/8, 8). These are referred to as ‘small’ intensity signals in the remainder of this chapter. We also create a set of ‘large’ signals, again using the Donoho & Johnstone (1994) signals, but with (min, max) equal to (1/128, 128). We also have a ‘large’ constant signal of intensity 64.

We denote our underlying intensity by $\boldsymbol{\lambda}$ and generate signals of Poisson variables $\mathbf{v} = \text{Poi}(\boldsymbol{\lambda})$ for each of the signals. We judge the success of Gaussianisation of our HF transforms by comparing the residuals $\mathcal{F}_{\boldsymbol{\alpha}}\mathbf{v} - \mathcal{F}_{\boldsymbol{\alpha}}\boldsymbol{\lambda}$ and $\mathcal{F}_{\underline{\alpha}}\mathbf{v} - \mathcal{F}_{\underline{\alpha}}\boldsymbol{\lambda}$, with those of the Box-Cox (BC) transform ($\mathcal{B}\mathbf{v} - \mathcal{B}\boldsymbol{\lambda}$) and the identity (ID) transform ($\mathbf{v} - \boldsymbol{\lambda}$). We compare the transforms by considering the Q-Q plots of each of the residuals and we also test the residuals using the Kolmogorov-Smirnov (KS) test of Gaussianity.

Figure 8.2 shows the mean of the Q-Q plots for the small intensity signals, taken over 100 sample signals. We also show a histogram of the values of $\underline{\alpha}$ maximising the likelihood for each signal for the constrained HFT$\underline{\alpha}$.

The Q-Q plots of the transformations indicate that over all the signals the HFT$\boldsymbol{\alpha}$ (plotted in green) out performs the other transforms, signified by the comparative straightness of the line. The plot for the constant signal is still stepped, but less so than the other transforms (and is only a slight improvement on the constrained HFT$\underline{\alpha}$). For the Blocks signal, the

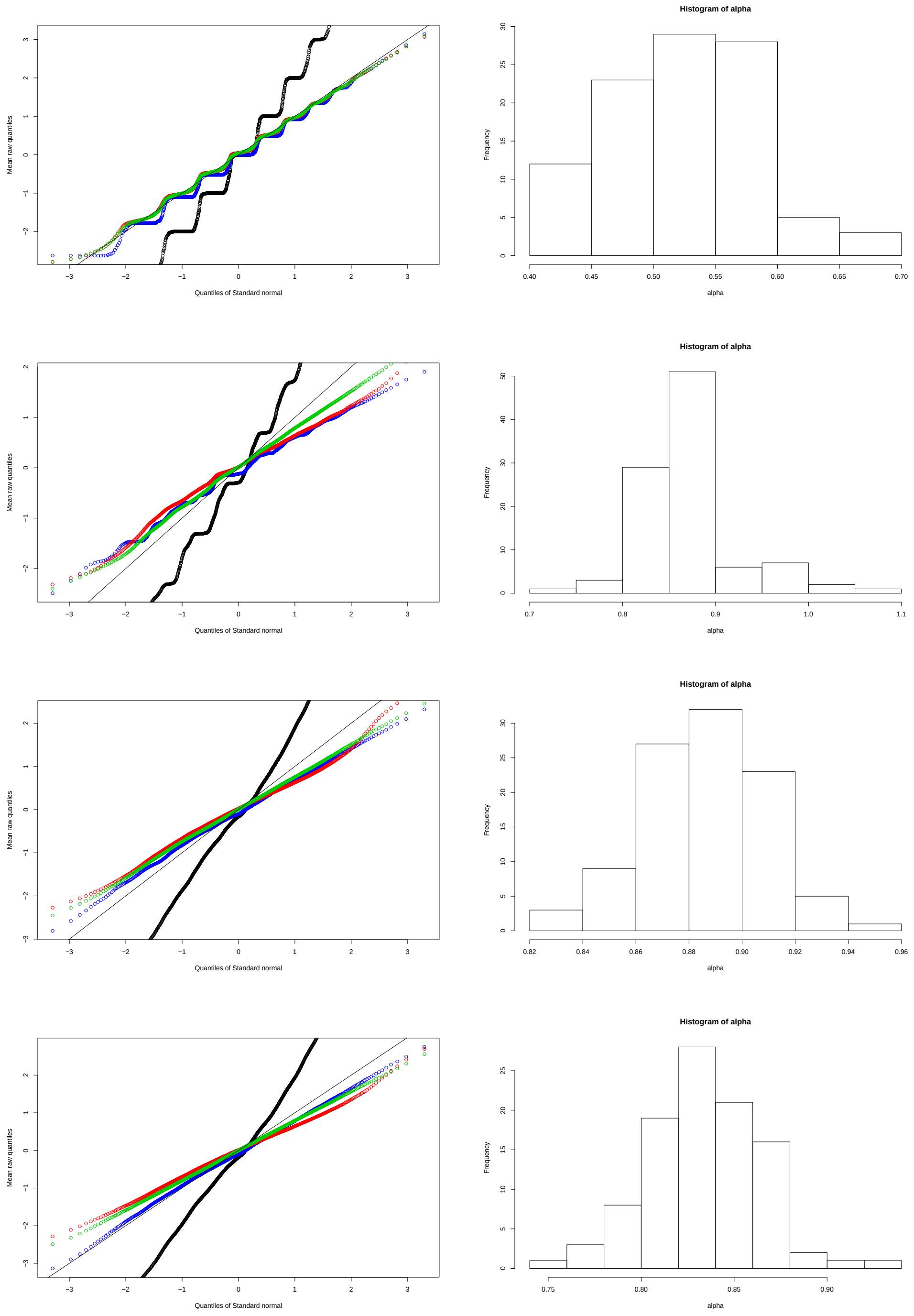


Figure 8.2: Underlying intensity signal from top: Constant; Blocks; Heavisine and Doppler. Left plots: QQ-Plots. Black: Identity transform; Blue: Box-Cox transform; Red: HFT$\underline{\alpha}$; Green: HFT$\boldsymbol{\alpha}$. Solid line has slope 1, indicating unit variance. Constant intensity = 4, all others have (min,max) of (1/8, 8). Right: Corresponding histogram of $\underline{\alpha}$ values.

	KS Test				Variance			
Signal	ID	BC	HFT$\underline{\alpha}$	HFT $\boldsymbol{\alpha}$	ID	BC	HFT$\underline{\alpha}$	HFT$\boldsymbol{\alpha}$
Constant	1.1e-13	5.1e-10	2.7e-4	$\boxed{0.0027}$	4.01	1.02	0.95	0.95
Blocks	4.7e-6	9.1e-7	0.0013	$\boxed{0.27}$	4.02	0.47	0.45	0.63
Heavisine	0.0039	0.39	0.15	$\boxed{0.71}$	4.19	0.60	0.50	0.58
Doppler	0.0064	0.57	0.38	$\boxed{0.70}$	4.45	0.77	0.48	0.63

Table 8.2: Mean p-values of Kolmogorov-Smirnov Gaussianity test for 'small' intensity signals. The best overall p-value for each signal is boxed (note that other transforms may still have statistically significant p-values).

	Mean Estimated α_j										
Signal	$\underline{\alpha}$	$\alpha_1 0$	α_9	α_8	α_7	α_6	α_5	α_4	α_3	α_2	α_1
Constant	0.52	0.52	0.52	0.52	0.51	0.52	0.52	0.50	0.40	0.37	0.01
Blocks	0.87	0.67	0.65	0.69	0.83	1.04	1.65	1.72	2.17	2.09	1.43
Heavisine	0.89	0.81	0.69	0.64	0.62	0.69	0.82	3.26	1.43	2.83	0.32
Doppler	0.83	0.70	0.61	0.63	0.71	0.92	1.24	1.70	2.49	2.09	0.96

Table 8.3: Values of α_i for the 'small' intensity signals for both constrained HFT$\underline{\alpha}$ and the general HFT$\boldsymbol{\alpha}$.

Q-Q plot of the HFT$\underline{\alpha}$ is also a lot smoother than that of the Box-Cox transform. The Q-Q plot of the Box-Cox transform is, however, closer to the solid line for the Heavisine and Doppler signals, indicating that the transformed residual variance closer to 1 than for the other transforms.

Table 8.2 shows the mean p-values of the KS-test under the null hypothesis that the residuals are drawn from a Gaussian distribution (and the alternative that they are not). For reference, we also show the variance of the transformed residuals. The values coincide with the interpretations of the Q-Q plots that the general HF transform performs best at Gaussianisation compared to the other transforms. The p-value is significant when transforming the constant signal using the general HF transform, although this was apparent in the stepped nature of the Q-Q plot. For the Heavisine and Doppler signals, the KS test does not reject the hypothesis of Gaussianity for both the Box-Cox and the constrained general HF transforms.

The optimal values of α for both HF transforms are given in Table 8.3. We note that for the constant signal, both transforms have α_i close to 0.5, indicating that the variance should also have been stabilised around 1 (which is true, from Table 8.2). We again observe the

	KS Test				Variance			
Signal	ID	BC	HFT$\underline{\alpha}$	HFT$\boldsymbol{\alpha}$	ID	BC	HFT$\underline{\alpha}$	HFT$\boldsymbol{\alpha}$
Constant	0.073	0.12	0.42	$\boxed{0.76}$	64.18	13.26	0.55	0.76
Blocks	0.048	$\boxed{0.25}$	1.2e-5	1.7e-4	62.61	1.78	0.0083	0.39
Heavisine	0.020	0.36	0.077	$\boxed{0.75}$	66.00	6.80	0.16	1.18
Doppler	0.0047	0.21	0.55	$\boxed{0.79}$	70.19	20.47	0.24	1.06

Table 8.4: Various statistics for 'large' intensity signals. The best overall p-value from the KS test for each signal is boxed (note that other transforms may still have statistically significant p-values).

	Mean Estimated α_j										
Signal	α	α_10	α_9	α_8	α_7	α_6	α_5	α_4	α_3	α_2	α_1
Constant	0.89	0.56	0.56	0.56	0.56	0.56	0.55	0.54	0.51	0.50	0.43
Blocks	1.54	0.92	1.00	1.14	1.26	1.45	1.56	1.57	1.72	1.74	1.49
Heavisine	0.77	0.46	0.46	0.49	0.56	0.81	1.15	1.38	1.08	1.54	0.56
Doppler	0.69	0.44	0.50	0.64	0.78	0.93	1.18	1.20	1.37	1.30	0.96

Table 8.5: Values of α_i for the 'large' intensity signals for both constrained HFT$\underline{\alpha}$ and the general HFT$\boldsymbol{\alpha}$.

variation in values of α_i for larger i for the general HFT.

Figure 8.3 shows the Q-Q plots of the four transforms for the larger signals. For all but the Blocks signal (which we will discuss separately), both the HF transforms are closer to the solid line than the other transforms. Furthermore, the lines appear at least as straight as the Box-Cox and identity transformations (except for the constrained HF for the Heavisine). This is further supported by testing the residuals for Gaussianity, as the p-values of the KS test indicate in Table 8.4.

Excluding the Blocks signal, the HFT$\boldsymbol{\alpha}$ has a less significant p-value than the other transforms. It should be noted, however, that with these larger underlying intensity signals, it is only the identity transformation on the Doppler signal which is significant at the 5% level (as with large mean values for our Poisson signals, the Central Limit Theorem comes into effect).

Table 8.5 shows the optimal values of α_j for both HF transformations. Compared with the 'small' intensity signals, $\underline{\alpha}$ is larger for the constant signal but smaller for both Heavisine and Doppler signals. For the HFT$\boldsymbol{\alpha}$, the optimal $\boldsymbol{\alpha}$ are again close to 0.5 for the constant intensity. The optimal $\boldsymbol{\alpha}$ for the Heavisine and Doppler signals are smaller than those for the

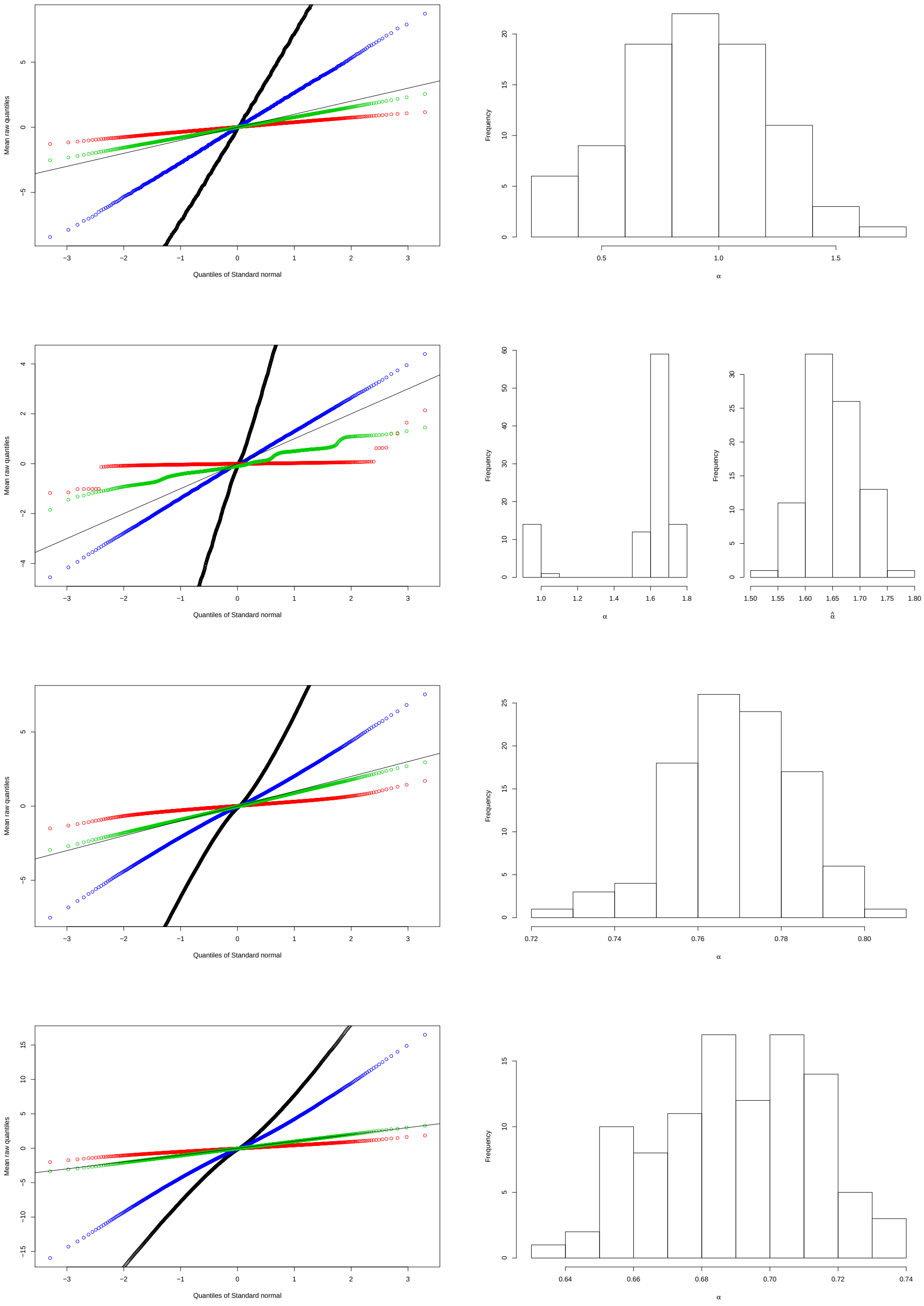


Figure 8.3: Underlying intensity signal from top: Constant; Blocks; Heavisine and Doppler. Left plots: QQ-Plots. Black: Identity transform; Blue: Box-Cox transform; Red: HFT$\underline{\alpha}$; Green: HFTα. Solid line has slope 1, indicating unit variance. Constant intensity = 64, all others have (min,max) of (1/128, 128). Right: Corresponding histogram of $\underline{\alpha}$ values.

smaller signals and again appear to increase in value for the coarser levels (with a somewhat erratic behaviour for α_1).

The Blocks signal for larger intensities exhibits unexpected behaviour while using the both HF transforms. Furthermore, the histogram of the optimal values of $\underline{\alpha}$ show two distinct clusters of values (Figure 8.3, second row, right column). The first cluster accounts for 15 out of the 100 repetitions. The histogram of the second cluster is plotted separately next to the original.

8.2.3 Problems in Methodology

We now briefly discuss problems which have arisen so far in this study. We comment on initial investigations which have been carried out but do not describe them in detail due to them being preliminary. They provide an initial indication as to where the problems may be occurring, but are not conclusive.

The above simulations on the Donoho & Johnstone (1994) signals did not include a sequence generated from the Bumps signal. When generating a Poisson sequence with the underlying intensity being the Bumps signal, a large proportion of the original data is zero. We attempted to transform our signals as with the other intensities, but found the likelihood estimate to be unbounded.

Initial investigations suggest that when the number of zero values within the signal increase, the log-likelihood becomes unbounded (as α increases). In particular, it is the Jacobian component of the likelihood in (8.2.8) which becomes unbounded. The cause of this is unknown and simulating data with an increasing number of zero points would be a suitable next step. Also, consideration of the estimation of the Jacobian and the effect which zero points have on the HF transformed variables will further help to understand the reasons for this unbounded behaviour.

We also found that often, no values exist for the Jacobian for large values of α and for negative α (and in particular when the signal intensity is small). This is caused by the approximation of the Jacobian (8.2.5) becoming too large and being replaced by an infinite value in R. We discuss this further in Section 8.4.

The last problem we observed was with the Blocks intensity signal for the HFT$\underline{\alpha}$, where there were two distinct regions where the parameter was optimal (Figure 8.3, second row). Initial investigations suggest that there are local ‘peaks’ occurring when calculating the Jacobian, and these are being treated as optima. Again, further investigation of the causes of this within the Jacobian is required and is further discussed in Section 8.4.

8.3 Data-Driven Haar-Fisz Transformation for Gaussianisation

In this section we detail initial investigations into adaptations of the DDHFT for Gaussianisation. It is conceptually and computationally straightforward to apply the same maximum likelihood techniques dependent upon a parameter to fit a curve to local estimates of the mean and variance so that our general HF transforms become data-driven. We refer the reader to the algorithm for the DDHFT in Section 5.5.5 and that for the general HFT given in Section 8.2.3.

In the conversion from the original HF to DDHF transforms, the mean-variance function went from being fixed (and for Poisson data the identity transform) to functional, based on estimation from the data. Recall that the regression setup used is

$$\hat{\sigma}_i^2 = h(\mu_i) + \varepsilon_i,$$

and the smooth coefficients of the Haar transform act as pre-estimates of μ_i and the squared detail coefficients act as pre-estimates of σ_i^2. The estimate of h was found using non-decreasing isotonic regression from Johnstone & Silverman (2005a) and then used to form the *Fisz coefficients* $\mathbf{f}^j$ defined by

$$f^j = \frac{d^j}{h^{1/2}(s^j)}. \tag{8.3.10}$$

We wish to generalise the transformation to incorporate a Gaussianising parameter as with the constrained Haar-Fisz transform (HFT$\underline{\alpha}$). As with the HFT$\underline{\alpha}$, we modify the transform so that the denominator $h^{1/2}(s^j)$ is raised to the β instead of being fixed at $1/2$ (other possible modifications are discussed in Section 8.3.2). Under the same assumptions of Gaussianity as with the HFT$\underline{\alpha}$, we can again use maximum likelihood techniques (as detailed in Section 8.2.3) to choose β such that the resulting signal transformation is most Gaussian. The *Fisz coefficients* $\mathbf{f}^j$ thus become

$$f^j = \frac{d^j}{h^\beta(s^j)}, \tag{8.3.11}$$

where $h()$ is the estimate of the mean-variance function using isotone regression and β is chosen by maximising the log-likelihood function given in (8.2.8) (but with the obvious change in parameter lettering). We refer to this transformation as DDHFTβ in the remainder of this chapter.

We also considered a further transformation where the mean-variance function was instead estimated with kernel regression smoothing using a global plugin bandwidth (using the `lokern` package from R). See Brockmann *et al.* (1993) for more details. The method uses a kernel estimator to fit a function to the mean-variance relationship and is dependent upon a global bandwidth. Again, we selected this global bandwidth parameter using maximum likelihood techniques, as with the general HF transform. The larger the bandwidth, the smoother the mean-variance function will be. This method was limited as no matter what values the bandwidth took, the function estimate was still estimating *the* mean-variance function of the data. When we generalise to attempt to Gaussianise the data, we no longer wish to have the actual function which fits the data, we only want *a* mean-variance function which serves best to Gaussianise the data. Thus the transform did not compare well and we do not report results using this method.

8.3.1 Comparison with Box-Cox and the Constrained HFT

We compare our data-driven Haar-Fisz transforms using the same sets of intensity signals from Section 8.2.2 with both the ‘small’ and ‘large’ intensity range. The idea of our transformations is to produce data which can be represented as ‘signal plus noise’ where the noise is Gaussian. We can then apply a denoiser to estimate the signal and invert our transformation to obtain an estimate of the (known) underlying intensity.

For each of the signals which are corrupted with Poisson noise, we apply the Box-Cox, HFT$\underline{\alpha}$, DDHFTβ transforms. We do not compare the HFT$\boldsymbol{\alpha}$ due to the considerable amount

Signal	Small Intensity					
	HFT$\underline{\alpha}$		DDHFTβ		BC	
	Mean	SE	Mean	SE	Mean	SE
Constant	$\boxed{0.023}$	0.022	0.071	0.061	0.060	0.033
Blocks	$\boxed{0.558}$	0.073	1.329	0.557	0.616	0.096
Heavisine	$\boxed{0.312}$	0.131	1.120	0.480	0.340	0.089
Doppler	$\boxed{0.604}$	0.091	1.287	0.517	$\boxed{0.604}$	0.084

Table 8.6: Mean MSE and standard errors (SE) of small intensity signals for different transformations.

Signal	Large Intensity					
	HFT$\underline{\alpha}$		DDHFTβ		BC	
	Mean	SE	Mean	SE	Mean	SE
Constant	$\boxed{0.023}$	0.019	0.107	0.180	0.065	0.034
Blocks	15.622	2.925	47.640	57.72	$\boxed{8.432}$	1.838
Heavisine	$\boxed{9.636}$	1.071	13.221	5.837	11.715	2.169
Doppler	$\boxed{26.937}$	2.086	28.229	2.229	29.321	2.217

Table 8.7: Mean MSE and standard errors (SE) of large intensity signals for different transformations.

of time required for suitable optimisation, compared to the other transformations. Also, although the p-values of the KS test were not *as* non-significant for the HFT$\underline{\alpha}$ compared to the HFT$\boldsymbol{\alpha}$, many were still not significant at the 5% level. We estimate the underlying intensity using EbayesThresh wavelet thresholding from Johnstone & Silverman (2005a) (we first take the wavelet decomposition of the signals using the Haar wavelet for the constant and Blocks signals, and Daubechies least-asymmetric wavelet with 10 vanishing moments for the Doppler and Heavisine signals). We then invert our smooth signals back to the original data domain and compare the estimation of the underlying intensity with the known intensity using the mean square error (MSE) as defined in (7.5.16).

Table 8.6 and Table 8.7 give the mean MSE over 100 sample paths for the small and large intensities respectively. Note the further transformation which is included in the tables, DDHFTδ. We define this in Section 8.3.2.

Over all intensities and signals, the DDHFTβ performs poorly, having a much higher MSE than both the HFT$\underline{\alpha}$ and the Box-Cox transform (with the only exception being the ‘large’ Doppler intensity, where it is better than Box-Cox). The standard errors of the MSE values for the DDHFTβ are very large, suggesting the MSE values vary considerably. The

	Peak intensity=8		Peak intensity=128	
Signal	α	β	α	β
Constant	0.49	0.95	0.87	0.99
Blocks	0.86	4.54	1.57	2.02
Heavisine	0.88	4.45	0.77	1.68
Doppler	0.83	4.10	0.68	1.52

Table 8.8: Mean optimal parameters from the HFT$\underline{\alpha}$ and the DDHFTβ

mean optimal parameters for both Haar-Fisz transforms are given in Table 8.8. For the smaller intensity, we see the high parameter values for all functions except the constant signal. The optimisation algorithms had maximum parameter value for β of 5, which suggests that the likelihood function for these signals are unbounded. We have increased this limit and have again observed parameter values close to the maximum range.

We leave investigations into the unbounded likelihood as future work, but considering the poor performance of the DDHFTβ over the other signal (where the likelihood appears to be bounded), we instead consider alternative modifications.

8.3.2 Further Work: Other Models

We first consider the behaviour of the mean-variance estimate from both the HFT$\underline{\alpha}$ and the DDHFTβ, for when the likelihood appears bounded. We look at the constant signal of intensity 4 and use both methods to obtain a mean-variance function. Figure 8.4 shows the local estimates of the mean and variance, along with our two function estimates.

For this example, the optimal parameter values are $\alpha = 0.59$ and $\beta = 0.93$. The values of the isotone regression estimate not included in the plot increase up to a variance of 18. As the mean increases, the estimate from the DDHFTβ gets larger at a faster rate than the HFT$\underline{\alpha}$.

Over the range of values plotted, the mean-variance function for the HFT$\underline{\alpha}$ appears fairly straight. As a further model for our data-driven method, we propose a different transform where the modification to the *Fisz coefficients* is defined by

$$f^j = \frac{d^j}{\delta h^{1/2}(s^j)}. \tag{8.3.12}$$

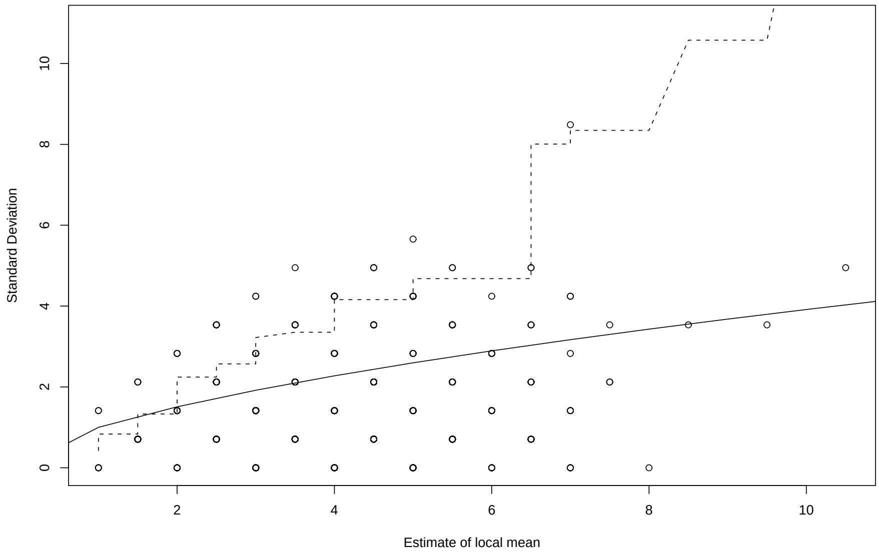


Figure 8.4: Local estimates of mean and variance of underlying constant signal. Solid line: $h()$ estimated using HFT$\underline{\alpha}$. Dashed line: $h()$ estimated using DDHFTβ. Maximum value for variance of DDHFTβ is 18.

This modification estimates the mean-variance curve using isotone regression as with the original DDHFT, but then then multiplies it by a constant, δ. We refer to this transformation as DDHFTδ in the remainder of this chapter. We include this further modification to compare the transformation to previous techniques and as a suggestion of next steps in improving the Gaussianisation of the DDHFT.

We calculate the mean square error for the same simulated signals in Section 8.3.1 with the same smoothing methods. The mean MSE and standard errors can be seen in Table 8.9. Comparing these results to the the small intensities in Table 8.6, we see that our modification DDHFTδ results in a lower MSE than any of the other transformations for all but the constant signal (where the standard error is large compared to the mean MSE). For the larger signal intensities, the DDHFTδ produces the smallest MSE for the Heavisine signal and the MSE is only 6 and 13 per cent higher than that of the HFT$\underline{\alpha}$ for the Doppler and Constant signals respectively. Furthermore, it should be noted the poor performance which we have shown the HFT$\underline{\alpha}$ to have for the Blocks signal, and the improvement (in terms of MSE) which the DDHFTδ shows (and is only 6% worse than the Box-Cox transform).

	Peak intensity=8		Peak intensity=128	
Signal	Mean MSE	SE	Mean MSE	SE
Constant	0.026	0.024	0.026	0.025
Blocks	$\boxed{0.520}$	0.069	8.900	1.859
Heavisine	$\boxed{0.250}$	0.064	$\boxed{9.462}$	1.060
Doppler	$\boxed{0.599}$	0.070	28.685	2.348

Table 8.9: Mean MSE and standard errors (SE) using the DDHFTδ for both large and small intensities. Boxes imply smaller mean MSE than other previous methods.

We note, however, that although the DDHFTδ appears superior to the other transformations, the likelihood is often unbounded. Nevertheless, the results presented here are promising and further work investigating the likelihood function is required to shed light on the behaviour of the transformation.

8.4 Conclusions and Future Work

In this chapter we proposed a Haar-Fisz transform which primarily attempts to Gaussianise data. We defined two types of such a transform, a constrained model in which the mean-variance relationship is assumed to be the same for all levels of the wavelet decomposition, and a general transform where a different relationship was sought for each level. We compared these transforms to the identity transformation and the Box-Cox transformation over known intensities which had been corrupted with Poisson noise. We compared Q-Q plots and p-values for the Kolmogorov-Smirnov test for Gaussianity and found that the general HFTα outperformed all other transforms except for the Blocks with the ‘large’ intensity range. Furthermore, our constrained HFT compared well to the other transforms and benefited from a considerably shorter computational time compared with the general transform.

It must be noted, however that although we conclude that our methods were ‘most Gaussian’ by the significance of the K-S tests, many of the other mean p-values were also significant at the 5% level. Further simulations should be caried out to further compare the performance of our methods.

Although the general HFTα outperformed other transforms, it is highly computationally intensive, with long execution times. The main cause of this was the use of approximation to the Jacobian matrix. An explicit form of this, or further approximations and simplifications would be greatly beneficial. A explicit form of the Jacobian matrix might also remove the anomalies found in the procedure where the estimation of the Jacobian is computationally set to infinity. Further investigation into the sensitivity of $l_{\max}$ for different values of α_j could result in the tolerance of the optimising procedure being reduced (if the required level of accuracy of $l_{\max}$ is not affected), resulting in much faster computational times. As an example, we have observed that by reducing the number of decimal places to which $l_{\max}$ has been deemed to converge by one, has reduced computational time by half. This change appears to mostly effect the parameters of the 'coarser' levels of the HFT$\boldsymbol{\alpha}$.

We further proposed data-driven Haar-Fisz transforms for Gaussianisation, which raised the mean-variance function to an unknown parameter β, and a further model which multiplied the square root of the estimated function by an unknown parameter. Both parameters were again found using maximum likelihood techniques.

Our transforms were compared to others in terms of mean square error of intensity estimation of a known underlying signal, following Gaussianisation. Again, our methods were compared to that of Box-Cox and also to the HFT$\underline{\alpha}$ and for both the 'small' and 'large' intensity signals used previously in the chapter.

For the small signals, the DDHFTδ had a smaller MSE than the other transforms over all signals except the constant (where HFT$\underline{\alpha}$ performed better). For the Doppler signal, however, the improvement from the DDHFTδ was less than 1% smaller MSE compared to the HFT$\underline{\alpha}$ and Box-Cox transforms. The DDHFTδ showed a 7% and 24% reduction in mean MSE for the Blocks and Heavisine signals, respectively, compared to the next best performing method.

There were mixed results when comparing transforms for the larger intensity signals, with the HFT$\underline{\alpha}$ having the smallest mean MSE for two of the signals, and the DDHFTδ and Box-Cox both having the smallest mean MSE on one occasion. For the large signal intensity, the the improvement the 'best' performing transform has over the next best method is never more than a 5% reduction in mean MSE.

Another observation was that for the smaller intensity signal, the likelihood within the parameter estimation appeared to be unbounded. Further understanding of the nature of the likelihood estimation and how it is linked with the estimation of the Jacobian matrix is required for this work to progress. Due to these uncertainties, we remind the reader that methods and results presented in this chapter are to be considered as preliminary.

Future work could also consider other modifications of both the Haar-Fisz and data-driven Haar-Fisz transformations. Our second modification of the data-driven DDHFTδ suggests that more general transformations could improve Gaussianisation performance when the mean-variance function is both known and unknown. Furthermore, extension of the data-driven transforms to estimate this relationship for each level of the wavelet decomposition would be an obvious extension and likely (as seen by HFTα) to improve performance.

Chapter 9

Conclusions and Further Work

This chapter provides a summary of the work outlined within this thesis. We consider the main work from the research chapters, discussing the advantages and disadvantages of the methodology as well as discussing ideas for future work.

9.1 Backbench Opinion in the House of Commons using EDMs

Work in Chapters 3 and 4 reintroduced the idea of using Early Day Motions (EDMs) as a measure of backbench opinion. A new cohesion measure was defined in Chapter 3 which took into account the asymmetric signing of EDMs and was used to chart the cohesion of the three main political parties over the course of the 2005/06 parliamentary session. As a means of calibration, a basic probabilistic model was derived to create simulated cohesion levels to compare with the observed levels. The cohesion was then interpreted in terms of these calibration levels and compared against real life events with which a perceived party unity could be ascertained.

Extensions to this work could be the development of the calibration level to more accurately model the party cohesion and even forecasting the cohesion of the parties.

The cohesion measure was then used for feature selection. Having classified EDMs into different types and assigned each a weight, the cohesion measure was maximised and minimised by altering the weights. Those EDMs given high or low weights were deemed to cause party cohesion and separation respectively.

EDMs were used for feature selection again in Chapter 4. Using the multi-dimensional scaling solution and linear discriminant analysis to assign each MP to a party, issues were sought which caused the parties to appear more disjoint. An optimising criteria was formed by summing the number of erroneously classified MPs. This acted as a measure of overlap between parties. Similar to the moving time window of the cohesion levels, this feature selection looked at windows of 200 EDMs moving across the session. This allowed for quicker computation and a less static feel of the data (although with an overlap of 100 EDMs it was still somewhat static).

The second use of EDMs for feature selection was not as successful as the first. This was due to the complexity of the methodology and subsequent computational time required for results. We presented preliminary results and suggested possible steps to increase efficiency of the computation. These included increasing the optimisation search area to uncover more optimal solutions, or decreasing the size of the data set to reduce the number of variables. A mixture of the two, as well as an increase in computational power or efficiency would help, but then it may also be tempting to increase the number of windows over the session to get a less static feel to the output.

Both examples of feature selection indicate the wealth of information contained in the EDM data set. The classification of EDMs into different issues was to a certain extent subjective and considering secondary classification issues would increase the information gained from the data.

Chapter 4 also included a brief investigation into the effect that propensity to sign has on an MPs position in the scaling solution. Initial investigations suggest that this is indeed prevalent up to and including the second dimension, but that more dimensions are needed to effectively capture the information within the data.

9.2 Coalition Mortality Rates in Iraq

Chapter 6 introduced data on the number of coalition deaths in the recent Iraqi conflict. We showed how current methods of intensity estimation are not suitable for the data as they assume a certain degree of Gaussianity, and we proposed using the recently developed data-driven Haar-Fisz transform (DDHFT) to variance stabilise the data. We then applied different smoothing methods to obtain intensity estimates. Our methods were shown to perform well using statistical tests for Gaussianity and heteroskedascity and they also outperformed the much used Box-Cox transform. We concluded that the mean level of intensity increased until about January 2005, leveled off until around June 2005 and then slightly decreased before leveling off again until the end of the series. We also tentatively concluded that the number of non-hostile deaths is inversely related to the intensity of hostile deaths.

Further extensions to this analysis could be to look at clusters within the data. There are periods where deaths appear to increase in intensity over a few days and then return to a lower level. Time series models could also be constructed based on the number of attacks, rather than the casualty rate. The ratio of attacks to casualties would help build up a clearer picture of the intensity of conflict. Such information, however, would be hard to accurately obtain.

As mentioned within the chapter, using methods from Spirling (2007) could also be used to pick out areas of increase intensity of deaths as well as providing a method of comparing coalition and civilian deaths. Similarity between these data sets could provide further measures regarding the conflict.

9.3 DDHFT for Negative Data

We detailed the central England temperature (CET) data set in Chapter 7 and examined its mean-variance relationship. We observed separate regions of positive and negative correlation and modified the Haar-Fisz and data-driven Haar-Fisz transforms so that this data could be suitably transformed and variance stabilised. We proposed two versions of the negative DDHFT, which depended on different assumptions being made about the distribution of the positive and negative data points, and suggested a bootstrap test for deciding between the two. We used our methodology to transform the CET data and then used smoothing methods to obtain an underlying intensity estimate. We concluded that since 1970, the temperature appears to be increasing at a faster rate than previously.

We compared our transforms to the Box-Cox transform using a set of test signals which were transformed and then smoothed. The mean square error of the smooth signal compared to the known underlying intensity was shown to be smaller for the DDHFT methods. A natural extension of this method would be to automatically select the 'turning point' of the mean-variance relationship (the point at which it changed from a negative to positive relationship). Further modifications could be made which allow for a more flexible mean-variance function.

9.4 Maximum Likelihood Techniques for Haar-Fisz Transforms

In Chapter 8 we looked at Haar-Fisz transforms from the viewpoint of wishing to primarily Gaussianise the data, as opposed to stabilising the variance of it. We introduced a parameter into the Haar-Fisz transform which replaced the square root function (for Poisson data) and used a similar derivation of the maximum likelihood estimator as with the Box-Cox transform. We proposed two such transforms: one where the parameter is constant over all wavelet levels and one where it was allowed to vary. Considering Q-Q plots and p-values from Kolmogorov-Smirnov tests for Gaussianity, our transforms were shown to perform well compared to Box-Cox. The transform which allowed for different parameter for each wavelet decomposition level outperformed the other transforms, but had a much longer computational time.

We also showed initial work into adapting the methods for a data-driven transform. We suggested some possible transforms and compared their performance (in terms of mean square error of a test signal estimate) to that of the one parameter Haar-Fisz and Box-Cox transforms. Our methods did not perform well and we suggested a further, simpler model. Although this appeared to perform well, it had an unbounded likelihood.

The methods outlined in this chapters are initial investigations and there are many avenues left to explore. We highlighted the computational time that the HF transform took when the parameters are assumed to be different for each level of the wavelet decomposition. Considering the sensitivity of the optimisation criteria to the likelihood, and the effect this has on the parameters, may provide some efficiencies or simplifications to the methods. Also, our modifications to the DDHFT showed that a simpler translation of the mean-variance function may indeed produce better results.

We have shown how the transforms can Gaussianise count data more effectively than the Box-Cox transform. Many problems which were found focused on the estimation of the Jacobian. Further development of an algebraic form of this, or simplifications within the derivation of the likelihood function may increase the effectiveness of these transformations.

Appendix A

A.1 Examples of Early Day Motions

A.1.1 Debated EDM

A Motion put down by Rt Hon Margaret Thatcher, the then Leader of the Opposition, censuring the Government. When this Motion was debated on 28 March, it was agreed to, leading to a General Election.

EDM Number: 351

Date: 22.03.1979

NO CONFIDENCE IN HER MAJESTYS GOVERNMENT

That this House has no confidence in Her Majestys Government.

Total Number of Signatures: 6

A.1.2 Early Day Motion 1646

Date: 14.02.2006

SMOKING IN THE HOUSE OF COMMONS

Tabled by: Julia Goldsworthy

That this House notes that right hon. and hon. Members voted to ban smoking in all public places including private members' clubs on 14th February 2006; further notes that the will of the House may not apply in the House itself since it is a royal palace; further notes that this means that staff working in the Smoking Room could still be exposed to the harmful effects of second-hand smoke; and calls for this anomaly to be rectified by the House authorities as soon as possible.

Total Number of Signatures (at time of writing): 69

A.2 Obtaining and Classifying Data

Data was downloaded from the internet by using Unix functions and converted into a suitable form using Perl scripts. Classification of EDM types was performed by hand by the authors or a team of coders under close supervision by the authors. Two coders independently classified each session into primary and secondary (where appropriate) issues. Where the two agreed on a primary issue, or a primary issue from one and a secondary from the other, the corresponding classification was used. Where there was disagreement, a third classifier was used and the process repeated. Where no agreement could be found from three coders, the authors took the final decision.

Appendix B

B.1 Smoothing Methods Used

S1 Haar decimated wavelet shrinkage using the EbayesThresh threshold choice of Johnstone & Silverman (2005a).

S2 Kernel regression smoothing using a local plugin bandwidth (using `lokern` package from R). See Brockmann *et al.* (1993).

S3 Translation-invariant, basis-averaging over complex-valued Lina-Mayrand wavelet (with 5 vanishing moments) using multiwavelet style threshold as described by Barber & Nason (2004).

B.2 Empirical Bias Results for the Data-Driven Haar-Fisz Transform for Finite Sample Sizes

In the following sections we individually detail the signals which we use to compare the bias of the two transformations. We first give a general overview of the procedure.

We define a sequence of intensities λ_i for $i = 1, \ldots, N$, where N is a power of 2. We use these to generate our simulated data **X**, defined by:

$$X_i = \text{Poi}(\lambda_i), \tag{B.1}$$

for $i = 1, \ldots, N$.

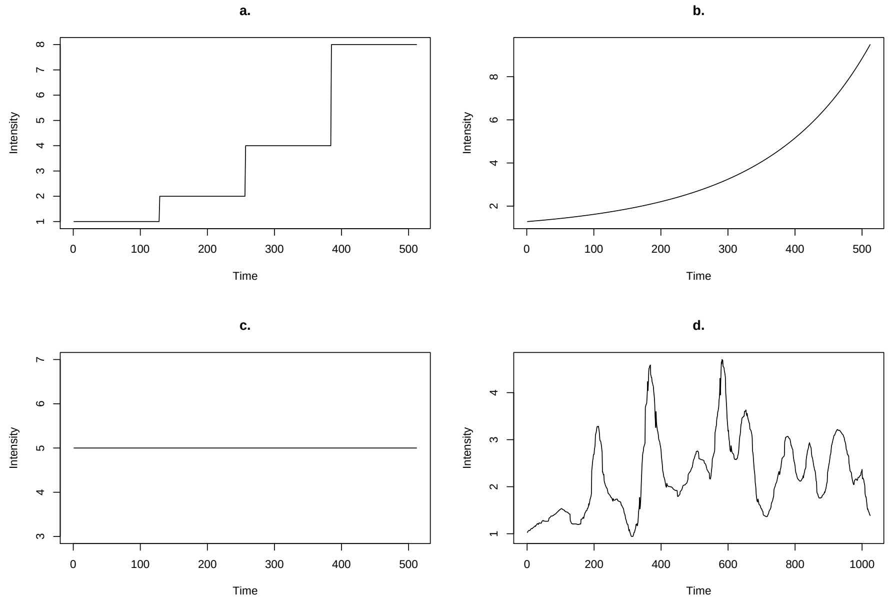


Figure B.1: Underlying intensity of test signals

For each signal, we take both the DDHF and Box-Cox transformations. (For DDHF, we use functions from the R package `DDHFm`. For Box-Cox, we add a constant of 1 to the data to ensure positivity and use the functions from the R package `car` to estimate the parameter and to transform the data.)

Kernel regression smoothing using a local plugin bandwidth, from Brockmann *et al.* (1993) is then used to smooth the transformed data (using the `lokern` package from R). We next invert both of our sequences and compare our estimates to the known underlying intensities.

We repeat this process 100 times and take the mean of the intensity estimations for each signal. The bias is the difference between the known signal intensity and the estimated one. As these values can be negative, we report the sum (over all points) of the square of the bias.

We next describe our intensity signals and resulting bias calculations.

B.2.1 Piecewise Constant Intensity

We first use an underlying intensity which is a piecewise constant of length 512. The constant regions are equal in length and take the values 1,2,4, and 8 respectively. The intensity can be seen in figure B.1(a).

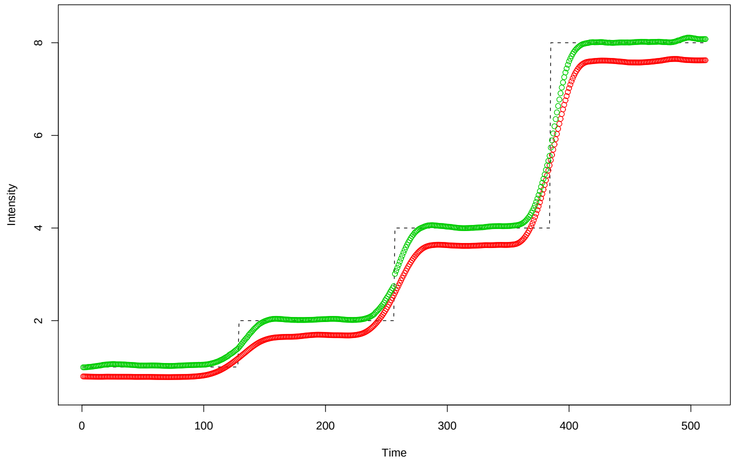


Figure B.2: Intensity estimates of piecewise constant function. Dashed line: known intensity. Green: DDHF. Red: Box-Cox.

We simulate 100 time series and use the method described above to obtain a mean signal intensity, shown in figure B.2. The sums of squared bias for Box-Cox and DDHF transforms are 127.08 and 59.61 respectively.

B.2.2 Exponential Intensity

For this simulation, our underlying intensity takes the values

$$\lambda_i = e^{k_i},$$

where k_i is the square of the sequence from 0.5 to 1.5 of length 512, shown in figure B.1(b).

We once again create a 100 sequences of Poisson random variables with intensity λ_i, which we transform, smooth and invert. Figure B.3 shows the mean of intensity estimates for our two transformations.

The sums of squared bias for Box-Cox and DDHF transforms are 40.0 and 0.47 respectively.

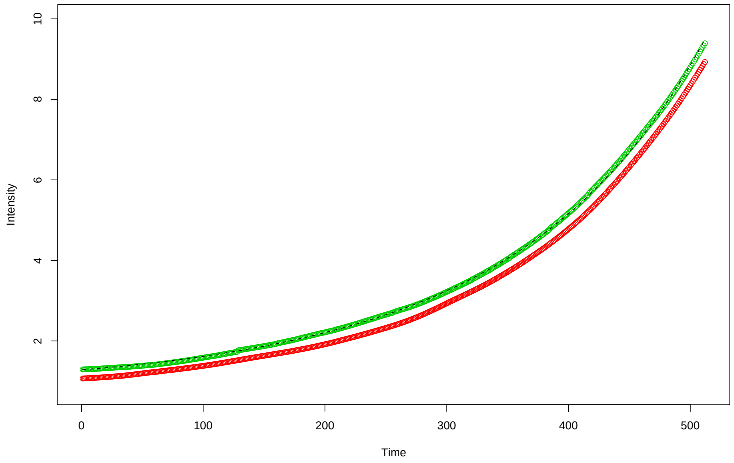


Figure B.3: Intensity estimates of exponential function. Dashed line: known intensity. Green: DDHF. Red: Box-Cox.

B.2.3 Constant Intensity

For our third simulation, we create a Poisson signal of constant intensity $\lambda_i = 5$, of length 512. Figure B.1(c) shows the mean estimated intensity for both transformations. The sum of squares of the bias is 18.36 and 0.39 for the Box-Cox and DDHF transforms respectively.

B.2.4 Iraq Data

Finally, we consider the bias of the transformations on data taken from Nason & Bailey (2008). The data is an estimation of the underlying intensity of daily mortality rates amongst coalition forces in Iraq for the 1024 days since the beginning of the invasion in March 2003. Of the three methods of estimation used in the paper, we use the results from the **S2** method, using kernel regression from Brockmann *et al.* (1993).

The intensity can be see in figure B.1(d.) We once again use this signal as our λ_i to create 100 simulated Poisson signals. Figure B.5 shows the mean estimated intensity of the signals from both transformations. The sum of squared bias for the Box-Cox and DDHF transforms is 117.81 and 30.16 respectively.

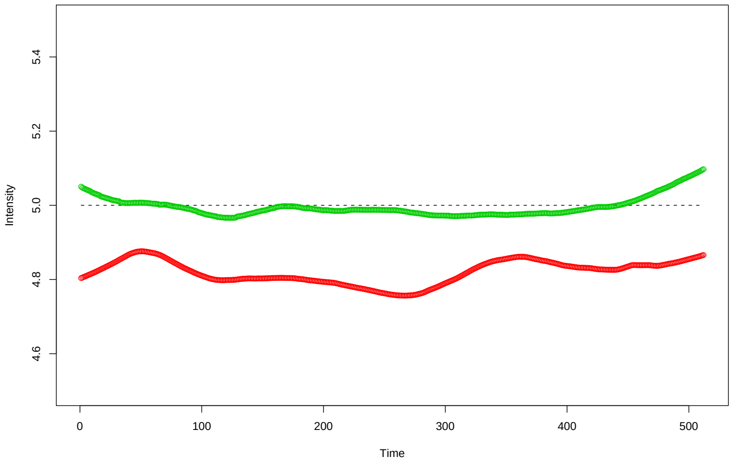


Figure B.4: Intensity estimates of constant function. Dashed line: known intensity. Green: DDHF. Red: Box-Cox.

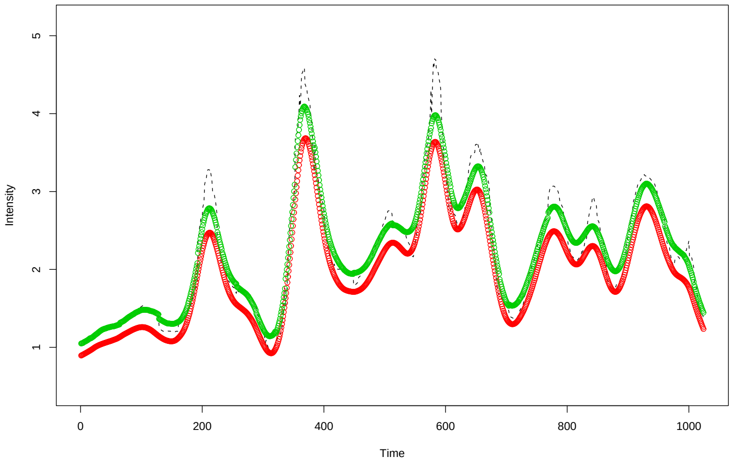


Figure B.5: Intensity estimates of Iraq data. Dashed line: known intensity. Green: DDHF. Red: Box-Cox.

B.2.5 Conclusions

For all four of our test sequences, the empirical transformation bias from the DDHF transform is significantly lower than that of the Box-Cox transformation. Furthermore, Theorem 3 from Fryzlewicz (2007) shows that the DDHF procedure, using a Nadaraya-Watson estimate of h, is asymptotically unbiased.

Appendix C

C.1 Convergence of $\hat{t}_i^j$

In this appendix we prove the convergence of $t_{k_l}^j$ as $l \to \infty$ and derive its limit, as stated in 7.4.9.

From equation (7.4.3), we have

$$t_k^j = \frac{|\hat{s}_{2k-1}^{j+1}| + |\hat{s}_{2k}^{j+1}|}{2}.$$

Substituting for the values given in (7.4.10) and (7.4.11) gives

$$t_k^j = \left(\left| \hat{s}_k^j + \hat{f}_k^j \sqrt{t_j^k} \right| + \left| \hat{s}_k^j - \hat{f}_k^j \sqrt{t_j^k} \right| \right) /2. \qquad \text{(C.1)}$$

For clarity, we remove annotation from around the variables as they do not change within our proof. We add the iteration index l to each of the t's, where $l \in \mathbb{N}$. Thus, equation C.1 becomes:

$$t_l = \left(\left| s + f\sqrt{t_{l-1}} \right| + \left| s - f\sqrt{t_{l-1}} \right| \right) /2. \qquad \text{(C.2)}$$

We next prove that the sequence t_l converges to the limit T, as $l \to \infty$. Furthermore, we show that T takes the values

$$T = \begin{cases} |s|, & \text{if } (f)^2 \leq |s|, \\ |f\sqrt{t}|, & \text{if } (f)^2 \geq |s| \end{cases} \qquad \text{(C.3)}$$

where t is the value calculated in the 'forward' step in (7.4.3) (and given above). We write

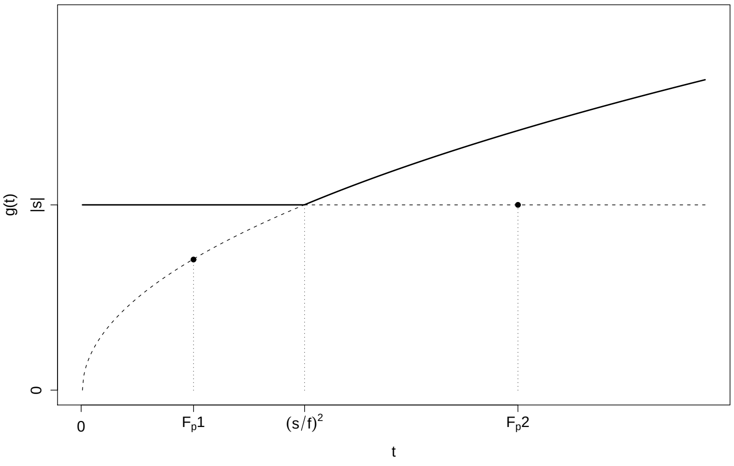


Figure C.1: Possible values of g(t). Solid line: values of g(t) within function constraints. Dashed line: possible values of g(t) which lie on the components of g(t), but outside the constraints.

equation C.2 in functional form to give

$$g(t) = \left(\left|s + f\sqrt{t}\right| + \left|s - f\sqrt{t}\right|\right)/2. \tag{C.4}$$

This is equivalent to

$$g(t) = \begin{cases} |s| & \text{if } s \geq f\sqrt{t}, \\ |f\sqrt{t}| & \text{if } s < f\sqrt{t}, \end{cases} \tag{C.5}$$

and is plotted as a solid line in figure C.1. $g(t)$ is said to have converged when $|g(t)-t| < \varepsilon$ for some small $\varepsilon > 0$.

The function $g(x)$ is made up of two components. The first, when $s \geq f\sqrt{t}$ is a constant value $g(t) = |s|$. So providing the constraints hold, $g(t) = t = |s|$ is constant, and has thus converged.

The second component of $g(t)$ is for the condition $s < f\sqrt{t}$ and in which case, $g(t)$ takes the value $|f\sqrt{t}|$. It can be shown that the square root function has an attracting fixed point (see for example Strogatz (1994)) and thus will converge so that $g(t) = t$.

We have therefore shown that t_l in (C.2) converges as $l \to \infty$. We next show that the limit of this convergence, denoted by T, takes the values given in (C.3).

Clearly, if $s \geq f\sqrt{t}$, the only value T can take is $|s|$. If $s \leq f\sqrt{t}$, then the attracting fixed point which is the limit T, will lie on the curve $g(t) = |f\sqrt{t}|$. Suppose this point is such that $g(t) = t = \alpha$. Then

$$\begin{aligned} & \alpha &=& \ |f\sqrt{\alpha}|, \\ \Rightarrow & \alpha^2 &=& \ f^2\alpha, \\ \Rightarrow & \alpha^2 - f^2\alpha &=& \ 0, \\ \Rightarrow & \alpha &=& \ 0 \text{ or } f^2. \end{aligned}$$

So $g(t) = t = f^2$. We also note the situation when $|s| = f^2$. In this scenario, we have that $|f\sqrt{t}| = |s|$ so the fixed points are identical. Therefore, as $l \to \infty$, t_l converges and takes the values given in (C.3).

In the last part of this proof we check that points which lie on either $g(t) = |s|$ or $g(t) = |f\sqrt{t}|$, but are not within the limits given in (C.5) are not attractive fixed points (so $g(t)$ has not converged). For an initial point t_1, let us suppose the latter so that $g(t_1) = |f\sqrt{t_1}|$. Say, for example, that $g(t_1)$ is point Fp1 in figure C.1, so that $t_1 = f^2$. Therefore we also have $t_1 \leq (s/f)^2$ so that

$$\begin{aligned} f^2 \leq s^2/f^2 \Rightarrow f^4 &\leq s^2, \\ \Rightarrow f^2 &\leq |s|. \end{aligned}$$

But from (C.3), if $f^2 \leq |s|$ then $g(t_1)$ must take the value $|s|$, so t_1 has not converged.

Lastly, if $g(t_1) = |s|$ but $t_1 \geq (s/f)^2$, we have $|s| \geq (s/f)^2$. An example of this is indicated by point Fp2 on figure C.1. This implies that $f^2 \geq |s|$. But if this is so, the attracting fixed point is at $g(t_l) = |f\sqrt{t_l}|$ and thus t_1 has not converged.

Therefore, any point which has value $g(t) = |s|$ or $g(t) = |f\sqrt{t}|$ but for which the constraints in (C.5) but do not hold, will converge to the values given in (C.3).

Bibliography

Al-Osh, M., & Alzaid, A. 1987. First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, **8**, 261–275.

Al-Osh, M., & Alzaid, A. 1988. Integer-valued moving average (INMA) process. *Statistical Papers*, **29**, 281–300.

Andrews, D.F., & Herzberg, A.M. 1985. *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer-Verlag.

Anscombe, F. 1948. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, **35**(3–4), 246–254.

Atkinson, A.C. 1987. *Plots, Transformations, and Regression*. Oxford: Oxford University Press.

Attiná, F. 1990. The voting behaviour of the European Parliament members and the problem of Europarties. *European Journal of Political Research*, **18**, 557–579.

Aubury, M., & Luk, W. 1995. Binomial Filters. *The Journal of VLSI Signal Processing*, **12**(1), 35–50.

Bailey, D., & Nason, G.P. 2008. Cohesion of major political parties. *British Politics*. Note: To appear.

Baker, N. (Liberal Democrat Shadow Environment Secretary). 2004. http://www.politics.co.uk/issueoftheday/norman-baker-proper-investment-in-renewables-would-rule-out-need-nuclear-power-$368309$367007.html. September.

Barber, S., & Nason, G.P. 2004. Real nonparametric regression using complex wavelets. *Journal of the Royal Statistical Society B*, **66**, 927–939.

BBC Website. 2007. `http://news.bbc.co.uk/1/hi/world/middle_east/737483.stm`.

Berrington, H. 1968. Partisanship and dissidence in the nineteenth-century House of Commons. *Parliamentary Affairs*, **21**(4), 338–374.

Berrington, H. 1982. *The Politics of the Labour Party*. London: George Allen & Unwin. Chap. 3.

Berrington, H.R. 1973. *Backbench opinion in the House of Commons, 1945-55*. Oxford: Pergamon.

Box, G.E.P., & Cox, D.R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society B*, **26**, 211–246.

Breiman, L., & Friedman, J. 1985. Estimating optimal transformations for mulitiple regression and correlation. *Journal of the American Statistical Association*, **80**, 580–598.

Breusch, T., & Pagan, A. 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, **47**, 1287–1294.

Brockmann, M., Gasser, T., & Herrmann, E. 1993. Locally adaptive bandwidth choice for kernel regression estimators. *Journal of the American Statistical Society*, **88**, 1302–1309.

Bromhead, P.A. 1962. Backbench opinion in the House of Commons, 1955-59 - Finer S.E., Berrington, H.B., Bartholomew, D.J. *Sociological Review*, **10**, 349–351.

Burrus, C., Gopinath, R., & Guo, H. 1998. *Introduction to Wavelets and Wavelets Transformations: A Primer*. New Jersey: Prentice-Hall.

Cameron, A., & Trivedi, K. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Chang, W. 1983. On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, **32**(3), 267–275.

Chatfield, C., & Collins, A.J. 1996. *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Childs, S., & Withey, J. 2004. Women representatives acting for women: sex and the signing of Early Day Motions in the 1997 British parliament. *Political Studies*, **52**, 552–564.

CNN Website. 2007. *U.S. and Coalition POW/MIA*. `http://edition.cnn.com/SPECIALS/2003/iraq/forces/pow.mia/`.

Conover, W.J. 1971. *Practical Nonparametric Statistics*. New York: Wiley.

Conservative Party. 2001. *2001 Conservative Party General Election Manifesto: Time for Common Sense*. London: Conservative Party.

Conservative Research Department. 2004. *Conservative Party Disability Consultation*. London: Conservative Party.

Conte, A. 2005. *Security in the 21st Century. The United Nations, Afghanistan and Iraq*. England: Ashgate.

Cowley, P., & Stewart, M. 1997. Sodomy, slaughter, Sunday shopping and seatbelts - free votes in the House of Commons, 1979 to 1996. *Party Politics*, **3**(1), 119–130.

Cox, G. 1987. *The Efficient Secret: The Cabinet and the Development of Political Parties in Victorian England*. Cambridge: Cambridge University Press.

Cox, T.F., & Ferry, G. 1993. Discriminant analysis using non-metric multidimensional scaling. *Pattern Recognition*, **26**(1), 145–153.

Cromwell, V. 1982. Mapping the political world of 1861: A multidimensional analysis of House of Commons' division lists. *Legislative Studies Quarterly*, **7**, 281–297.

Crossman, R.H.S. 1961. How poor are the poor? *Manchester Guardian*, 15 December.

Daubechies, I. 1992. *Ten Lectures on Wavelets*. Philadelphia: SIAM.

Davison, A.C., & Hinkley, D.V. 1997. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

Donoho, D.L., & Johnstone, I.M. 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Fellows, Sir E. 1962. Backbench opinion in the House of Commons, 1955-59 by Finer S.E., Berrington, H.B., Bartholomew, D.J. *Parliamentary Affairs*, **15**, 244–245.

Finer, S.E., Berrington, H.R., & Bartholomew, D.J. 1961. *Backbench Opinion in the House of Commons, 1955-59*. Oxford: Pergamon.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Fisz, M. 1955. The limiting distribution of a function of two independent random variables and its statistical application. *Colloquium Mathematicum*, **3**, 138–146.

Fletcher, R., & Powell, M.J.D. 1963. A rapidly convergent decent method for minimization. *Computer Journal*, **6**(2), 163–168.

Franklin, M.N., & Tappin, M. 1977. Early Day Motions as unobtrusive measures of backbench opinion in Britain. *British Journal of Political Science*, **7**, 49–69.

Friedman, J., & Meulman, J. 2004. Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society B*, **66**, 815–839.

Friedman, J., & Tibshirani, R. 1984. The Monotonic Smoothing of Scatterplots. *Technometrics*, **26**(3), 243–250.

Fryzlewicz, P. 2007. *Data-driven wavelet-Fisz methodology for nonparametric function estimation*. Submitted for publication.

Fryzlewicz, P., & Nason, G.P. 2004. A Haar-Fisz algorithm for Poisson intensity estimation. *Journal of Computational and Graphical Statistics*, **13**, 621–638.

Fryzlewicz, P., Sapatinas, T., & Subba Rao, S. 2006. A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, **93**, 687–704.

Fryzlewicz, P., Delouille, V., & Nason, G.P. 2007. GOES-8 X-ray sensor variance stabilization using the multiscale data-driven Haar-Fisz transform. *Journal of the Royal Statistical Society C*, **56**, 99–116.

Gnanadesikan, R., Kettenring, J.R., & Maloor, S. 2007. Better alternatives to current methods of scaling and weighting data for cluster analysis. *Journal of Statistical Planning and Inference*, **137**(11), 3482–3496.

Hazan, R.Y. 2005. *Cohesion and Discipline in Legislators*. London and New York: Routledge.

Heinen, A. 2003. Modelling time series count data: an autoregressive conditional Poisson model. *Core Discussion Paper No. 2003-66*.

Hencke, D. 2006. Poll cash race leads to secret deals. *The Guardian*, 14th March.

Hill, M.O., & Gauch, H.G. 1980. Detrended correspondence analysis: an improved ordination technique. *Vegetatio*, **43**, 47–58.

Hix, S., Noury, A., & Roland, G. 2005. Power to the parties: cohesion and competition in the European Parliament, 1979–2001. *British Journal of Political Science*, **35**, 209–234.

Hoiland, K., Laane, C.M.M., & Medbo, J.I. 2004. Multivariate analysis of materials found on a sentenced man and on the scene of the crime. *Law, Probability and Risk*, **3**(3–4), 193–209.

House of Commons Information Office. 2003a. *House of Commons Factsheet P3*. London: Office of Public Sector Information.

House of Commons Information Office. 2003b. *House of Commons Factsheet P9*. London: Office of Public Sector Information.

Howard, A. 1962. Backbench opinion in the House of Commons, 1945-55. *New Statesman*, 12 January.

Hurst, G. 2006. *Charles Kennedy: A Tragic Flaw*. London: Politico.

Jansen, M. 2001. *Noise Reduction by Wavelet Thresholding*. New York: Springer.

Jansen, M. 2006. Multiscale Poisson data smoothing. *Journal of the Royal Statistical Society Series B*, **68**, 27–48.

Johnstone, I.M., & Silverman, B.W. 2004. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, **32**(4), 1594–1649.

Johnstone, I.M., & Silverman, B.W. 2005a. EbayesThresh: R programs for empirical Bayes thresholding. *Journal of Statistical Software*, **12**(8), 1–38.

Johnstone, I.M., & Silverman, B.W. 2005b. Empirical bayes selection of wavelet thresholds. *The Annals of Statistics*, **33**(4), 1700–1752.

Jung, R.C., Kukuk, M., & Liesenfield, R. 2006. Time series of count data: modeling, estimation and diagnostics. *Computational Statistics and Data Analysis*, **51**, 2350–2364.

Kendall, D.G. 1971. Seriation from abundence matrices. *In:* Hodson, F.R., Kendall, D.G., & Tautu, P. (eds), *Mathematics in the Archaeological and Historical Sciences*. UK: Edinburgh University Press.

Kendall, M., Stuart, A., & Ord, J. 1983. *The Advanced Theory of Statistics*. 4 edn. Vol. 3. London & High Wycombe: Charles Griffin & company Limited.

Krzanowski, W., & Marriott, F. 1995. *Kendall's Library of Statistics 2, Multivariate Analysis volume 2*. Arnold.

Leece, J., & Berrington, H. 1977. Measurements of backbench attitudes by Guttman scaling of Early Day Motions: a pilot study, Labour, 1968–69. *British Journal of Political Science*, **7**, 529–541.

Liberal Democrats. 2001. *Freedom, justice, honesty: manifesto for a liberal and democratic Britain: general election 2001*. London: Liberal Democrat Party.

Linton, O., Chen, R., Wang, N, & Hardlem, W. 1997. An analysis of transfromations for additive nonparametric regression. *Journal of the American Statistical Association*, **92**, 1512–1521.

Lloyd, T. 1977. Backbench opinion in the House of Commons, 1945-55 - Berrington, H. *Canadian Historical Review*, **158**, 242–243.

Lowell, A.L. 1919. *The Government of England*. Vol. 2. New York: Macmillian.

MacDonald, I., & Zucchini, W. 1997. *Hidden Markov and Other Models for Discrete-Valued Time Series*. New York: Chapman and Hall.

Manley, G. 1953. The mean temperature of central England, 1698–1952. *Quarterly Journal of the Royal Meteorological Society*, **79**, 242–261.

McKenzie, E. 1988. Some ARMA models for dependent sequences for Poisson counts. *Advances in Applied Probability*, **20**, 822–835.

McLean, I. 1995. *Party, Parliament and Personality: Essays Presented to Hugh Berrington*. New York: Routledge. Chap. 8.

Mitchell, M. 1998. *An Introduction to Genetic Algorithms*. Cambridge, Mass.: MIT Press.

Motakis, E., Nason, G., Fryzlewicz, P., & Rutter, G. 2006. Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach. *Bioinformatics*, **22**, 2547–2553.

Nadaraya, E.A. 1964. On estimating regression. *Theory of Probability and Its Applications*, **10**, 186–190.

Nason, G.P. 2001. Early Day Motions: exploring backbench opinion during 1997-2000. *Technical Report, Dept. Mathematics Bristol. 01:11*.

Nason, G.P. 2008. *Wavelet Methods in Statistics with R*. New York: Springer-Verlag.

Nason, G.P., & Bailey, D. 2008. Estimating the intensity of conflict in Iraq. *Journal of the Royal Statistical Society Series A*.

Nelder, J.A., & Mead, R. 1965. A simplex-method for function minimization. *Computer Journal*, **7**(4), 308–313.

Owens, J.E. 2003. Explaining party cohesion and discipline in democratic legislatures: purposiveness and contexts. *The Journal of Legislative Studies*, **9**(4), 12–40.

Panofsky, H., & Brier, G. 1968. *Some Applications of Statistics to Meteorology*. Pennsylvania: University Park.

Parker, D., Legg, T., & Folland, C. 1992. A new daily central England temperature series, 1772-1991. *International Journal of Climatology*, **12**, 317–342.

Performance and Innovation Unit. 2002. *The Energy Review*. London: Labour Party.

Podani, J, & Milklos, I. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*, **83**(12), 3331–3343.

Poole, K. T. 2005. *Spatial Models of Parliamentary Voting*. New York: Cambridge University Press.

Poole, K.T., & Rosenthal, H. 1997. *Congress: a political-economic history of roll call voting*. New York: Oxford University Press.

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. 1992. *Numerical Recipes in C*. 2 edn. Cambridge: Cambridge University Press.

Rahat, G. 2007. Determinants of party cohesion: evidence from the case of the Israeli Parliament. *Parliamentary Affairs*, **60**(2), 279–296.

Read, M., Marsh, D., & Richards, D. 1994. Why did they do it? Voting on homosexuality and capital punishment in the House of Commons. *Parliamentary Affairs*, **47**(3), 374–386.

Rice, S.A. 1928. *Quantitative Methods in Politics*. New York: Knopf.

Richards, P.G. 1962. Backbench opinion in the House of Commons, 1955-59 by Finer S.E., Berrington, H.B., Bartholomew, D.J. *Public Administration*, **40**(3), 337–339.

Roberts, L., Lafta, R., Garfield, R., Khudhairi, J., & Burnham, G. 2004. Mortality before and after the 2003 invasion of Iraq. *The Lancet*, **364**, 1857–1864.

Robinson, V. 2003. *Spreading the 'Burden'?: A Review of Policies to Dispurse Asylum Seekers and Refugees*. Bristol: The Policy Press.

Simonoff, J.S. 1996. *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Spirling, A. 2007. "Turning points" in the Iraq conflict: reversible jump Markov chain Monte Carlo in political science. *The American Statistican*, **61**(4), 315–320.

Spirling, A., & McLean, I. 2006. The rights and wrongs of roll calls. *Government and Opposition*, **41**(4), 581–588.

Strogatz, S.H. 1994. *Nonlinear Dynamics and Chaos*. USA: Westview Press.

The Economist. 2006 (August). *Northern Ireland*. http://www.economist.com /research/backgrounders/displaybackgrounder.cfm?bg=832536.

The Poynter Institute. 2001. *September 11, 2001*. Andrews McMeel Publishing.

Tibshirani, R. 1988. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Society*, **83**, 394–405.

Turner, J.E. 1963. Backbench opinion in the House of Commons, 1955-59 by Finer S.E., Berrington, H.B., Bartholomew, D.J. *Administrative Science Quarterly*, **8**(1), 104–108.

Vidakovic, B. 1999. *Statistical Modeling by Wavelets*. New York: John Wiley & Sons.

von Bortkiewicz, L. 1898. *Das Gesetz der kleinen Zahlen*. Leipzig: Teubner.

Wand, M., & Jones, M. 1995. *Kernel Smoothing*. New York: Chapman & Hall.

Watson, G.S. 1964. Smooth regression analysis. *Sankhya, Series A*, **26**, 359–372.

Watt, N. 2003. MPs to grill cabinet on WMD. *The Guardian*, 4th June.

Whitaker, B. 2004. Arab world mourns "Father of nation". *The Guardian*, 12th November.

White, M. 2005. Cameron's new Conservatism. *The Guardian*, 7th December.

Winkelmann, R. 2003. *Economic Analysis of Count Data*. New York: Springer.

Wintour, P. 2006. Blair wins on education - but at a cost. *The Guardian*, 16th March.

Wintour, P., Ahmed, K., Vulliamy, E., Taynor, I., & Saraj, J. 2001. It's time for war, Bush and Blair tell Taliban. *The Observer*, 7th October.

Zeger, S. 1988. A regression model for time series of counts. *Biometrika*, **75**, 621–629.

Zhang, B., Fadili, M.J., & J-L., Starch. 2006. Multi-scale variance stabilizing transform for multi-dimensional Poisson count image denoising. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1329–1332.