Statistical Models For Rank Data

A thesis submitted to the University of Dublin, Trinity College

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Statistics, University of Dublin, Trinity College

September 2006

Isobel Claire Gormley

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Isobel Claire Gormley.

Isobel Claire Gormley

Dated: September 16, 2006

Abstract

Rank data arise when a set of judges rank some or all of a set of objects. Rank data emerges in many areas of society; the list of the world's most cited scientists or the final ordering of horses in a race provide examples of such data. Irish society generates a wealth of rank data in two specific contexts: applicants to Irish third level educational institutions rank degree courses in order of preference and under the Irish electoral system voters rank candidates in order of preference.

The relationships that may exist between the set of objects ranked and between the judges who rank them are explored in this thesis. The set of applicants to Irish third level institutions in the year 2000 are investigated to determine if groups of similar applicants exist and if so, what characteristics they share. Voters and candidates from the 1997 Irish presidential election and from the 2002 Irish general election are examined. The (dis)similarities that the Irish electorate deem to exist between candidates are revealed.

Complex rank data models are developed which take account of the ranked nature of the data. Mixture models, and extensions thereof, are used to model heterogeneous populations which generate rank data; a latent space model is also proposed which locates the ranked objects in an unobservable space. Model fitting is performed in both classical and Bayesian frameworks. Unique model fitting techniques are necessary due to the complex nature of the models.

Examining fitted model parameters provides insight to the underlying mechanisms which drive Irish social opinions. Applicants to Irish third level institutions are influenced by course discipline, an institutions' geographical location and by course prestige. Evidence of both candidate orientated and politically driven voters is also presented.

Acknowledgements

First and foremost I would like to express my gratitude to my supervisor Dr Brendan Murphy. I have worried more about adequately expressing my thanks to him in this thesis than about the chapters within it! Brendan has given me every opportunity to broaden my knowledge: he taught me as much statistics as I could absorb, provided numerous opportunities to attend and present at conferences all over the world and even the chance to live abroad. He has always been helpful and patient and always had time to answer a query, no matter how small. He has been fun to work with and has taught me as much about life as about statistics. Thanks must also go to Brendan and his wife Trish for their endless kindness in Seattle.

Secondly I would like to thank my parents and family. My parents have provided myself and my siblings with every opportunity and endless support in life and it is to them we owe all we have. Thank you to them for everything they have done and continue to do for all of us. Thanks to Brian, Mary, Maeve, Kevin and Stephen for putting up with my years of studentdom! I must also thank Mary for cleaning out her study and taking me in.

Thank you to all in the Department of Statistics in Trinity College Dublin. They were always kind and helpful and made my years there very enjoyable. Thanks to all the postgraduates and postdocs, both past and present, for the chats and interesting times. Thanks to all in the Department of Statistics in the University of Washington, Seattle, especially to Professor Adrian Raftery and members of his Working Group in Model-based Clustering. Many thanks to the postgraduates there also for introducing me to soccer and softball; you made my months in the USA great fun.

Thanks to the Irish Research Council for Science, Engineering and Technology for amply supporting me and for providing me with the chance to attend conferences throughout the world. Thanks also to Trinity College for the travel grant they awarded. Thanks to Joan Burton and the CAO who provided the data from which this thesis emerged.

Finally, my biggest thanks to my dear Michael Brides who has supported me through it all. Thanks for being there with me through the ups and downs of the past few years and over the miles that often separated us. Thank you for your patience and your support.

Isobel Claire Gormley
University of Dublin, Trinity College
September 2006

Contents

Abstract iii

Acknowledgements iv

List of Tables x

List of Figures xii

Publications xiv

Chapter 1 Introduction 1

1.1 The Central Applications Office System 2

1.2 Irish Elections 3

1.3 Overview of Chapters 4

1.4 Research Contributions 7

Chapter 2 Rank data 8

2.1 Central Applications Office Data 8

2.2 Irish Voting Data 10

2.2.1 The 1997 Presidential Election 10

2.2.2 The 2002 General Election 12

2.2.3 The Vote Counting Process 14

Chapter 3 Statistical Methodology 18

3.1 The Plackett-Luce Model for Rank Data 18

3.2 Benter's Model for Rank Data 21

3.3 Classical Inference 22

3.3.1 The EM Algorithm 23

3.4 Bayesian Inference 25

3.4.1 Prior Elicitation 26

3.4.2 Markov Chains 27

3.4.3 Markov Chain Monte Carlo Methods 29

3.5 Model Selection Techniques 34

Chapter 4 Mixtures of Plackett-Luce Models 37

4.1 Mixture Models 37

4.2 The Plackett-Luce Model 38

4.3 Model Fitting 39

4.3.1 The MM Algorithm 42

4.3.2 Estimation of Support Parameters 44

4.3.3 The EM/MM Algorithm 45

4.4 Provision of Parameter Standard Errors 46

4.4.1 Computation of the Empirical Information Matrix for a Mixture of Plackett-Luce Models 48

4.4.2 Multiparameter Standard Errors 49

4.5 Model Extension: Inclusion of a Noise Component 49

4.6 Model Comparison 50

4.7 Mixtures of Plackett-Luce Models for CAO Applicants 50

4.7.1 The Geographical Effect 53

4.7.2 The ‘Points Race’ 54

4.7.3 Examination of Component Parameters 54

4.7.4 Examination of Posterior Component Membership Probabilities 58

4.7.5 The Gender Effect 60

4.7.6 Clustering of Applicants 62

4.8 Conclusions 63

Chapter 5 Mixtures of Benter Models 66

5.1 The Benter Model 66

5.2 Homogeneous Benter Models 67

5.3 Mixtures of Benter Models 71

5.3.1 Mixture Constraints 71

5.3.2 Fitting Mixtures of Benter Models 72
5.3.3 Maximization with Respect to Support Parameters. 74
5.3.4 Maximization with Respect to Dampening Parameters. 76
5.3.5 The EM/MM Algorithm. 77
5.4 Estimation of Standard Errors 78
5.5 Model Comparison 79
5.6 Analysis of the Irish Electorate 80
5.6.1 The 1997 Presidential Election 80
5.6.2 The 2002 General Election 85
5.7 Conclusions 89

Chapter 6 A Grade of Membership Model for Rank Data 90

6.1 Model Specification 90
6.1.1 The Plackett-Luce Model 91
6.1.2 The Grade of Membership Model 91
6.1.3 Prior and Posterior Distributions 92
6.2 Parameter Estimation 94
6.2.1 The Gibbs Sampler 94
6.2.2 The Metropolis Within Gibbs Sampler 95
6.2.3 Surrogate Proposal Distributions 96
6.3 Model Features 99
6.3.1 Label Switching 99
6.3.2 Model Selection 100
6.4 Application of the GoM Model to the Irish Electorate 101
6.4.1 Support for the Presidential Candidates 102
6.4.2 Mixed Membership Parameters for the Electorate 105
6.5 Conclusions 108

Chapter 7 A Mixtures-of-Experts Model for Rank Data 109

7.1 Mixtures-of-Experts Models 109
7.2 Generalized Linear Models 111
7.3 Fitting a ME Model. 113
7.3.1 The E Step 114

7.3.2 The M Step 115
7.3.3 Estimation of Gating Network Parameters 116
7.3.4 The EM/MM Algorithm 117
7.4 Standard Errors for ME Parameters 118
7.4.1 Score Function for Benter Support Parameters 119
7.4.2 Score Function for Dampening Parameters 119
7.4.3 Score Function for Gating Network Parameters 119
7.5 Application of ME Models to Irish Presidential Opinion Poll Data . . 120
7.6 Conclusions 127

Chapter 8 A Latent Space Model for Rank Data 128
8.1 Model Specification 129
8.1.1 The Latent Space Model 129
8.1.2 The Plackett-Luce Model 130
8.2 Model Fitting 131
8.2.1 Estimation of Voter and Candidate Latent Locations 131
8.2.2 Invariant Configurations 133
8.2.3 Dimensionality 134
8.3 A Latent Space Model for Irish Voting Data 135
8.3.1 The 2002 General Election: Dublin North Constituency 135
8.3.2 The 2002 General Election: Dublin West Constituency 139
8.3.3 The 2002 General Election: Meath Constituency 146
8.3.4 The 1997 Irish Presidential Opinion Polls 151
8.4 Conclusions 155

Chapter 9 Conclusions and Further Work 158
9.1 Conclusions 158
9.1.1 Central Applications Office Data 158
9.1.2 Irish Voting Data 159
9.2 Further Work 160

Appendix A Data Sources 163

Appendix B Propositions 164

List of Tables

2.1 Irish presidential candidates in 1997. 11
2.2 Dublin North candidates in the 2002 general election. 13
2.3 Dublin West candidates in the 2002 general election. 14
2.4 Meath candidates in the 2002 general election. 15
2.5 Transfer of votes in the Dublin West constituency. 17

4.1 The twenty-two components detected when analyzing all CAO applicants. 52
4.2 Most probable courses ranked by a health sciences applicant. 56
4.3 Courses whose selection gives highest probability of belonging to the health sciences component. 59
4.4 Resulting components from analyzing applicants by gender. 61
4.5 Summary statistics associated with clustering uncertainty of applicants. 63

5.1 Presidential election dampening parameters. 69
5.2 Maximum likelihood estimates of support parameters for the 2002 Dublin West constituency. 69
5.3 Numbers of parameters in mixture models. 80
5.4 Support parameter estimates for the 1997 presidential opinion polls. . 82
5.5 Fifteen component mixture of Benter models fitted to Dublin West data. 85

7.1 Covariates recorded in the 2/10 IMS poll. 121
7.2 BIC values, number of expert networks and associated covariates for ME models. 122

7.3 Gating network parameter estimates and associated confidence intervals under the ME model. 125

8.1 Dublin North latent space model selection criteria. 136

8.2 Proportion of the variance explained by each principal component when PCA is applied to the Dublin North candidate locations. 136

8.3 Dublin West latent space model selection criteria. 140

8.4 Proportion of the variance explained by each principal component when PCA is applied to the Dublin West candidate locations. 140

8.5 Details of Dublin West candidates. 141

8.6 Model selection criteria for the optimal latent space model for the Meath constituency. 146

8.7 Principal component variances for the Meath candidates configurations.147

8.8 Latent space model selection criteria for presidential exit poll data. . 151

8.9 Variance of principal components fitted to presidential exit poll data. 152

List of Figures

1.1 Sample headlines from national Irish newspapers. 2

4.1 An example of a linear minorizing surrogate function. 43

4.2 Uncertainty in the clustering of female and male applicants. 63

5.1 Support parameters for the 1997 presidential opinion polls. 68

5.2 Mosaic plot of Plackett-Luce mixture fitted to the 1997 presidential exit poll. 84

5.3 Mosaic plot of Benter mixture fitted to the Dublin West data. 86

5.4 Mosaic plot of groups 1 and 3 support parameters within the Benter mixture fitted to the Dublin West data. 86

6.1 Model selection criterion for GoM models. 102

6.2 Box and whisker plots of the posterior mean support parameter estimates for the 1997 exit poll data. 103

6.3 Trace plots of samples of support parameters from the 1997 exit poll. 104

6.4 Mixed membership parameters for three randomly sampled voters from the 1997 exit poll. 106

7.1 Tree like structure of a mixtures-of-experts model. 110

7.2 Support parameters under the mixtures-of-experts model. 124

8.1 The one dimensional configuration of the Dublin North candidate means. 137

8.2 The two dimensional configuration of the Dublin North candidate means. 138

8.3 The one dimensional configuration of the Dublin West candidate means. 141

8.4 The two dimensional configuration of the Dublin West candidate means. 142

8.5 The three dimensional configuration of the Dublin West candidate means. 143

8.6 A matrix of scatter plots of the three dimensional configuration of the Dublin West candidate means. 144

8.7 One dimensional configuration of the Meath candidate locations. . . . 149

8.8 Two dimensional configuration of the Meath candidate locations. . . . 150

8.9 One dimensional configuration of the 1997 presidential candidate locations. 152

8.10 Chronological trace of the spatial movement of each of the 1997 presidential candidates in one dimension. 153

8.11 Two dimensional configuration of the 1997 presidential candidate locations. 156

8.12 Chronological trace of the spatial movement of each of the 1997 presidential candidates in two dimensions. 157

Publications

Some of the material presented in this thesis has been taken from the author's following publications which are co-authored by Dr Thomas Brendan Murphy.

Gormley, I.C. and Murphy, T.B. (2005), Exploring Heterogeneity in Irish Voting Data: A Mixture Modelling Approach., Technical Report no. 05/09, Department of Statistics, Trinity College Dublin.

Gormley, I.C. and Murphy, T.B. (2006) 'Analysis of Irish Third-Level College Applications Data.', *Journal of the Royal Statistical Society, Series A.* **169**, Part 2, 361–379.

Murphy, T.B. and Gormley, I.C. (2006), Discussion of 'Bayesian paleoclimate reconstruction.', Haslett et al., *Journal of the Royal Statistical Society, Series A.* **169**, Part 3, 434–435.

Gormley, I.C. and Murphy, T.B. (2006), Discussion of 'Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity.', Raftery et al., *Bayesian Statistics 8.*

Gormley, I.C. and Murphy, T.B. (2006), 'A Latent Space Model for Rank Data.', *Statistical Network Analysis: Models, Issues, and New Directions. Lecture Notes in Computer Science.*

Chapter 1

Introduction

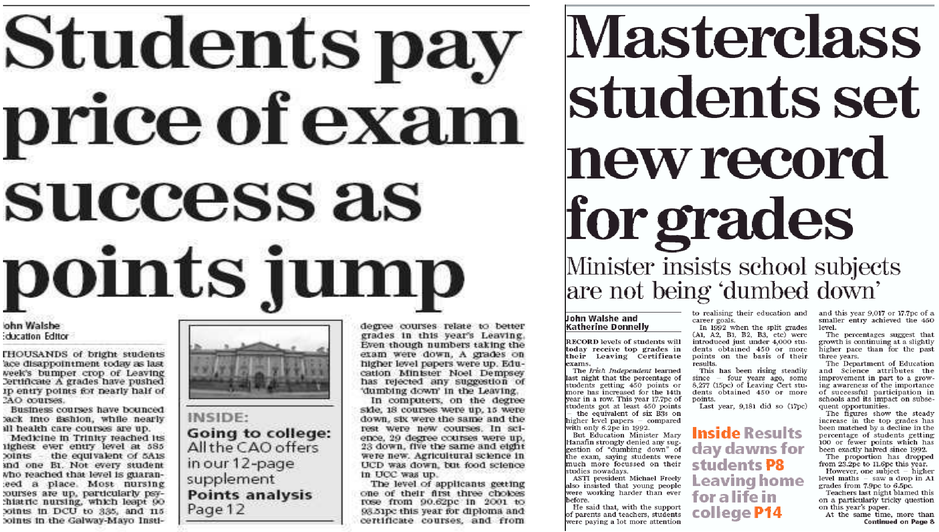
Rank data arises when a set of judges rank some or all of a set of objects. Rank data emerges naturally in many areas of society; the list of the world's most cited scientists, the ranking of internet search results by relevance or the final ordering of horses in a race all provide examples of such data.

Two aspects of Irish society produce a wealth of rank data. Applicants to institutions of third level education in Ireland apply directly to the Central Applications Office (CAO). On an application form, an applicant may rank up to ten degree programs in order of preference. Thus each year thousands of applicants with different characteristics and vocations generate a rich source of rank data. Another source of rank data in Ireland are presidential and general elections. Irish presidential elections employ an electoral system known as the Single Transferable Vote system (STV); a similar system known as Proportional Representation by means of a Single Transferable Vote (PR-STV) is used in general elections. Under these electoral systems a voter's ballot form consists of a ranking of some or all of the electoral candidates in order of preference. Opinion polls, typically conducted prior to elections, generate similar preference data. The aim of this thesis is to explore the relationships that may or may not exist between the set of objects ranked and between the judges who rank them within the contexts of Irish third level college applications and Irish elections.

1.1 The Central Applications Office System

The Irish college applications system involves prospective college students ranking up to ten degree courses in order of preference prior to sitting their final second level examinations (Leaving Certificate). Applications are processed by the Central Applications Office (CAO) who deal with applications for all third level degree programs in Ireland.

The method of gaining entry to third level education, as managed by the CAO, is a much debated subject among the Irish media, students, parents and education circles. Many aspects of the CAO system appear annually as headlines in the Irish media (see Figure 1.1) – national front pages carry stories of fluctuating entry points requirements and volatile applicant numbers, particularly for the weeks surrounding the announcement of who is admitted to each course.



Students pay price of exam success as points jump

ohn Walshe
ducation Editor

THOUSANDS of bright students ace disappointment today as last week's bumper crop of Leaving Certificate A grades have pushed up entry points for nearly half of CAO courses.

Business courses have bounced back into fashion, while nearly all health care courses are up.

Medicine in Trinity reached its highest ever entry level at 585 points – the equivalent of 5A1s and one B1. Not every student who reached that level is guaranteed a place. Most nursing courses are up, particularly psychiatric nursing, which leapt 90 points in DCU to 335, and 115 points in the Galway-Mayo Insti-

INSIDE:
Going to college: All the CAO offers in our 12-page supplement
Points analysis Page 12

degree courses relate to better grades in this year's Leaving. Even though numbers taking the exam were down, A grades on higher level papers were up. Education Minister Noel Dempsey has rejected any suggestion of 'dumbing down' in the Leaving.

In computers, on the degree side, 18 courses were up, 15 were down, six were the same and the rest were new courses. In science, 29 degree courses were up, 23 down, five the same and eight were new. Agricultural science in UCD was down, but food science in UCC was up.

The level of applicants getting one of their first three choices rose from 90.62pc in 2001 to 93.51pc this year for diploma and certificate courses, and from

(a)

Masterclass students set new record for grades

Minister insists school subjects are not being 'dumbed down'

John Walshe and
Katherine Donnelly

RECORD levels of students will today receive top grades in their Leaving Certificate exams.

The *Irish Independent* learned last night that the percentage of students getting 450 points or more has increased for the 14th year in a row. This year 17.7pc of students got at least 450 points – the equivalent of six B3s on higher level papers – compared with only 6.2pc in 1992.

But Education Minister Mary Hanafin strongly denied any suggestion of "dumbing down" of the exam, saying students were much more focussed on their studies nowadays.

ASTI president Michael Freely also insisted that young people were working harder than ever before.

He said that, with the support of parents and teachers, students were paying a lot more attention to realising their education and career goals.

In 1992 when the split grades (A1, A2, B1, B2, B3, etc) were introduced just under 4,000 students obtained 450 or more points on the basis of their results.

This has been rising steadily since – four years ago, some 8,277 (15pc) of Leaving Cert students obtained 450 or more points.

Last year, 9,181 did so (17pc) and this year 9,017 or 17.7pc of a smaller entry achieved the 450 level.

The percentages suggest that growth is continuing at a slightly higher pace than for the past three years.

The Department of Education and Science attributes the improvement in part to a growing awareness of the importance of successful participation in schools and its impact on subsequent opportunities.

The figures show the steady increase in the top grades has been matched by a decline in the percentage of students getting 100 or fewer points which has been exactly halved since 1992.

The proportion has dropped from 23.2pc to 11.6pc this year.

However, one subject – higher level maths – saw a drop in A1 grades from 7.9pc to 6.5pc.

Teachers last night blamed this on a particularly tricky question on this year's paper.

At the same time, more than

Continued on Page 8

Inside Results day dawns for students P8
Leaving home for a life in college P14

(b)

Fig. 1.1: Sample headlines from the national 'Irish Independent' newspaper from the days surrounding the announcement of the Leaving Certificate results. Figure 1.1(a) is the front page headline from Monday August 23rd 2004; Figure 1.1(b) is the front page headline from Wednesday August 16th 2006.

Detractors suggest that applicants are influenced by the annual media hype and rank courses according to entry requirements, ensuring they study a current 'high profile' course and therefore they may ignore their vocational callings. They claim artificial demand is created for courses deemed to be of high social standing. Supporters insist it is a fair system where each applicant is dealt with in a consistent and transparent manner. The supporters claim that the so called ‘points race’ for entry to courses is media generated and has no significant affect on applications.

If students are actually selecting courses according to their prestige rather than by vocational callings, then there should exist groups of applicants where the discipline of their ranked courses are quite different, but the common feature of their selected courses is that they have high points requirements. Therefore, if the points race drives applicants’ choices, then groups of applicants ranking high points requirement courses together (such as Law, Medicine, Pharmacy, Dentistry and Actuarial Science) should be present, but where the courses are from different disciplines. On the other hand, if the system does work in its intended manner, then applicants should belong to groups where the discipline of their ranked courses is consistent.

This thesis focuses on analyzing the set of degree course applications made through the CAO in the year 2000; there is a separate applications system for diploma and nursing courses. Details of the CAO data are provided in Chapter 2. The resulting groupings of applicants reveal that applicants generally appear to be driven by their vocational interests as discipline emerges as the defining characteristic of applicant groups. The geographical location of the institution to which an applicant applies also transpires to have a significant influence on course choice. Crucially however, some weight is added to the CAO system detractors’ arguments. A deeper analysis of the revealed groups highlights a subtle influence of the required points on the applicants’ choices. A separate analysis of the male and female data suggests applicants of different gender have different course choice behaviours.

1.2 Irish Elections

In elections, the electorate exhibit different voting behaviours by choosing to vote for different candidates. The difference in voting behaviour may be due to allegiance to a political party, choosing familiar candidates, choosing geographically local candidates or one of many other reasons. The different voting behaviours lead to a collection of votes from a heterogenous population.

This thesis focuses on studying Irish elections because the votes recorded under the Single Transferable Vote electoral system contain information on the preferences that the voters have for the candidates. The aim is to provide an exploratory analysis of the Irish electorate and their opinions. This greatly adds to the understanding of how the Irish electoral system works in practice.

Two elections are analyzed — the 1997 Irish presidential election and the 2002 Irish general election. These elections are quite different in character; the general election elects the government and party politics are believed to play an important role, whereas party politics are believed to only play a minor role in the presidential election. Full details of both elections and of the electoral system are provided in Chapter 2.

It is observed that there is strong political party support in the general election, because voters tend to give their high preferences to candidates from the same political party or to parties of a particular persuasion. There is also evidence however of candidate orientated voters within the Irish electorate in both the presidential and general elections.

1.3 Overview of Chapters

A brief outline of the research conducted follows.

Chapter 2: Rank Data

The models and methods described in this thesis are applied to rank data which naturally arise within the context of Irish society. The circumstances which give rise to these rank data sets and their specific features are detailed in this chapter. Full details of the intricate counting process employed in Irish elections is also given.

Chapter 3: Statistical Methodology

Statistical models for rank data are necessary for the appropriate modelling of the Irish CAO and election data sets. Details of the Plackett-Luce model and of Benter's model for rank data are provided. Model fitting is achieved in both the classical and Bayesian frameworks throughout this thesis — the theory and methods employed in both paradigms are discussed in this chapter. Finally, model selection is an fundamental element of statistical modelling; many different criteria and techniques are available. A discussion of those examined and implemented in this work is given.

Chapter 4: Mixtures of Plackett-Luce Models

In Chapter 4 mixture models are used to investigate the presence of groupings in the set of Irish third level college applications. A finite mixture model assumes that the population consists of a finite collection of components. It is assumed that the probability of belonging to component k is π_k. In addition, an observation within component k has a probability density $f(\cdot|\underline{p}_k)$, where $\underline{p}_k$ are unknown parameters. By estimating the values of π_k and $f(\cdot|\underline{p}_k)$ this model-based approach allows a complete clustering of the applicants to be made. The mixture models that are employed use the Plackett-Luce model for ranked data as the probability densities within each component.

Models are fitted within the classical framework with extensive use made of the EM algorithm. Mathematical intractability causes problems when implementing the EM algorithm for rank data and thus a compound EM/MM algorithm is used.

Chapter 5: Mixtures of Benter Models

A similar mixture modelling approach is proposed for modelling the heterogeneous Irish electorate. Voting patterns are modelled using Benter's model for rank data; mixtures of Benter models are fitted by maximum likelihood using the EM algorithm. Issues with fitting mixture models using the EM algorithm are discussed and variants of the EM algorithm are also considered. Mixtures of Benter models are fitted to data from the 1997 presidential election and to data from the 2002 general election.

Chapter 6: A Grade of Membership Model for Rank Data

A grade of membership model is similar to a mixture model in that it models a heterogeneous population as a collection of 'extreme profiles'. However, the grade of membership model allows each member of the population have a probability of belonging to each extreme profile rather than only forming a hard partition of the population as mixture models do. The grade of membership model is incorporated with the Plackett-Luce model for rank data which is used to model the voters' preference data. This is fitted within the Bayesian paradigm to data from the 1997 Irish presidential election.

Chapter 7: A Mixture-of-Experts Model for Rank Data

Mixture-of-experts models build further on the structure implemented by mixture models by taking account of both the observations and associated covariates when modelling a heterogeneous population. The aim in this chapter is to perform an exploratory analysis of the Irish electorate to determine which social factors influence voting patterns and what the induced voting patterns are. Both the votes cast and the covariates associated with the voters are modelled. The preferences expressed are modelled via Benter's model for rank data. Data from the 1997 Irish presidential opinion polls are analyzed. Model fitting is conducted in the classical framework again making use of a compound EM/MM algorithm.

Chapter 8: A Latent Space Model for Rank Data

Early chapters focus on modelling and exploring the heterogeneous nature of a set of judges who generate rank data. The latent space model introduced in this chapter provides another tool for exploring such a population. The focus however is no longer on examining the heterogeneous nature of the judges but on estimating the relative locations of the judges and the objects they rank in a latent space. This model is fitted to a range of data sets from both the 1997 Irish presidential election and from the 2002 Irish general election. The relative spatial locations of the candidates in the latent space are suggestive of the type of relationships that may exist between the candidates as viewed by the electorate.

The Plackett-Luce model for rank data is again employed to model the ranked nature of the electorate's votes. The latent space model is fitted within the Bayesian paradigm; typical latent space model issues such as identifiability and dimensionality arise.

1.4 Research Contributions

The following are the main contributions made by the research contained in this thesis:

1. The development of statistical models for heterogeneous populations, the members of which have expressed preferences on a number of objects. Models which have the scope to incorporate covariates associated with members of the population have also been presented.

2. A latent space model for objects which have been ranked has been completed. The ranked objects have a location in an unobservable space where their relative positions are suggestive of the relationships between the objects.

3. Maximization issues associated with complex rank data models have been overcome via the MM algorithm.

4. The provision of suitable proposal distributions for sampling within a Bayesian framework via ideas from the MM algorithm has been developed.

Chapter 2

Rank data

The models and methods described in future chapters are applied to rank data which naturally arise within the context of Irish society. The circumstances which give rise to these rank data sets and their specific features are detailed here.

2.1 Central Applications Office Data

A college application which is made through the CAO allows an applicant rank up to ten degree courses in order of their preference. Course places are subsequently offered using both these ordered choices and the applicants' grades. The Central Applications Office (CAO) data set was collected in the year 2000 and consists of the course choices of 53757 applicants to degree courses offered in Irish third level institutions. A total of 533 degree courses were selected by the applicants. The gender of each applicant is known — there were 29338 female and 24419 male applicants in the year 2000. Characteristic features of these data include the large number of applicants giving preferences for a large number of courses and the constraint that applicants are restricted in the number of courses they may rank.

Typically, seven or eight subjects are taken for the Leaving Certificate examination. Once graded the best six examination results are used to produce a 'points' score; each grade A1, A2, B1, ..., NG (No Grade) has an associated number of points. Subsequent to examination grading, the CAO fixes a universal points requirement for each degree program. Applicants are subsequently offered a place in their highest preference course for which they have achieved the points requirement; in the case of applicants being tied for the last available positions in a course, random allocation is used to choose which applicant is offered a place.

It is worth emphasizing that applicants do not know the required course points requirement prior to completing their application or to taking their examinations. The points requirement is influenced by the examination results of applicants who applied for the course and by the number of available positions in the course. Some courses have minimum entry standards, for example, a sufficient standard of mathematics may be required for an engineering degree. However, the actual subjects taken at Leaving Certificate level do not have an effect on the applicants points score nor does previous examination performance; a few courses have interviews but these are not common. The subjects Irish, English and Mathematics taken at Leaving Certificate level are entry requirements for Irish applicants for many courses but the remaining subjects are the student's choice. In addition, the Leaving Certificate can be taken several times without having any effect on an application.

International applicants are dealt with in the same manner. For example, the UK final secondary level A-Level results are converted into points — these are totalled and subsequently such applicants are allocated a course by the same method as Irish Leaving Certificate students. The college applications system used in Australia is also similar; applicants rank up to nine courses which are processed by a Universities Admission Centre (UAC) (see `http://www.uac.edu.ac`). Both the Irish CAO and Australian UAC systems can be likened to the 'Tote' in horse-racing betting (`http://www.tote.ie`). In the Tote, no punter knows the odds on any horse prior to all bets being placed. Similarly under the CAO system no applicant knows the points requirements for any course prior to the publication of all examination results. Extensive details of the college applications system are available on the CAO web page (`http://www.cao.ie`).

In 1997, the Minister for Education and Science set up the "Commission on the Points System" to review the current college applications system. This led to the publication of a report (Hyland, 1999) which reviews the system and makes a series of recommendations concerning its future. A series of four research reports were also published in conjunction with the commission's final report. Of particular interest is the report of Tuohy (1998) who studies the college application data using exploratory techniques; this work is the closest to the analysis presented in this thesis. Of some interest is the report of Lynch et al. (1999) who investigates the predictive performance of the points awarded to applicants in determining overall performance in higher education. These reports received an enormous amount of coverage in the Irish media and were discussed at length by the public. The general conclusion of the exercise was although the current system is not perfect, it works very well in practice. Clancy (1995) studies the admissions (rather than applications) data for students in Irish third level institutions, but his work is closely related to this analysis.

2.2 Irish Voting Data

Irish general (governmental) elections employ an electoral system known as Proportional Representation by means of a Single Transferable Vote (PR-STV). Since proportional representation is not possible in single seat elections, Irish presidential elections employ the Single Transferable Vote (STV) system. Under both systems voters rank some (or all) of the electoral candidates in order of preference. The votes are totalled through a series of counts, where candidates are eliminated, their votes are distributed, and surplus votes are transferred between candidates. An in depth description of the electoral system, including the method of counting votes is given in Sinnott (1999) and good introductions to the Irish political system are given in Coakley and Gallagher (1999) and Sinnott (1995). Further, an illustrative example of the manner in which votes are counted and transferred follows in Chapter 2.2.3.

2.2.1 The 1997 Presidential Election

The eighth (and current) President of Ireland, Mary McAleese, was originally elected in 1997. The number of candidates in the 1997 presidential election was larger than usual. There were five candidates that year: Mary Banotti, Mary McAleese, Derek Nally, Adi Roche, and Rosemary Scallon. Some candidates were endorsed by political parties and others were independent candidates (see Table 2.1). Mary Banotti, Derek Nally and Adi Roche were considered to be liberal candidates where Mary McAleese and Rosemary Scallon were deemed the more conservative candidates;

Derek Nally entered the election race at a later stage than the other four candidates.

Table 2.1: The five candidates who ran for Irish presidency in 1997 and their endorsing political parties. Mary McAleese was subsequently elected.

Candidate	**Endorsing Party**
Mary Banotti	Fine Gael (FG)
Mary McAleese	Fianna Fáil, Progressive Democrats and Sinn Féin (FF, PD and SF)
Derek Nally	Independent (Ind)
Adi Roche	Labour (Lab)
Rosemary Scallon	Independent (Ind)

Seven opinion polls and an exit poll, taken on polling day, were completed during the election campaign.

Four of the opinion polls were conducted by Irish Marketing Surveys (IMS) during the two months prior to the election. Approximately 1100 respondents, drawn from 100 sampling areas, were interviewed for each poll. Interviews took place at randomly located homes with individuals selected according to a socioeconomic quota. A range of sociological questions were asked of each respondent as was the respondent's voting preference, if any, for each of the candidates. These preferences were in effect utilized as each respondent's vote. Everyone included in the poll data expressed at least one preference — in fact each poll has slightly more than the required 1100 respondents.

The other three opinion polls were conducted by the Market Research Bureau of Ireland (MRBI), again during the two month electoral campaign. A similar sampling methodology as used in the IMS polls was employed — 100 Primary Sampling Units (PSU's) were selected from census data, and from each PSU 10 interviews were conducted using a random route procedure. The sample was quota controlled by age, gender, and socioeconomic class. Each of the three MRBI polls contained 'missing' data — an average of 150 respondents in each poll either replied don't know, won't vote or refused to give their preferences. Examining such voters and their covariates is an area of further research; only those respondents who expressed at least one preference are modelled here.

On the day of the presidential election, October 30th 1997, Lansdowne Market Research conducted an exit poll where 2498 voters were interviewed at 150 polling stations in all 41 Irish constituencies.

As all of these polls were conducted using similar methodology the comparison of respondents of different polls is deemed to be justified.

A detailed description of the entire presidential election campaign, including the nomination and selection of candidates, is given by Marsh (1999). The sources of the poll data are given in Appendix A.

2.2.2 The 2002 General Election

Ireland had its most recent general election on May 17th, 2002. In 2002 the Irish electorate was composed of forty two constituencies; the 1997 Dublin West constituency had been subdivided into the Dublin West and Dublin Mid-West constituencies. One hundred and sixty six politicians were elected to be members of Dáil Éireann (the Irish parliament). This election saw the introduction of electronic voting, for the first time, in three constituencies (Dublin North, Dublin West, and Meath). The remaining thirty nine constituencies had paper ballots. The electronic votes cast are analyzed in this work.

Dublin North.

In the Dublin North constituency twelve candidates campaigned for four parliamentary seats. The total electorate was 72908 and only 43942 valid votes were cast. The two major Irish political parties of Fianna Fáil and Fine Gael had multiple representatives — Fianna Fáil had three and Fine Gael two. The candidates who ran in the Dublin North constituency and their political affiliations are detailed in Table 2.2.

Dublin West.

In the Dublin West constituency three seats were to be filled with nine candidates running for election. The nine candidates represented eight political parties, with Fianna Fáil having two candidates. The electorate was 53780 and there was a total of 29988 valid votes cast. Table 2.3 details the candidates and their associated political parties.

Meath.

Five seats in Dáil Éireann were to be filled from the constituency of Meath and fourteen candidates ran for these seats within the constituency. The fourteen candidates represented seven political parties, with the major Irish parties of Fianna Fáil and Fine Gael each having three candidates (see Table 2.4). The electorate was 108717 and there was a total of 64081 valid votes cast.

The voting data from the Dublin North, Dublin West and Meath constituencies are publicly available and the sources are given in Appendix A. These data were previously analyzed using exploratory data analysis techniques by Laver (2004).

Table 2.2: The twelve candidates who ran for election in the Dublin North constituency. An asterisk * before the name of a candidate indicates that the candidate was subsequently elected in the 2002 election.

Candidate	**Abbreviation**	**Party**
BOLAND, Cathal	Bol	Fine Gael (FG)
DALY, Clare	Dal	Socialist Party (SP)
DAVIS, Mick	Dav	Sinn Féin (SF)
*GLENNON, Jim	Gle	Fianna Fáil (FF)
GOULDING, Ciaran	Gou	Independent (Ind)
KENNEDY, Michael	Ken	Fianna Fáil (FF)
OWEN, Nora	Owe	Fine Gael (FG)
QUINN, Eamon	Qui	Independent (Ind)
*RYAN, Seán	Rya	Labour (Lab)
*SARGENT, Trevor	Sar	Green Party (GP)
WALSHE, David	Wal	Christian Solidarity Party (CSP)
*WRIGHT, G.V.	Wri	Fianna Fáil (FF)

Table 2.3: The nine candidates who ran for election in the Dublin West constituency. An asterisk * before the name of a candidate indicates that the candidate was subsequently elected in the 2002 election.

Candidate	Abbreviation	Party
BONNIE, Robert	Bon	Green Party (GP)
*BURTON, Joan	Bur	Labour (Lab)
DOHERTY-RYAN, Deirdre	Do-Ry	Fianna Fáil (FF)
*HIGGINS, Joe	Hig	Socialist Party (SP)
*LENIHAN, Brian	Len	Fianna Fáil (FF)
McDONALD, Mary Lou	McD	Sinn Féin (SF)
MORRISSEY, Tom	Mor	Progressive Democrats (PD)
SMYTH, John Thomas	Smy	Christian Solidarity Party (CSP)
TERRY, Sheila	Ter	Fine Gael (FG)

2.2.3 The Vote Counting Process

A brief overview of the vote counting process is given here. For illustrative purposes, the transfer of votes in the Dublin West constituency, where there were three seats available for election, is shown in Table 2.5. Under the PR-STV electoral system a constituency specific 'quota' of votes is calculated which depends on the number of seats available and the number of valid votes cast. Specifically the quota is

$$\text{quota} = \frac{\text{total valid votes in the constituency}}{\text{number of seats to be filled } + 1} + 1.$$

Thus for the Dublin West constituency the quota was calculated to be 7498. Once any candidate at any counting stage obtained or exceeded 7498 votes this candidate was elected.

As detailed in Table 2.5, in the first stage of the counting process the number of first preference votes obtained by each candidate is totalled. As candidate Lenihan got 8086 first preference votes, which is more than the quota, he was the first candidate to be elected. Candidates Bonnie and Smyth got the lowest number of first preferences and, as neither would ever be able to exceed the quota of votes required, were eliminated from the race. Thus at the second stage of counting

Table 2.4: The fourteen candidates who ran for election in the Meath constituency. An asterisk * before the name of a candidate indicates that the candidate was subsequently elected in the 2002 election.

Candidate	**Abbreviation**	**Party**
*BRADY, Johnny	By	Fianna Fáil (FF)
*BRUTON, John	Bt	Fine Gael (FG)
COLWELL, Jane	Cl	Independent (Ind)
*DEMPSEY, Noel	Dp	Fianna Fáil (FF)
*ENGLISH, Damien	Eg	Fine Gael (FG)
FARRELLY, John	Fr	Fine Gael (FG)
FITZGERALD, Brian	Ft	Independent (Ind)
KELLY, Tom	Kl	Independent (Ind)
OBRIEN, Pat	Obr	Independent (Ind)
OBYRNE, Fergal	Oby	Green Party (GP)
REDMOND, Michael	Rd	Christian Solidarity Party (CSP)
REILLY, Joe	Rl	Sinn Féin (SF)
*WALLACE, Mary	Wl	Fianna Fáil (FF)
WARD, Peter	Wd	Labour (Lab)

their 748 and 134 votes respectively were transferred to the candidates given the second place preference on those ballot forms. Seventy five of these votes were non-transferrable i.e. no second place preferences were expressed. Lenihan's 588 votes in excess of the quota were transferred at the third stage of counting to those candidates given second place preferences on those ballot forms. The 588 of Lenihan's votes that were transferred were randomly selected from his 8086 first preference votes. At the fourth stage of the counting process, after the previous transfers, candidate McDonald was not be able to reach the quota and was thus eliminated from the race. Her 2524 votes were then transferred to the next most preferred remaining candidates detailed on each of the ballots. 487 of these were non-transferrable votes. Subsequent to the transfer of McDonald's votes, Higgins' 7853 votes exceeded the quota and thus he was elected. At the fifth stage of the counting process Morrissey was eliminated and his 2662 votes were transferred to those remaining candidates ranked next on the ballot forms — 359 of Morrissey's votes were non-transferrable. At the sixth and final stage of counting Doherty-Ryan had the least number of votes and as her elimination left only one candidate, Burton was elected. Thus Lenihan was elected outright on first preference votes, but Higgins and Burton were subsequently elected during the counting process.

While the PR-STV system has many proponents, it also has many opponents. Sinnott (1995) describes some of the potential problems with the PR-STV system in an Irish context. Other potential flaws are explained in Katz (1984) and Brams and Fishburn (1984).

It has been argued that the PR-STV voting system puts too little emphasis on the political parties and too much emphasis on the candidates (Katz, 1984; Blais, 1991) and thus can lead to fracticious governments; this potential problem is examined in Sinnott (1995) where it is concluded that this problem does not manifest itself to a great degree in Irish elections.

Table 2.5: The transfer of votes in the Dublin West constituency. The numbers marked in boldface indicate that the candidate was elected. Three seats were available. The - symbol indicates that a candidate has been eliminated from the election. The quota required for guaranteed election for this constituency is 7498.

Candidate	**Party**	**Count**					
		1	2	3	4	5	6
Bonnie, R.	GP	748	—	—	—	—	—
			-748				
Burton, J.	Lab	3810	4020	4079	4375	5125	**6300**
			+210	+59	+296	+750	+1175
Doherty-Ryan, D.	FF	2300	2386	2698	3056	3728	—
			+86	+312	+358	+672	-3728
Higgins, J.	SP	6442	6660	6731	**7853**	**7853**	**7853**
			+218	+71	+1122		
Lenihan, B.	FF	**8086**	**8086**	**7498**	**7498**	**7498**	**7498**
				-588			
McDonald, M.	SF	2404	2498	2524	—	—	—
			+94	+26	-2524		
Morrissey, T.	PD	2370	2480	2554	2662	—	—
			+110	+74	+108	-2662	
Smyth, J.	CSP	134	—	—	—	—	—
			-134				
Terry, S.	FG	3694	3783	3829	3982	4863	5669
			+89	+46	+153	+881	+806
Non-transferable			75	75	562	921	2668
			+75	+0	+487	+359	+1747

Chapter 3

Statistical Methodology

In future chapters the data outlined in Chapter 2 is modelled and analyzed to gain a better understanding of the mechanisms which give rise to it. Statistical models are constructed and inferences are drawn on the parameters of these models. Several different models are fitted using a range of methods from both the classical and Bayesian paradigms.

In this chapter the rank data models employed throughout are detailed as are the statistical frameworks in which they are fitted. Model choice is an intrinsic part of the analysis and the various model selection techniques employed are also detailed here.

3.1 The Plackett-Luce Model for Rank Data

When modelling rank data an appropriate rank data model is required. Several models have been proposed in the literature to model rank data. The Bradley-Terry model (Bradley and Terry, 1952) examines competition between a set of individuals as a set of pairwise comparisons from which an 'ability parameter' can be inferred and thus a ranking of the competitors can be formed. Future chapters have a slightly different context in that all competitors are simultaneously compared with each other, rather than in a pairwise manner. Fienberg and Larntz (1976) examine a log linear representation of the Bradley-Terry model whose advantage is that it can be easily generalized to deal with multivariate extensions of the Bradley-Terry model. Other possible models for rank data are described in Critchlow (1985), Diaconis

(1988) and Marden (1995). Fligner and Verducci (1993) provide details of a variety of applications of models for ranking data. More recently, Bradlow and Fader (2001) model the simultaneous movement of multiple items up and down a ranking over time within a Bayesian framework with an exploding multinomial-logit likelihood. Johnson et al. (2002) take a Bayesian latent variable approach to modelling rank data from multiple evaluators who may use different ranking criteria. They include parameters in their hierarchical model to accommodate ties within the rankings. Graves et al. (2003) model car racing results by using a combination of the Bradley-Terry model with the Luce model and Stern's model to form their 'attrition model' which estimates driver ability. A step-wise approach is taken where the probability of a driver finishing in last place is examined and from this the final permutation of drivers is built. Hunter (2004) discusses unique fitting techniques for the Bradley-Terry model.

Multi-stage ranking models (Marden, 1995, Section 5.6) have a nice interpretation in terms of sequentially choosing items in order of preference. One parsimonious multi-stage ranking model which is easily interpretable is the Plackett-Luce model (Plackett, 1975).

Plackett (1975) motivated the Plackett-Luce model in terms of modelling horse races where a vector of probabilities for each horse winning is used to construct a probability model for the finishing order. Similar characteristics can be identified between horse races and the process of ranking third level courses or electoral candidates; for example, once an object has been chosen it cannot be selected again, and following a choice being made the probability of any remaining object being selected at the next stage is altered.

The Plackett-Luce model is parameterized by a *support parameter*

$$\underline{p} = (p_1, p_2, \dots, p_N)$$

where N denotes the total number of objects from which the judges choose and $\sum_{j=1}^{N} p_j = 1$. The probability of object 1 being ranked in first position is p_1. The probability of object 2 being ranked second, given that object 1 is ranked first, is $p_2 / \sum_{j \neq 1} p_j$. That is, it is equal to the probability that object 2 is ranked first when all objects except object 1 are available for selection. The probability of object 3 being ranked third, given that objects 1 and 2 are selected first and second, is

$p_3/\sum_{j\notin\{1,2\}} p_j$. The process continues to give the other placing probabilities. That is, each ranking is modelled as the product of the probabilities of each chosen object being ranked first where, at each preference level, the probabilities are appropriately normalized.

Let n_i be the number of objects ranked by judge i and let $c(i,t)$ denote the course/candidate ranked at the tth level by judge i. The Plackett-Luce model then suggests the probability of judge i's ranking $\underline{x}_i = (c(i,1),\ldots,c(i,n_i))$ is

$$\begin{aligned}\mathbf{P}\{\underline{x}_i|\underline{p}\} &= \prod_{t=1}^{n_i}\mathbf{P}\{\text{Object } c(i,t) \text{ being ranked in position } t|\text{Available objects}\}\\ &= \frac{p_{c(i,1)}}{\sum_{s=1}^{N} p_{c(i,s)}}\cdot\frac{p_{c(i,2)}}{\sum_{s=2}^{N} p_{c(i,s)}}\cdots\frac{p_{c(i,n_i)}}{\sum_{s=n_i}^{N} p_{c(i,s)}} \qquad (3.1)\\ &= \prod_{t=1}^{n_i}\frac{p_{c_{(i,t)}}}{\sum_{s=t}^{N} p_{c(i,s)}}\end{aligned}$$

where, for $s > t$, the sequence of objects $c(i,s)$ is any arbitrary ordering of the unselected objects.

The Plackett-Luce model assumes a weak dependence between the objects ranked at different levels. In particular, $\mathbf{P}\{c(i,t) = j\}$ depends on $\{c(i,1),\ldots,c(i,t-1)\}$ but is independent of their order. This appears to be a reasonable assumption as the ranking process involves deciding if an object should be placed higher or lower than other alternatives and not on the specific level at which the other alternatives are ranked. Thus at each choice level in the Plackett-Luce model the support parameter probabilities are adjusted such that they account for the objects already ranked.

Rosén (1972) proposed an approximation for the Plackett-Luce model when studying sampling with unequal probabilities. Rosén proved that for sufficiently large N the difference between the probability of ranking object j in first position and the probability of ranking object j in any lower position tends to zero i.e. for large N the choices are approximately independent for $t = 1,\ldots,n_i$. Thus the probability of voter i's ranking could be modelled as

$$\mathbf{P}\{\underline{x}_i|\underline{p}\} \approx p_{c(i,1)}p_{c(i,2)}\cdots p_{c(i,n_i)}.$$

Implementation of this approximation is discussed in Chapter 6.

The Plackett-Luce model is said to exhibit independence from irrelevant alternatives (see Train, 2003) as the ratio of the probabilities of choosing one alternative over another is independent of all other available alternatives and independent of the choice level. While it can be argued that such models are unrealistic in some situations, in the applications detailed in Chapter 2 the model appears to provide a realistic representation of the choice process. A detailed description of the relationship between independence from irrelevant alternatives and rank data models is given in Marden (1995, Section 5.13.1).

3.2 Benter's Model for Rank Data

The Plackett-Luce model suffers from the property that the probability of an object with a low support parameter being ranked highly is too small. Similarly, under the Plackett-Luce model

$$\mathbf{P}\{\text{choosing object } j \text{ at level } t\} > \mathbf{P}\{\text{choosing object } j \text{ at level } s\}$$

for any $t > s$; this is not always a good model for many ranking contexts. Benter (1994), within the context of modelling horse races, proposed a variant of the Plackett-Luce model to overcome these issues. The Benter model has two parameters: a *support parameter* $\underline{p} = (p_1, p_2, \ldots, p_N)$ where $\sum_{j=1}^{N} p_j = 1$ and a *dampening parameter*

$$\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$$

where N denotes the total number of objects available for selection. The support parameter p_j represents the probability of object j being given a first preference; the dampening parameters model the way in which some preferences may be chosen less carefully than other preferences. Under the Benter model, the probability of judge i's ranking $\underline{x}_i$ is:

$$\mathbf{P}\{\underline{x}_i | \underline{p}, \underline{\alpha}\} \quad = \quad \prod_{t=1}^{n_i} \frac{p_{c(i,t)}^{\alpha_t}}{\sum_{s=t}^{N} p_{c(i,s)}^{\alpha_t}}$$

where n_i denotes the total number of preferences expressed by judge i. It is reasonable to assume that $0 \leq \alpha_t \leq 1$, which makes lower preference choices at least as random as higher preference ones (see Proposition 1, Appendix B). In any case, $\alpha_1 \equiv 1$ and $\alpha_N \equiv 0$ for all models; this avoids over parameterization of the model.

Under the Benter model the log odds of selecting object j over object l at choice level t is $\alpha_t \log(p_j/p_l)$. Thus the tth level dampening parameter α_t can be interpreted as how the log odds of selecting object j over object l is affected by the selection being made at choice level t. Since α_1 is constrained to be 1 for identifiability reasons, at the first choice level the log odds is unaffected. An α_2 value of 0.8, for example, indicates that the log odds is 'dampened' by a fifth to model the manner in which the second selection was made with less certainty than the first. Thus the probabilities in the Benter model have greater entropy than the Plackett-Luce model (see Proposition 1, Appendix B). The Plackett-Luce model is in fact a special case of the Benter model with $\underline{\alpha} = \underline{1} \equiv (1, 1, \ldots, 1)$. The estimation of dampening parameters is of interest as the care with which judges express their preferences is an attribute of the ranking process which is of interest.

As with the Plackett-Luce model, one concern associated with choice models is the issue of independence from irrelevant alternatives (IIA). The Benter model exhibits IIA within choice levels as the ratio of the probability of choosing one alternative over another is independent of the other available alternatives. In the Benter model the ratio of the probabilities varies with choice level due to the dampening parameter in the model. While it can be argued that IIA is an unsatisfactory property in some situations, again in the applications under study here the models appear to give a realistic representation of the choice process.

3.3 Classical Inference

Both the manner in which the rank statistical models are constructed and the way in which inferences are subsequently drawn can be conducted within a classical or Bayesian framework.

Classical inference is based on a frequency interpretation of probability i.e. the probability of an event x is the proportion of times that x has been observed to occur in an infinite sequence of trials. Bayesian inference takes a subjective view of probability which measures the degree of belief an individual has in a proposition. Both the data observed and any prior knowledge the individual has about the proposition are involved in the degree of belief.

More specifically, the fundamental difference between classical and Bayesian inference is that model parameters are treated as random variables when conducting

Bayesian inference whereas parameters take (unknown) unique values in the classical paradigm (O'Hagan and Forster, 2004).

In 1922, Fisher introduced the method of maximum likelihood to make inferences about the parameters of an underlying model given a data set. Maximum likelihood methods lie within the classical framework as parameters are treated as unknown unique values and no prior information is taken into account when drawing inferences from the parameter estimates.

Definition 1 *Assume $\underline{x} = (x_1, \ldots, x_M)$ is a sample of size M drawn from a probability density $f(x|\theta)$. The joint density of the observed data is then $f(x_1, \ldots, x_M|\theta)$. As a function of the parameter θ and treating $\underline{x}$ as fixed the **likelihood function** is*

$$\mathcal{L}(\theta) = f(x_1, \ldots, x_M|\theta) \quad \blacksquare$$

Under the maximum likelihood method the value of θ which maximizes $\mathcal{L}(\theta)$ is reported as the the maximum likelihood estimator $\hat{\theta}$. Often it is straight forward to maximize the likelihood $\mathcal{L}(\theta)$ with respect to the model parameters. In future chapters however this is not the case and numerical techniques are employed to maximize the likelihood. In Chapters 4, 5 and 7 the expectation-maximization (EM) algorithm is employed as a method of producing maximum likelihood estimators (MLEs).

3.3.1 The EM Algorithm

Dempster et al. (1977) introduced the EM algorithm as a technique to produce MLEs for problems where the data is incomplete. Latent variables or missing data points can be an intrinsic feature of the problem under investigation or they can be artificially imputed. Such missing data both makes implementation of the EM algorithm feasible and can often have useful interpretations.

In principle, the EM process is straight forward to implement and has broad applicability. It is a two step iterative algorithm consisting of an expectation (E) step followed by a maximization (M) step. Generally, during the E step the expected value of the log likelihood of the complete data (i.e. the observed and unobserved data) is computed. In the M step the expected log likelihood is maximized producing

MLEs of the model parameters. In practice, the imputation of latent variables often makes maximization of the expected likelihood feasible. The parameter estimates produced during the M step are then used in a new E step and the cycle continues until convergence.

While the EM algorithm will not decrease the observed data likelihood function (see Dempster et al. (1977)) there is no guarantee that the resulting sequence of parameter estimates will converge to the MLE. Multiple runs of the algorithm from different parameter starting values should help avoid local maxima. Dempster et al. (1977) defined the generalized EM (GEM) algorithm which increases the value of the likelihood at the M step without actually maximizing it. Many other variants of the EM algorithm have also emerged; for example the expectation and conditional maximization (ECM) algorithm (Meng and Rubin, 1993) and the stochastic EM (SEM) algorithm (Celeux and Diebolt, 1985) among others. McLachlan and Krishnan (1997) provides an excellent review.

An important feature of iterative algorithms is the determination of convergence. Aitken's accleration criterion (Böhning et al., 1994; Lindsay, 1995; McLachlan and Peel, 2000) was employed throughout the following chapters as a convergence criterion. Denote by $\{l^{(t)}\}$ the sequence of log likelihood values which emerge from the M steps of the EM algorithm. Assuming l^* is the limiting value of these log likelihoods it follows that

$$\begin{aligned} l^{(t+1)} - l^* &\approx a(l^{(t)} - l^*) \qquad \text{for } 0 \leq a \leq 1 \\ \Rightarrow l^{(t+1)} - l^{(t)} &\approx (1-a)(l^* - l^{(t)}). \end{aligned}$$

Thus if $a \approx 1$ a small increase in the log likelihood value does not necessarily imply that the EM algorithm is close to convergence. It can be shown that estimating a by the ratio of successive increments leads to the Aitken acceleration estimate of the limiting log likelihood value

$$l^* \approx l^{(t)} - \frac{\{l^{(t+1)} - l^{(t-1)}\}}{\{l^{(t)} - l^{(t-1)}\}}\{l^{(t+1)} - l^{(t)}\}.$$

The EM algorithm should be stopped if $|l^{(t+1)} - l^*| < \epsilon$ where ϵ is a pre-specified tolerance level. Since this criterion is not an exact indicator of convergence multiple runs of the algorithm with random starts must be employed.

Within Chapters 4, 5 and 7 the specific calculations performed by the EM algorithm are fully detailed.

3.4 Bayesian Inference

Under the Bayesian philosophy prior beliefs about various hypotheses are updated in light of relevant observed data to produce posterior beliefs. Bayesian analysis is subjective in nature in that one's personal prior beliefs need not agree with another's prior beliefs.

Assume interest lies in the value of the $k \geq 1$ dimensional parameter θ which describes the underlying mechanism of the process of interest. The Bayesian method generally comprises of the following steps:

1. **Summarizing prior knowledge**

   Prior knowledge or information about the parameter values can be based on personal experience or on an expert's opinion. Such beliefs are characterized by a probability density $\mathbf{P}\{\theta\}$ known as the prior density. The manner in which prior distributions are selected and specified (i.e. prior elicitation) requires careful attention (see Chapter 3.4.1).

2. **Formation of the likelihood function**

   Assume M relevant data values $\mathbf{x} = (\underline{x}_1, \ldots, \underline{x}_M)$ are observed which depend on the unknown parameter θ. The likelihood function (see Definition 1) of the data $\mathcal{L}(\theta) = \mathbf{P}\{\mathbf{x}|\theta\}$ is formed by calculating the joint probability density of the observed data given the parameters. When $\mathbf{P}\{\mathbf{x}|\theta\}$ is considered as a function of the data given θ it is a probability density and thus properties such as integration to unity hold. Alternatively, when $\mathbf{P}\{\mathbf{x}|\theta\}$ is considered as a function of the parameter θ given the data such properties do not necessarily hold and $\mathbf{P}\{\mathbf{x}|\theta\}$ is known as the likelihood function.

3. **Formation of the posterior**

   Bayes theorem provides the tool for combining the two sources of information i.e. the prior knowledge $\mathbf{P}\{\theta\}$ about the parameters and the likelihood function which expresses the relationship between the data and θ. Bayes theorem states

$$\begin{aligned}\mathbf{P}\{\theta|\mathbf{x}\} &= \frac{\mathbf{P}\{\theta\}\mathbf{P}\{\mathbf{x}|\theta\}}{\mathbf{P}\{\mathbf{x}\}} \\ &\propto \mathbf{P}\{\theta\}\mathbf{P}\{\mathbf{x}|\theta\} \\ &\propto \text{ prior } \times \text{ likelihood}\end{aligned}$$

where

$$\mathbf{P}\{\mathbf{x}\} = \begin{cases} \int \mathbf{P}\{\theta\}\mathbf{P}\{\mathbf{x}|\theta\}d\theta & \text{in the continuous case} \\ \sum_\theta \mathbf{P}\{\theta\}\mathbf{P}\{\mathbf{x}|\theta\} & \text{in the discrete case.}\end{cases}$$

Bayes theorem constructs the posterior density $\mathbf{P}\{\theta|\mathbf{x}\}$ which is a summary of all knowledge about the parameter θ subsequent to observing $\mathbf{x}$.

4. **Inference**

   The posterior distribution is a comprehensive inference statement about the model parameter θ. Any summary of the posterior distribution is useful eg. moments, quantiles, highest posterior regions and credible intervals (see Lee (2004)).

3.4.1 Prior Elicitation

Prior elicitation is the process of constructing a prior distribution which reflects your background information. Two aspects require consideration when constructing a prior distribution; firstly the choice of prior distribution and secondly the specification of the hyperparameters of the prior distribution.

The type of distribution chosen is generally governed by mathematical tractability constraints. Such requirements often necessitate the use of conjugate priors.

Definition 2 *A class $\prod$ of prior densities is said to form a **conjugate family** if the posterior density $\mathbf{P}\{\theta|\mathbf{x}\}$ is in the class $\prod$ for all $\mathbf{x}$ whenever $\mathbf{P}\{\theta\} \in \prod$.*

Thus when using conjugate priors the only change when updating the prior distribution to the posterior distribution is a change of parameter values. In cases where a conjugate prior is not justifiable or where it is infeasible, selection of the type of distribution should not be governed by mathematical tractability; sampling from a non-standard posterior distribution is possible using numerical methods.

Selection of the prior hyperparameters also requires care. Crude methods such as eliciting opinions on various properties (eg. moments) of the parameter θ and equating these opinions to their theoretical value (as a function of θ) can be used. Sensitivity analysis is important in prior elicitation i.e. the effect of changes in values of the prior's hyperparameters on any subsequently calculated posterior distribution should be investigated.

When eliciting priors it may also be the case that the expert/researcher cannot provide much background information about the parameter θ. Thus a suitable 'non-informative' prior is necessary. Specifying large variance hyperparameters induces a prior distribution which is flat over realistic values of θ. If the range of possible values of θ is finite, a uniform distribution on this range would be a suitable non-informative prior. If the parameter takes values over an infinite range the 'improper' prior

$$\mathbf{P}\{\theta\} = \frac{1}{c} \qquad \text{where } -\infty < \theta < \infty \text{ and } c = \text{constant}$$

is often used. This is an improper prior in that it's integral is infinite. Although the prior is not a valid probability density function it is possible that the posterior density may be.

3.4.2 Markov Chains

In order to make inferences about the posterior distribution of interest, directly sampling from the posterior itself allows the calculation of approximations of integration-based summaries (such as posterior moments and marginal densities). For non-standard posterior distributions sampling is not always a straight forward procedure. Markov chain Monte Carlo (MCMC) methods are a collection of techniques which allow (in an asymptotic sense) the sampling of dependent observations from a density of interest. They cope easily with high dimensional problems and non-standard posterior distributions.

A key notion embedded in MCMC theory is that of a stochastic process and a Markov chain.

Definition 3 *A **stochastic process** is a set of random variables $\{x_t; t \in T\}$ where T is called the index set. Each x_t takes a value (i.e. a **state**) in a set S known as the **state space**.* ▮

A discrete time stochastic process occurs when T is a countable set. In what follows a discrete state space is assumed but the theory is easily extended to the case where the state space S is continuous.

Definition 4 *A stochastic process $\{x_t; t \in T\}$ is called a **Markov chain** with countable state space S if:*

1. $\mathbf{P}\{x_t \in S\} = 1 \;\; \forall \; t \geq 0$

2. *the distribution of x_{t+1} is independent of all previous random variables but x_t i.e.*

$$\mathbf{P}\{x_{t+1}|x_0, x_1, \ldots, x_t\} = \mathbf{P}\{x_{t+1}|x_t\}$$

*i.e. the **Markov property** holds.* ▮

A Markov chain where the distribution of x_{t+1} given x_t is independent of t is said to be a homogeneous chain. In such a case

$$P_{xy} = \mathbf{P}\{x_{t+1} = y|x_t = x\}$$

is known as the transition probability. As P_{xy} is a probability it follows that $P_{xy} \geq 0 \;\forall x, y \in S$ and $\sum_{y \in S} P_{xy} = 1 \;\forall x \in S$. The transition matrix

$$P \;\; = \;\; \begin{pmatrix} P_{11} & P_{12} & \ldots & P_{1s} \\ P_{21} & P_{22} & \ldots & P_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ P_{s1} & P_{s2} & \ldots & P_{ss} \end{pmatrix}$$

is known as the transition or Markov matrix. The t-step transition probability P^t_{xy} denotes the probability of moving from state x to state y in t steps.

Under certain conditions P^t_{xy} converges asymptotically to a unique distribution $\pi(\cdot)$. $\pi(\cdot)$ is known as the stationary (or invariant) distribution of the Markov chain and is independent of both t and the starting state x_0. To explain the conditions necessary for the convergence of a Markov chain to a stationary distribution some definitions are required.

Definition 5

1. A Markov chain is **irreducible** if for any starting state $x \in S$ there exists t such that $P^t_{xy} \geq 0 \forall y \in S$. In other words for any starting state x the chain can eventually reach every region of the state space S with positive probability.

2. A Markov chain is said to be **periodic** if it cycles between disjoint subsets of the state space S. Any Markov chain which is not periodic is called **aperiodic**.

3. A state x is said to be **recurrent** if the Markov chain starting in state x returns to state x with probability 1. If the expected time to return to state x is finite x is said to be positive recurrent. A state which is not recurrent is said to be **transient**. It follows that a Markov chain is said to be (positive) recurrent if all states are (positive) recurrent.

4. A Markov chain with stationary distribution $\pi(\cdot)$ is defined to be **ergodic** if it is irreducible, aperiodic and positive recurrent. $\blacksquare$

For an ergodic Markov chain $|P^t_{xy} - \pi(\cdot)| \to 0$ as $t \to \infty$ for any initial state x_0. Thus the limiting distribution of the chain is the stationary distribution. An ergodic Markov chain tends to sample from $\pi(\cdot)$ as the number of transitions in the chain tends to infinity.

3.4.3 Markov Chain Monte Carlo Methods

Integration based summaries of posterior densities can be formed using Monte Carlo integration. Monte Carlo integration asserts that the expectation $\mathbf{E}[f(\underline{x})]$ can be approximated by

$$\mathbf{E}[f(\underline{x})] \approx \frac{1}{M} \sum_{i=1}^{M} f(x_i)$$

where the points $\underline{x} = (x_1, \ldots, x_M)$ are independent samples drawn from the density of interest. Basic sampling techniques such as the acceptance-rejection technique and the inverse transform method are available but for non-standard distributions these become complicated. Press et al. (1996) provide details of sampling from a range of commonly used densities.

To sample from high dimensional or complex distributions Markov chain Monte Carlo (MCMC) techniques are used. Suppose it was possible to simulate from a homogeneous, ergodic Markov chain whose unique stationary distribution $\pi(\cdot)$ was in fact the posterior (or target) distribution of interest. Subsequent to a burn-in period of the Markov chain (i.e. a period of transitions of the chain after which the stationary distribution $\pi(\cdot)$ is reached) any random variables sampled from the chain will be dependent samples drawn from an approximation of the target distribution $\pi(\cdot)$. Such samples could be used in Monte Carlo integration to estimate expectations; since the sampled random variables are in fact dependent, Monte Carlo integration can only be employed if the samples are representative of the full support of the target distribution.

Construction of a Markov Chain

To construct a Markov chain whose stationary distribution is the required posterior distribution of interest the relevant transition probabilities are required. It turns out (see Tierney (1994)) that probabilities which satisfy a condition known as detailed balance or time reversibility i.e.

$$\pi(x_t)\mathbf{P}\{x_{t+1}|x_t\} \quad = \quad \pi(x_{t+1})\mathbf{P}\{x_t|x_{t+1}\}$$

are suitable transition probabilities. Detailed balance means that the probability of being in state x_t at time t and moving to state x_{t+1} at time $t+1$ (when the initial probabilities are given by the stationary distribution $\pi(x_t)$) is the same as starting at x_{t+1} and ending in state x_t.

To illustrate why the appropriate stationary distribution is reached when detailed balance holds, denote by r_{xx} the probability that the chain remains in state x. Then

$$\mathbf{P}\{x_{t+1} = y|x_t = x\} = P_{xy} = P^*_{xy} + r_{xx}\delta_{xy}$$

where $P^*_{xx} = 0$ and

$$\delta_{xy} \quad = \quad \begin{cases} 1 & \text{if } y = x \\ 0 & \text{otherwise.} \end{cases}$$

Since the chain must either remain in the same state or move to a new state it follows that

$$1 = \sum_y P_{xy} \quad = \quad \sum_y P^*_{xy} + r_{xx} \qquad (3.2)$$

Hence

$$\begin{aligned}\sum_y \pi(x_t)P_{xy} &= \sum_y \pi(x_t)P^*_{xy} + \sum_y \pi(x_t)r_{xx}\delta_{xy} \\ &= \sum_y \pi(x_{t+1})P^*_{yx} + \pi(x_{t+1})r_{yy} \quad \text{(by detailed balance)} \\ &= \pi(x_{t+1})(1-r_{yy}) + \pi(x_{t+1})r_{yy} \quad \text{(by (3.2))} \\ &= \pi(x_{t+1}).\end{aligned}$$

Thus given transition probabilities P_{xy} which satisfy detailed balance the marginal distribution of x_{t+1} can be obtained where x_t comes from the stationary distribution $\pi(\cdot)$. Once x_t is sampled from the stationary distribution any subsequent samples will also be from the stationary distribution. This proves that given detailed balance holds the stationary distribution is $\pi(\cdot)$ but is not a proof that P^t_{xy} will converge to the stationary distribution; see Gilks et al. (1996) for details.

The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a Markov chain method to simulate multivariate distributions. It was first proposed by Metropolis et al. (1953), generalized by Hastings (1970) and is now commonly referred to as the Metropolis-Hastings algorithm. Chib and Greenberg (1995) provide an introductory article; Gilks et al. (1996) provides further detail. Suppose the posterior distribution $\pi(\cdot)$ is the stationary distribution to be sampled from. Given that the Markov chain is in state x_t, the algorithm begins by drawing a proposal state, y, for state x_{t+1} from a candidate density $q(\cdot|x_t)$. If the density $q(\cdot|x_t)$ satisfies detailed balance, then the algorithm detailed below proceeds safe in the knowledge that (asymptotically) sampled values will be drawn from the desired posterior distribution. More commonly it is the case that

$$\pi(x_t)q(x_{t+1}|x_t) \;>\; \pi(x_{t+1})q(x_t|x_{t+1}) \tag{3.3}$$

which means detailed balance does not hold and that the chain moves from state x_t to state x_{t+1} more often than it moves from x_{t+1} to x_t. An ‘acceptance probability’ $\alpha(x_t, x_{t+1})$ is introduced which is the probability of a move from x_t to x_{t+1}. Since too many moves are made from x_t to x_{t+1}, and as $\alpha(\cdot,\cdot)$ is a probability, $\alpha(x_{t+1}, x_t)$

is made as large as possible i.e. $\alpha(x_{t+1}, x_t) = 1$. To satisfy detailed balance it must follow that

$$\begin{aligned} \pi(x_t)q(x_{t+1}|x_t)\alpha(x_t, x_{t+1}) &= \pi(x_{t+1})q(x_t|x_{t+1})\alpha(x_{t+1}, x_t) \\ \Rightarrow \quad \alpha(x_t, x_{t+1}) &= \frac{\pi(x_{t+1})q(x_t|x_{t+1})}{\pi(x_t)q(x_{t+1}|x_t)}. \end{aligned}$$

If inequality (3.3) is reversed $\alpha(x_t, x_{t+1}) = 1$ and the derivation of $\alpha(x_{t+1}, x_t)$ follows. Thus to ensure detailed balance holds the acceptance probability is set to be

$$\alpha(x_t, x_{t+1}) = \min\left[1, \frac{\pi(x_{t+1})q(x_t|x_{t+1})}{\pi(x_t)q(x_{t+1}|x_t)}\right]. \tag{3.4}$$

The probability of moving from state $x_t = x$ to state $x_{t+1} = y$ is $P_{xy} = q(y|x)\alpha(x, y)$; it follows that the probability of remaining in state x is $r_{xx} = 1 - \sum_y q(y|x)\alpha(x, y)$. These transition probabilities make up the transition matrix.

In summary the Metropolis-Hastings algorithm proceeds as follows:

1. Given the chain is in state x_t, for state x_{t+1} generate a proposal state y from a candidate density $q(\cdot|x_t)$.

2. Compute the acceptance probability $\alpha(x_t, x_{t+1})$.

3. Generate a value $u \sim \text{Uniform}(0, 1)$.

4. If $u \leq \alpha(x_t, x_{t+1})$ then define $x_{t+1} = y$, otherwise define $x_{t+1} = x$.

5. Return the sequence $x_1, \ldots, x_M$ where M is the number of iterations performed subsequent to burn in.

If the candidate density employed is symmetric i.e. $q(x_{t+1}|x_t) = q(x_t|x_{t+1})$ the acceptance probability reduces to

$$\alpha(x_t, x_{t+1}) = \min\left[1, \frac{\pi(x_{t+1})}{\pi(x_t)}\right]$$

which is the form originally proposed by Metropolis et al. (1953). If such a symmetric density is employed all uphill moves are accepted (i.e. when $\pi(x_{t+1}) > \pi(x_t)$); some downhill moves are also accepted (i.e. when $\pi(x_{t+1}) < \pi(x_t)$). A ‘random walk’ Metropolis algorithm has candidate density of the form $q(x_{t+1}|x_t) = q(|x_{t+1} - x_t|)$. This is a symmetric density and thus the algorithm also reduces to the Metropolis algorithm. Many other families of candidate-generating densities are available (see Chib and Greenberg (1995) and Lee (2004)). The rate of convergence of the chain to the stationary distribution depends on the relationship between the candidate density and the target density; Gilks et al. (1996) provides more detail on strategically choosing $q(\cdot|\cdot)$.

The choice of parameters within the candidate density is of critical importance. Assume the chain is currently located near the mode of the target density. If the spread of the candidate distribution is small then generated candidates will be close to the current state. It will therefore take the chain a long time to explore the full support of the target density (i.e. poor 'mixing' will occur) and low probability areas will be undersampled. This is often indicated by a high acceptance rate (i.e. the percentage of moves accepted). Conversely, if the spread of the distribution is too large, proposed states will have low probability of being accepted giving low acceptance rates. Roberts et al. (1994) and Chib and Greenberg (1995) provide further discussion of this issue.

The Gibbs Sampler

A single component Metropolis-Hastings algorithm divides a K dimensional random variable x into blocks and updates x block by block. A special case of the single component Metropolis-Hastings algorithm is the Gibbs sampler. The term 'Gibbs sampling' arose in Geman and Geman (1984) who analyzed the Gibbs distribution in an image-processing context. Gelfand and Smith (1990) later revealed its general applicability within mainstream statistics.

Let $\mathbf{x} = (x_1, \ldots, x_K)$ denote a random variable with joint density $\pi(x_1, \ldots, x_K)$. Interest lies in sampling from the marginal density

$$\pi(x_1) = \int \ldots \int \pi(x_1, \ldots, x_K) dx_2 \ldots dx_K.$$

Denote by $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_K)$. In Gibbs sampling the proposal distribution for updating the ith component of $\mathbf{x}$ with target density $\pi(\cdot)$ is

$$q(y_i|x_i, \mathbf{x}_{-i}) = \pi(y_i|\mathbf{x}_{-i})$$

i.e. the full conditional distribution of the ith component of $\mathbf{x}$ conditional on the remaining components of $\mathbf{x}$. Substituting this proposal density into (3.4) gives an acceptance probability of 1. Thus in the Gibbs sampler all moves proposed are accepted.

Explicitly the Gibbs sampler takes the following steps:

1. Choose arbitrary starting values $\mathbf{x}^{(0)}$ for $\mathbf{x}$. Let $t = 1$.

2. Repeat over a burn in period and until convergence:

   - Generate $x_1^{(t)}$ from $\pi(y_1|x_2^{(t-1)}, \ldots, x_K^{(t-1)})$.
   - Generate $x_2^{(t)}$ from $\pi(y_2|x_1^{(t)}, x_3^{(t-1)}, \ldots, x_K^{(t-1)})$.
     $\vdots$
   - Generate $x_K^{(t)}$ from $\pi(y_K|x_1^{(t)}, \ldots, x_{K-1}^{(t)})$.
   - Let t = t+1.

$x_i^{(t)}$ is effectively a sample point from the marginal $\pi(x_i)$ for large t. Thus the Gibbs sampler consists only of sampling from full conditional distributions. Casella and George (1992) provides a clear explanation of how and why the Gibbs sampler works.

3.5 Model Selection Techniques

In future chapters many different model types are fitted to rank data — a criterion is required for comparing the fitted models. Information criteria are often used to compare models; they are motivated by the aim of minimizing the Kullback-Leibler information (Kullback and Leibler, 1951) of the true model from the fitted model (see McLachlan and Peel (2000), Section 6.8). Many types of information criteria have been developed but their general structure is one which rewards model fit while penalizing model complexity.

The Bayesian Information Criterion (BIC) (Schwartz, 1978) is a widely used criterion which compares models. The usual justification for the use of BIC is that, for regular problems, it provides an approximation to the Bayes factor for comparing models under certain prior assumptions (Kass and Raftery, 1995). The BIC is defined to be

$$\text{BIC} \quad = \quad 2(\text{maximized likelihood}) - (\text{number of parameters}) \log(M) \qquad (3.5)$$

where M is number of data points. The first term on the right-hand side of (3.5) measures model fit; the second term is the penalty term which measures model complexity. Discussion of the use of BIC within the specific framework of mixture models is given in Chapter 4.6.

Many alternative model selection information criteria exist — McLachlan and Peel (2000) provide a good review. Akaike's Information Criterion (AIC) (Akaike, 1973, 1974)

$$\text{AIC } = 2(\text{maximized likelihood}) - 2(\text{number of parameters})$$

is similar to the BIC but AIC has a smaller penalty when $M > e^2$. AIC is often inconsistent and tends to overfit models. The Integrated Complete Likelihood (ICL) (Biernacki et al., 2000) is also similar to the BIC but the integrated complete likelihood is used rather than the integrated observed likelihood and the penalty term is heavier in ICL than in BIC.

The Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) uses a Bayesian measure of model fit. It penalizes the posterior mean deviance $(D(\cdot))$ of a model by the 'effective number of parameters' (p_D). The effective number of parameters is derived to be the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. Explicitly for data $\mathbf{x}$ and parameter θ the DIC is

$$\text{DIC} \;=\; \overline{(D(\theta))} \;+\; 2p_D$$

where $\quad D(\theta) = -2\log(\mathbf{P}\{\mathbf{x}|\theta\}) \quad$ and $\quad p_D = \overline{D(\theta)} - D(\overline{\theta})$. The criterion has an approximate decision theoretic justification.

Model selection tools which are not information criteria are also examined. Pritchard et al. (2000) suggested a Bayesian based criterion which emerged to be a variant of the DIC — they penalized the mean of the Bayesian deviance by a quarter of its variance rather than by the effective number of parameters.

$$\text{Pritchard et. al's criterion} \;=\; \overline{D(\theta)} + \text{var}(D(\theta))/4.$$

Cross-validated likelihood was also proposed as a model selection tool (Smyth, 2000). Models are judged on their performance in out-of-sample prediction, as estimated in a cross validation manner. It is proposed as a practical alternative to the Bayesian BIC approach; it is more computationally expensive however and also tends to overfit.

Different model selection tools were utilized in different contexts throughout future chapters.

Chapter 4

Mixtures of Plackett-Luce Models

Mixture models have recently come to the fore as a tool for providing theoretically solid model-based clustering techniques (see Banfield and Raftery (1993), Bensmail et al. (1997), Fraley and Raftery (1998) and Fraley and Raftery (1999) for example). The finite mixture model provides a model-based framework in which rigorous statements may be made about the presence of groups within a population and about the structure of these groups. Statements are based on statistical theory rather than being descriptive in nature. Motivation for the use of mixture models when clustering data is given by Aitkin et al. (1981) where they state "*when clustering samples from a population, no cluster method is a priori believable without a statistical model*".

To describe and illustrate the use of mixtures of Plackett-Luce models for rank data the CAO data as detailed in Chapter 2.1 is used. The population which generated this data contains students with many different characteristics and vocational callings. The course choices of these students are modelled using a mixture model, so that groups of students with different choice behaviour can be discovered. Much of the work in this chapter is reported in Gormley and Murphy (2006*c*).

4.1 Mixture Models

It is assumed that the course choices made by the CAO applicants form a sample from a heterogeneous population. This assumption is justified because of the differing backgrounds and interests of the applicants. Mixture models appropriately model situations where data are collected from heterogeneous populations. Therefore, it appears natural to use a mixture model to model college applications data.

A finite mixture model assumes that the population consists of a finite collection of K components (or groups). It is assumed that the (unknown) probability of belonging to component k is π_k — these values are termed the *mixing proportions.* By definition $\pi_k \geq 0$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. In addition, observation $\underline{x}_i$ within component k has a probability density $f(\underline{x}_i | \underline{p}_k)$, where $\underline{p}_k$ are unknown parameters. Hence, the resulting model for a single observation $\underline{x}_i$ is

$$f(\underline{x}_i) = \sum_{k=1}^{K} \pi_k f(\underline{x}_i | \underline{p}_k).$$

Given the data $\mathbf{x} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_M)$ for M (assumed independent) applicants and a mixture modelling framework the likelihood of the data is

$$L(\pi_1, \pi_2, \ldots, \pi_K; \underline{p}_1, \underline{p}_2, \ldots, \underline{p}_K | \mathbf{x}) \quad = \quad \prod_{i=1}^{M} \sum_{k=1}^{K} \pi_k f(\underline{x}_i | \underline{p}_k). \tag{4.1}$$

Extensive reviews of mixture modelling are given by McLachlan and Peel (2000) and Titterington et al. (1985); in addition, an excellent overview of using mixture models to produce model-based methods for clustering is given by Fraley and Raftery (2002). Previous applications of mixture models for analyzing rank data are given in Marden (1995, Section 10.2) and Murphy and Martin (2003) amongst others.

4.2 The Plackett-Luce Model

An appropriate density for each component of the mixture model must be specified. Each applicant's data consists of a ranking of up to ten courses. Hence, a model that is appropriate for modelling rank data is required. One such model is the Plackett-Luce model for rank data (see Chapter 3.1).

The Plackett-Luce model is parameterized by the support parameter $\underline{p} = (p_1, \ldots, p_N)$ where N denotes the total number of courses from which the applicants chose. p_j denotes the probability of course j being ranked in first position. Within the context of the CAO data, the maximum number of courses that may be ranked on an application form is denoted by $n = 10$.

Within the mixture modelling context, let $p_{kc(i,t)}$ denote the probability of the course chosen in tth position by applicant i being selected first, given that the applicant belongs to the kth component. The rank t of a selected course must be less than or equal to n_i, where n_i is the number of choices expressed by applicant i. The Plackett-Luce model then suggests the probability of applicant i's ranking conditional on belonging to component k is

$$\mathbf{P}\{\underline{x}_i|\underline{p}_k\} = \prod_{t=1}^{n_i} \frac{p_{kc(i,t)}}{\sum_{s=t}^{N} p_{kc(i,s)}}.$$

By fitting a mixture of Plackett-Luce models to the CAO application data homogeneous groups of applicants which follow characteristic Plackett-Luce densities will be highlighted. The estimation of both the number of groups in the population and the associated parameters of the characteristic Plackett-Luce densities is required.

4.3 Model Fitting

The EM algorithm (Dempster et al., 1977) is a widely used tool for obtaining maximum likelihood estimates in missing data problems; mixture models can be formulated as having the component membership of each observation as missing data. Maximization of the likelihood function is simplified by augmenting the data to include the missing membership variables. Furthermore, the EM algorithm provides estimates not only of the model parameters but also of the unknown component memberships of the observations.

The term 'complete' data refers to the combination of both the observed preferences and the missing membership variables. It is denoted by $(\mathbf{x}, \mathbf{z}) = \{(\underline{x}_1, \underline{z}_1), \ldots, (\underline{x}_M, \underline{z}_M)\}$, where $\underline{x}_i$ is applicant i's application and

$$\underline{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK}) \qquad \forall\ i = 1, \ldots, M$$

with

$$z_{ik} = \begin{cases} 1 & \text{if applicant } i \text{ belongs to component } k \\ 0 & \text{otherwise.} \end{cases}$$

The missing data $\mathbf{z}$ can be interpreted as an indicator of component membership. On convergence of the EM algorithm, the estimated values of z_{ik} are the conditional probabilities of applicant i belonging to component k.

The EM algorithm involves two steps, an expectation step (E step) followed by a maximization step (M step). In the context of finite mixture models, the expectation step estimates the unknown values of each z_{ik}. The maximization step then proceeds to maximize the complete data log likelihood to estimate the model parameters.

The complete data log likelihood is formulated as follows. Under the Plackett-Luce model, given that applicant i belongs to group k, the probability of applicant i's ranking is

$$\mathbf{P}\{\underline{x}_i|\underline{p}_k\} \;=\; \prod_{t=1}^{n_i} \frac{p_{kc(i,t)}}{\sum_{s=t}^{N} p_{kc(i,s)}}.$$

Accounting for the missing component membership indicator $\underline{z}_i$, it follows that the complete data log likelihood for applicant i is

$$\begin{aligned}\mathbf{P}\{\underline{x}_i, \underline{z}_i|\mathbf{p}\} &= \mathbf{P}\{\underline{x}_i|\underline{z}_i\}\mathbf{P}\{\mathbf{z}_i\} \\ &= \left[\prod_{k=1}^{K}\left\{\prod_{t=1}^{n_i} \frac{p_{kc(i,t)}}{\sum_{s=t}^{N} p_{kc(i,s)}}\right\}^{z_{ik}}\right]\left\{\prod_{k=1}^{K}(\pi_k)^{z_{ik}}\right\}.\end{aligned}$$

Hence, the complete data log likelihood for all applicants is

$$\begin{aligned} l &= \log\left[\prod_{i=1}^{M}\prod_{k=1}^{K}\left\{\pi_k \prod_{t=1}^{n_i} \frac{p_{kc(i,t)}}{\sum_{s=t}^{N} p_{kc(i,s)}}\right\}^{z_{ik}}\right] \\ &= \sum_{i=1}^{M}\sum_{k=1}^{K} z_{ik}\left\{\log \pi_k + \sum_{t=1}^{n_i}\log p_{kc(i,t)} - \sum_{t=1}^{n_i}\log\sum_{s=t}^{N} p_{kc(i,s)}\right\}.\end{aligned}$$

The EM algorithm is an iterative technique and continually repeats the E and M steps until convergence to stable estimates and/or a predetermined criterion is achieved. Aitken's acceleration criterion (see Chapter 3.3.1) was employed in this application as a convergence criterion.

Specifically, the EM algorithm proceeds as follows:

0. **Initialize:** Choose starting values for $\underline{\pi}^{(0)}$ and $\mathbf{p}^{(0)}$. Let $l = 0$.

1. **E step:** Compute the values

$$\hat{z}_{ik} = \frac{\pi_k^{(l)}\mathbf{P}\{\underline{x}_i|\underline{p}_k^{(l)}\}}{\sum_{k'=1}^{K}\pi_{k'}^{(l)}\mathbf{P}\{\underline{x}_i|\underline{p}_{k'}^{(l)}\}} \tag{4.2}$$

for $i = 1, \ldots, M$ and $k = 1, \ldots, K$ where the value $\hat{z}_{ik}$ is the estimated posterior probability of observation i belonging to group k. The likelihood values (4.1) necessary to calculate Aitken's acceleration criterion are easily obtained within this step from the numerator of (4.2).

2. **M step:** Maximize the function

$$\sum_{i=1}^{M}\sum_{k=1}^{K}\hat{z}_{ik}\left\{\log\pi_k+\sum_{t=1}^{n_i}\log p_{kc(i,t)}-\sum_{t=1}^{n_i}\log\sum_{s=t}^{N}p_{kc(i,s)}\right\}$$

   to yield new parameter estimates $\underline{\pi}^{(l+1)}$ and $\mathbf{p}^{(l+1)}$. Increment l by 1.

3. **Convergence:** Repeat the E step and M step until convergence (as deemed by Aitken's acceleration criterion). The final parameter values are the maximum likelihood estimates $\hat{\underline{\pi}}$ and $\hat{\mathbf{p}}$.

The E step is relatively straightforward when fitting a mixture of Plackett-Luce models. Optimization with respect to the membership proportions $\underline{\pi}=(\pi_1,\ldots,\pi_K)$ in the M step is also straightforward. To obtain $\pi_k^{(l+1)}$ the expected complete data log-likelihood function

$$Q(\underline{\pi}) \quad = \quad \sum_{i=1}^{M}\sum_{k=1}^{K}\hat{z}_{ik}\left\{\log\pi_k+\sum_{t=1}^{n_i}\log p_{kc(i,t)}-\sum_{t=1}^{n_i}\log\sum_{s=t}^{N}p_{kc(i,s)}\right\}$$

is maximized with respect to π_k, subject to the constraint $\sum_{k=1}^{K}\pi_k=1$. Thus, denoting a Lagrange multiplier by λ

$$\begin{aligned}\frac{\partial}{\partial\pi_k}\left\{Q(\underline{\pi}) \;-\; \lambda\left(\sum_{k'=1}^{K}\pi_{k'}-1\right)\right\} &= 0\\ \Rightarrow \quad \frac{\sum_{i=1}^{M}\hat{z}_{ik}}{\pi_k} \;-\; \lambda &= 0\end{aligned}$$

Since $\sum_{k=1}^{K}\pi_k=\frac{\sum_{k=1}^{K}\sum_{i=1}^{M}\hat{z}_{ik}}{\lambda}=1$ then $\frac{M}{\lambda}=1$ and hence

$$\pi_k^{(l+1)}=\frac{\sum_{i=1}^{M}\hat{z}_{ik}}{M}$$

for $k=1,\ldots,K$.

Optimization in the M step with respect to $\underline{p}_1,\underline{p}_2,\ldots,\underline{p}_K$ is more problematic; this optimization is discussed in Chapter 4.3.1.

4.3.1 The MM Algorithm

The M step of the EM algorithm aims to maximize

$$Q(\mathbf{p}) = \sum_{i=1}^{M}\sum_{k=1}^{K} \hat{z}_{ik} \left\{ \log \pi_k + \sum_{t=1}^{n_i} \log p_{kc(i,t)} - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)} \right\} \tag{4.3}$$

with respect to the support parameters p_{kj} (for $k = 1, \ldots, K$ and $j = 1, \ldots, N$), where $\mathbf{p} = (\underline{p}_1, \underline{p}_2, \ldots, \underline{p}_K)$. The term $-\sum_{t=1}^{n_i} \log \left\{ \sum_{s=t}^{N} p_{kc(i,s)} \right\}$ makes maximization of (4.3) in the usual manner difficult. However, Lange et al. (2000) provide a summary of a method called optimization transfer using surrogate objective functions which they later term the MM algorithm. The MM algorithm is a prescription for constructing optimization algorithms more so than a directly implementable algorithm.

In order to maximize an objective function the MM algorithm forms a surrogate function that minorizes the objective function.

Definition 6 *(See Figure 4.1.) A function $g(\theta|\theta^n)$ is said to* ***minorize*** *the function $f(\theta)$ at θ^n if:*

$$(i)\ f(\theta^n) = g(\theta^n|\theta^n) \qquad \textit{and} \qquad (ii)\ f(\theta) \geq g(\theta|\theta^n) \textit{ for all } \theta.$$

A function $g(\theta|\theta^n)$ is said to ***majorize*** *the function $f(\theta)$ at θ^n if $-g(\theta|\theta^n)$ minorizes $-f(\theta)$.* ▮

The idea behind the MM algorithm is that by iteratively optimizing a suitable surrogate function the objective function is driven uphill or downhill as is required. Maximizing a minorizing surrogate function produces a new parameter estimate θ^{n+1}, the sequence of which converges to a local maximum of the objective function. Thus in a maximization problem the initials MM stand for minorize/maximize and in a minimization problem MM stands for majorize/minimize. It emerges that the EM algorithm is in fact a special case of the MM algorithm: during the E step a surrogate function is formed by imputing the expected value of the missing data and then at the M step this surrogate function (or complete data log likelihood) is maximized. The relationship between the EM and MM algorithms is discussed in Lange et al. (2000) and Hunter and Lange (2004).

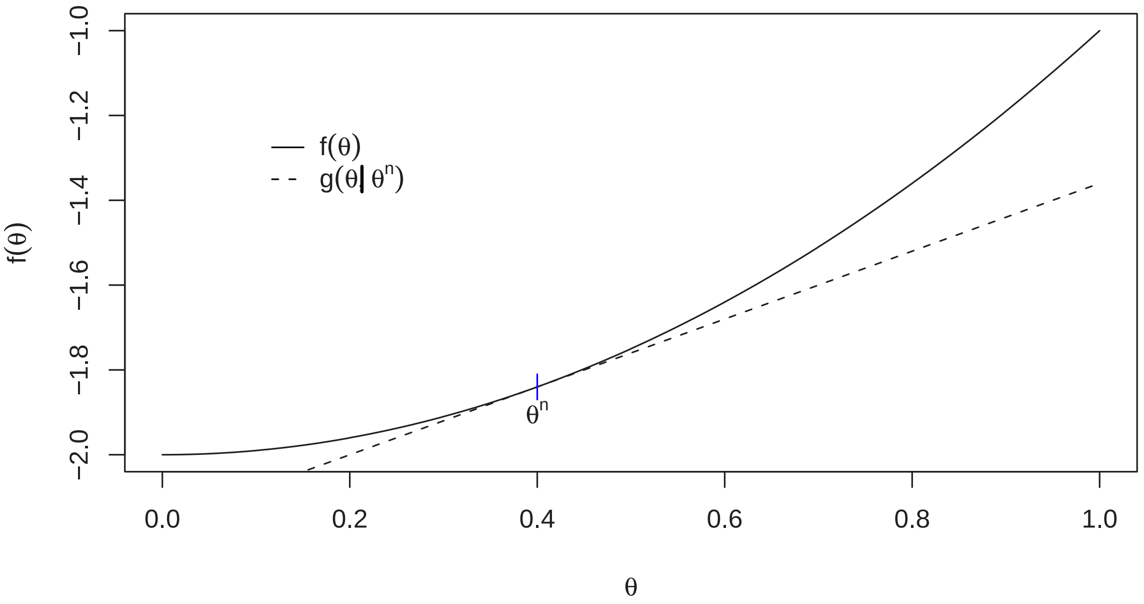


Fig. 4.1: An example of a linear minorizing surrogate function $g(\theta|\theta^n)$ which minorizes the objective function $f(\theta)$ at the parameter value θ^n.

The stability of the maximization MM algorithm relies on the ascent property

$$\begin{aligned} f(\theta^{n+1}) &= g(\theta^{n+1}|\theta^n) + f(\theta^{n+1}) - g(\theta^{n+1}|\theta^n) \\ &\geq g(\theta^n|\theta^n) + f(\theta^n) - g(\theta^n|\theta^n) \\ &= f(\theta^n). \end{aligned}$$

The MM algorithm is linearly convergent, the rate of which depends on how well the surrogate function approximates the objective function. If an objective function is strictly convex or concave, then the MM algorithm will converge to the unique optimal point, assuming it exists. If strict convexity or concavity does not hold then the MM algorithm will converge to a stationary point. Multiple random starting values for the algorithm are implemented to help avoid convergence to such a local optimum.

To construct surrogate functions mathematical properties of the function itself or of terms within the function are exploited. One such property is the supporting hyperplane property of a convex function.

Definition 7 *Suppose $f(\theta)$ is convex with differential $f'(\theta)$. Then the **supporting hyperplane property** of $f(\theta)$ states:*

$$f(\theta) \quad \geq \quad f(\theta^n) + f'(\theta^n)(\theta - \theta^n). \quad \blacksquare \tag{4.4}$$

The right hand side of inequality (4.4) provides a linear minorizing function which can be utilized as a surrogate function in an optimization transfer algorithm. Sometimes it is preferable to form a quadratic or higher order surrogate function. Expanding $f(\theta)$ using higher order expansions can yield such higher order functions. This expansion is demonstrated in Chapter 5.

4.3.2 Estimation of Support Parameters

A surrogate function which minorizes the expected complete data log likelihood (4.3) is required — maximization in the M step of the EM algorithm can then be transferred to maximization of the minorizing surrogate function. Iterative maximization provides a sequence of parameter estimates with increasing values of (4.3). This derivation is closely related to calculations given in Hunter (2004). The general reviews of the MM algorithm given by Lange et al. (2000) and Hunter and Lange (2004) are also of interest.

By (4.4), the strict convexity of the $-\log(\theta)$ function implies that

$$-\log(\theta) \quad \geq \quad -\log(\theta^n) + 1 - \frac{\theta}{\theta^n}.$$

Let $\theta = \sum_{s=t}^{N} p_{kc(i,s)}$. Thus,

$$-\log \sum_{s=t}^{N} p_{kc(i,s)} \quad \geq \quad -\log \sum_{s=t}^{N} p^{(l)}_{kc(i,s)} + 1 - \frac{\sum_{s=t}^{N} p_{kc(i,s)}}{\sum_{s=t}^{N} p^{(l)}_{kc(i,s)}}$$

where $p^{(l)}_{kc(i,s)}$ denotes the constant estimate of $p_{kc(i,s)}$ from the lth iteration of the algorithm. It follows that, up to a constant,

$$Q(\mathbf{p}) \quad \geq \quad q(\mathbf{p}) \quad = \quad \sum_{k=1}^{K} \sum_{i=1}^{M} \hat{z}_{ik} \left\{ \log \pi_k + \sum_{t=1}^{n_i} \left(\log p_{kc(i,t)} - \frac{\sum_{s=t}^{N} p_{kc(i,s)}}{\sum_{s=t}^{N} p^{(l)}_{kc(i,s)}} \right) \right\}.$$

Optimizing the surrogate function $q(\mathbf{p})$ yields new parameter values $\mathbf{p}^{(l+1)}$ which give a higher value for $Q(\mathbf{p})$; that is $Q(\mathbf{p}^{(l+1)}) \geq Q(\mathbf{p}^{(l)})$. The values converge to a maximum of Q with respect to p_{kj}.

Differentiation of q with respect to p_{kj} gives

$$\frac{\partial q}{\partial p_{kj}} \quad = \quad \sum_{i=1}^{M} \sum_{t=1}^{n_i} \hat{z}_{ik} \left(\frac{\mathbf{1}_{\{j=c(i,t)\}}}{p_{kj}} - \frac{\mathbf{1}_{[j \in \{c(i,t),\ldots,c(i,N)\}]}}{\sum_{s=t}^{N} p^{(l)}_{kc(i,s)}} \right) \qquad (4.5)$$

where $\mathbf{1}_{\{j=c(i,t)\}}$ is an indicator function such that

$$\mathbf{1}_{\{j=c(i,t)\}} = \begin{cases} 1 & \text{if } j = c(i,t) \\ 0 & \text{otherwise.} \end{cases}$$

Denoting

$$\delta_{ijs} = \begin{cases} 1 & \text{if } j = c(i,s) \text{ and } 1 \leq s \leq n_i \\ 1 & \text{if } j \neq c(i,l) \text{ for } 1 \leq l \leq n_i \text{ and } s = n+1 \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

for $i = 1, \ldots, M$, $j = 1, \ldots, N$ and $s = 1, \ldots, (n+1)$ and denoting

$$\omega_{kj} = \sum_{i=1}^{M} \sum_{t=1}^{n_i} \hat{z}_{ik} \mathbf{1}_{\{j=c(i,t)\}}$$

for $k = 1, \ldots, K$ and $j = 1, \ldots, N$ and by equating (4.5) to zero it follows that

$$\frac{\omega_{kj}}{p_{kj}} = \sum_{i=1}^{M} \sum_{t=1}^{n_i} \hat{z}_{ik} \left[\sum_{s=t}^{N} p_{kc(i,s)}^{(l)} \right]^{-1} \left[\sum_{s=t}^{(n+1)} \delta_{ijs} \right]$$

which implies that

$$p_{kj}^{(l+1)} = \frac{\omega_{kj}}{\sum_{i=1}^{M} \sum_{t=1}^{n_i} \hat{z}_{ik} \left[\sum_{s=t}^{N} p_{kc(i,s)}^{(l)} \right]^{-1} \left[\sum_{s=t}^{(n+1)} \delta_{ijs} \right]}$$

for $k = 1, \ldots, K$ and $j = 1, \ldots, N$.

By inserting this step in place of the M step of the EM algorithm maximum likelihood estimates of the Plackett-Luce support parameters can be obtained.

Many other techniques exist which optimize complex functions. The Newton-Rhapson method (Press et al., 1996) for example is often used. Each stage of the Newton-Rhapson technique involves a matrix inversion calculation which is generally computationally expensive. Since such an inversion step is avoided here and due to the manner in which the MM algorithm neatly fits into the structure of the EM algorithm, the use of the MM algorithm as an optimization technique was deemed to be justified.

4.3.3 The EM/MM Algorithm

In summary, to estimate the parameters of a mixture of Plackett-Luce models the steps of the EM algorithm with the MM algorithm embedded at the M step stage proceed as follows:

0. **Initialize:** Choose random starting values for $\underline{\pi}^{(0)}$ and $\mathbf{p}^{(0)}$. Let $l = 0$.

1. **E step:** Compute the values

$$\hat{z}_{ik} = \frac{\pi_k^{(l)} \prod_{t=1}^{n_i} \frac{p_{kc(i,t)}^{(l)}}{\sum_{s=t}^{N} p_{kc(i,s)}^{(l)}}}{\sum_{k'=1}^{K} \pi_{k'}^{(l)} \prod_{t=1}^{n_i} \frac{p_{k'c(i,t)}^{(l)}}{\sum_{s=t}^{N} p_{k'c(i,s)}^{(l)}}}.$$

2. **M step:** Calculate

$$\pi_k^{(l+1)} = \frac{\sum_{i=1}^{M} \hat{z}_{ik}}{M}$$

for $k = 1, \ldots, K$ and

$$p_{kj}^{(l+1)} = \frac{\omega_{kj}}{\sum_{i=1}^{M} \sum_{t=1}^{n_i} \hat{z}_{ik} \left[\sum_{s=t}^{N} p_{kc(i,s)}^{(l)} \right]^{-1} \left[\sum_{s=t}^{(n+1)} \delta_{ijs} \right]}$$

for $k = 1, \ldots, K$ and $j = 1, \ldots, N$. Increment l by 1.

3. **Convergence:** Repeat the E step and M step until convergence (as deemed by Aitken's acceleration criterion). The final parameter values are the maximum likelihood estimates $\hat{\underline{\pi}}$ and $\hat{\mathbf{p}}$.

Repeating this algorithm with multiple random starting values helps ensure the estimates obtained are global estimates.

4.4 Provision of Parameter Standard Errors

Early criticisms of the EM algorithm focussed on the algorithm's non-automatic production of standard errors (or of an approximate covariance matrix of the maximum likelihood estimates). However McLachlan and Krishnan (1997) and McLachlan and Peel (2000) detail methodology which provides approximate standard errors of parameter estimates derived during an EM algorithm. Moreover, in the case of independent data (which has been assumed in the case of the CAO applicants) standard errors can be produced without additional work beyond the necessary EM algorithm calculations.

In general, it is common practice to estimate the covariance matrix of the parameter estimates θ by the inverse of the expected information matrix $\mathcal{I}(\theta) = -\mathbf{E}\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta^T}\right]$. For independent data

$$\log L(\theta) \;=\; \sum_{i=1}^{M} \log L_i(\theta)$$

where in this context $L_i(\theta) = f(\underline{x}_i|\theta)$ is the Plackett-Luce density for applicant i. Denoting the score function of all applicants by $S(\mathbf{x}|\theta) = \frac{\partial \log L(\theta)}{\partial\theta}$ it follows that

$$S(\mathbf{x}|\theta) \;=\; \sum_{i=1}^{M} s(\underline{x}_i|\theta)$$

where $s(\underline{x}_i|\theta) = \frac{\partial \log L_i(\theta)}{\partial\theta}$. The expected information matrix is then

$$\begin{aligned}\mathcal{I}(\theta) &= -\mathbf{E}\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta)\right]\\ &= M\mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\log L_i(\theta)\right)^2\right] \qquad \text{(see Casella and Berger (1990,Section 7.3))}\\ &= Mi(\theta)\end{aligned}$$

where $i(\theta) = \mathbf{E}_\theta[s(\underline{x}_i|\theta)s^T(\underline{x}_i|\theta)] = \text{cov}_\theta[s(\underline{x}_i|\theta)]$ is the information contained in a single observation. Evaluating $i(\theta)$ empirically gives

$$\begin{aligned}\bar{i}(\theta) &= \frac{1}{M}\sum_{i=1}^{M} s(\underline{x}_i|\theta)s^T(\underline{x}_i|\theta) - \bar{s}\bar{s}^T\\ &= \frac{1}{M}\sum_{i=1}^{M} s(\underline{x}_i|\theta)s^T(\underline{x}_i|\theta) - \frac{1}{M^2}S(\mathbf{x}|\theta)S^T(\mathbf{x}|\theta)\end{aligned}$$

where $\bar{s} = \frac{1}{M}\sum_{i=1}^{M} s(\underline{x}_i|\theta)$. This leads to the empirical observed information matrix (Meilijson, 1989) $\mathcal{I}_e(\theta)$ where

$$\begin{aligned}\mathcal{I}_e(\theta) &= M\bar{i}(\theta)\\ &= \sum_{i=1}^{M} s(\underline{x}_i|\theta)s^T(\underline{x}_i|\theta) - \frac{1}{M}S(\mathbf{x}|\theta)S^T(\mathbf{x}|\theta)\end{aligned}$$

which can be used to approximate the observed information matrix. When $\theta = \hat{\theta}$ then $S(\mathbf{x}|\theta) = 0$ and

$$\mathcal{I}_e \;=\; \sum_{i=1}^{M} s(\underline{x}_i|\theta)s^T(\underline{x}_i|\theta).$$

Thus the covariance matrix of $\hat{\theta}$ can be approximated by the inverse of the empirical observed information matrix which in turn can be expressed in terms of the score functions of the complete data log likelihood. Computation of second order partial derivatives is therefore avoided.

4.4.1 Computation of the Empirical Information Matrix for a Mixture of Plackett-Luce Models

From the expected complete data log likelihood:

$$Q = \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \{ \log \pi_k + \sum_{t=1}^{n_i} \log p_{kc(i,t)} - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)} \}$$

the score functions with respect to each model parameter can be derived. The score function necessary to estimate the variance of the mixing proportions π_k for $k = 1, \dots, K$ is

$$\begin{aligned} s(\underline{x}_i | \pi_k) &= \frac{\partial Q}{\partial \pi_k} \\ &= \frac{\hat{z}_{ik}}{\pi_k}. \end{aligned}$$

Similarly, to estimate the variance of the support parameters p_{kj} for $k = 1, \dots, K$ and $j = 1, \dots, N$ the score function is

$$\begin{aligned} s(\underline{x}_i | p_{kj}) &= \frac{\partial Q}{\partial p_{kj}} \\ &= \hat{z}_{ik} \left[\sum_{t=1}^{n_i} \left\{ \frac{\mathbf{1}_{\{j=c(i,t)\}}}{p_{kj}} - \frac{\mathbf{1}_{[j=\{c(i,t),\dots,c(i,N)\}]}}{\sum_{s=t}^{N} p_{kc(i,s)}} \right\} \right] \\ &= \hat{z}_{ik} \left[\sum_{t=1}^{n_i} \left\{ \frac{\mathbf{1}_{\{j=c(i,t)\}}}{p_{kj}} - \frac{\sum_{s=t}^{(n+1)} \delta_{ijs}}{\sum_{s=t}^{N} p_{kc(i,s)}} \right\} \right] \end{aligned}$$

where δ_{ijs} is defined by (4.6).

Thus by formulating a matrix S which contains the score function for each parameter evaluated for each applicant, setting the empirical observation matrix $I_e(\hat{\theta}) = S^T S$ and taking the square root of the diagonal of $I_e(\hat{\theta})^{-1}$ the approximate standard errors of the membership proportions and the support parameters emerge.

4.4.2 Multiparameter Standard Errors

The posterior component membership probabilities $\pi_k p_{kj}$ for $k = 1, \ldots, K$ and $j = 1, \ldots, N$ (see Chapter 4.7.4) are of interest. The multiparameter delta method (see Wasserman, 2004,Chapter 5.5) may be used to infer the approximate distribution of such a new multiparameter.

Theorem 1 ***The multiparameter delta method.*** *Let $\underline{\theta} = (\theta_1, \ldots, \theta_n)$ and let $\underline{\hat{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ be the maximum likelihood estimate. Let $\phi = g(\theta_1, \ldots, \theta_n)$ be a function and denote by ∇g the gradient of g where $\nabla g(\underline{\hat{\theta}}) \neq 0$. Let $\hat{\phi} = g(\underline{\hat{\theta}})$. Then*

$$\frac{(\hat{\phi} - \phi)}{\hat{se}(\hat{\phi})} \rightsquigarrow N(0,1)$$

where

$$\hat{se}(\hat{\phi}) = \sqrt{(\hat{\nabla} g)^T \hat{J} (\hat{\nabla} g)}$$

where $\hat{J} = I^{-1}(\underline{\hat{\theta}})$ is the inverse of the Fisher information matrix. ■

Thus via the multiparameter delta method standard errors of multiparameter estimates may be inferred.

4.5 Model Extension: Inclusion of a Noise Component

The inclusion in the mixture of a 'noise' component with support parameter

$$\underline{p}_k = (p_{k1}, p_{k2}, \ldots, p_{kN}) = (1/N, 1/N, \ldots, 1/N)$$

was examined. Such a component 'soaks up' observations which have low probability of belonging to the other components and those observations who have equal preference for each course. The net result is that outlying observations have less of an effect on the overall results. This component is analogous to the Poisson noise component introduced in model-based clustering (Fraley and Raftery, 2002). Thus when choosing the appropriate number of components in the mixture model, whether or not a noise component should be included in the model was examined.

4.6 Model Comparison

Mixture models with differing numbers of components and the presence/absence of a noise component are fitted to the CAO data. A criterion for comparing the fitted models is required. Chapter 3 provides details on commonly used model selection techniques.

The Bayesian Information Criterion (BIC) is widely used to compare models (see Chapter 3.5). The BIC is defined to be

$$BIC = 2(\text{maximized likelihood}) - (\text{number of parameters}) \log(M).$$

The BIC can be viewed as a criterion which rewards model fit, but penalizes model complexity. The usual justification for the use of BIC is that, for regular problems, it provides an approximation to the Bayes factor for comparing models under certain prior assumptions (Kass and Raftery, 1995). Finite mixture models do not satisfy the regularity conditions for this approximation to be valid, but there is much in the literature to support its use in a mixture modelling context. Leroux (1992) showed that the number of components in the mixture, as estimated by the BIC, is at least as large as the true number of components, for large sample sizes. Keribin (1998, 2000) proved that the BIC is a consistent indicator, almost surely, of the number of components due to its appropriate penalizing term. In addition, the literature details many successful applications of the use of BIC as a model selection tool within the context of mixture models (see for example Fraley and Raftery (1998) and Dasgupta and Raftery (1998)).

The BIC consistently returned the most parsimonious and interpretable models and was used here as the main model selection tool.

4.7 Mixtures of Plackett-Luce Models for CAO Applicants

Mixtures of Plackett-Luce models were fitted to the CAO data with the number of components ranging from $K = 1$ to $K = 30$, within a maximum likelihood framework. Several random starting values for both the $\mathbf{p}$ and $\underline{\pi}$ parameters were employed in the EM algorithm with similar results. In addition, the option for allowing one of the components to be a noise component was also investigated. The mixture model with the highest BIC value was chosen and the resulting model was carefully examined.

The maximum likelihood estimates of the support parameters

$$\underline{\hat{p}}_k = (\hat{p}_{k1}, \hat{p}_{k2}, \ldots, \hat{p}_{kN})$$

for each component were examined and sorted into decreasing order. From these probabilities it is possible to determine which types of courses have highest probability of being selected by applicants from the kth component. By examining these probabilities, the component was given a summarizing label. Clearly, it would be expected that the most probable ranking is that in which the highest probable courses are selected.

In addition the estimated proportion of applicants belonging to each component $\hat{\pi}_k$ was recorded and examined.

When the full set of all CAO applicants is examined, the BIC values suggest that a twenty-two component mixture model should be used. The selected model had a noise component as one of the components. The mixing proportions π_k describe the percentage of the population assigned to each component. Table 4.1 gives the resulting twenty-two components in decreasing order of their mixing proportions.

Also reported in Table 4.1 are the approximate standard errors associated with the mixing proportions. The formation of the necessary $\{K + (N \times K)\} \times \{K + (N \times K)\} = 11748 \times 11748$ covariance matrix posed computational problems due to computing constraints. Thus, under the (rather large) assumption that the covariance of the Plackett-Luce parameters is zero, the empirical information matrix was formed (see Chapter 4.4.1) and the square root of the inverse of the diagonal terms used to provide approximate estimates of the standard errors.

An evaluation of Table 4.1 verifies the argument of supporters of the CAO system — the defining characteristic of the mixture components is the common discipline of the courses with high probabilities, as opposed to courses' common entry requirements. For example, the mixture model contains a component reflecting applicants who chose engineering courses, a component describing applicants who chose education courses and one for applicants who chose health science courses. There is no evidence, from the examination of the support parameter estimates, of a com-

Table 4.1: The names and proportions of the twenty-two components detected when the set of all applicants were analyzed. Approximate standard errors of the mixing proportion estimates are given in parentheses.

Component Name	**Proportion**	
Business & Marketing	0.08	(0.004)
Hospitality Management	0.08	(0.003)
Arts & Humanities	0.07	(0.003)
Biological Sciences	0.06	(0.003)
Business & Commerce	0.06	(0.003)
Communications & Media	0.06	(0.003)
Construction Studies	0.06	(0.003)
Computer Science (Ex-Dublin)	0.05	(0.002)
Social Science	0.05	(0.002)
Munster Based Courses	0.05	(0.002)
Computer Science (Dublin)	0.05	(0.002)
Engineering	0.04	(0.002)
Cork Based Courses	0.04	(0.002)
Galway & Limerick Based Courses	0.04	(0.001)
Education	0.03	(0.001)
Health Sciences	0.03	(0.001)
Art & Design	0.03	(0.001)
Law	0.03	(0.001)
Mathematical & Physical Sciences	0.03	(0.001)
Business & Languages	0.03	(0.001)
Music	0.02	(0.001)
Noise Component	0.002	(<0.001)

ponent representing applicants who appear to apply for high status (usually high points standard) courses. The resulting components suggest that CAO applicants do follow their vocational interests when applying to Irish institutions of third level education.

Interestingly, science based applicants are very distinctly partitioned. Applicants to biological sciences, engineering, mathematical sciences and health sciences are well segregated rather than constituting a single science component.

Also of note are the mixing proportion values. Ranking the components in order of mixing proportions indicates more applicants have a tendency to apply for humanities and business degrees than for more science based programs.

However, the results do require further examination and discussion; this is done in Chapters 4.7.1–4.7.5.

4.7.1 The Geographical Effect

The components reported in Table 4.1 reveal important traits within the population of applicants. Most obvious are the presence of components which highlight a geographical effect on applications.

Interestingly, five of the twenty-two components identified have a geographical basis. The Munster based courses and Cork based course components are epitomized by applicants who predominantly apply to institutions situated within the province of Munster or to institutions located in County Cork, respectively. The Galway and Limerick based component emerges from similarly motivated applicants. While possibly surprising that a geographical effect would be so well defined in such a relatively small island, readers acquainted with Irish society will be familiar with such a phenomenon. Firstly, many Irish students opt to live at home during their college studies; this differs from the situation in many other countries. Also, Irish people are very parochial and show strong affinity to their home region. People from Munster, and Cork in particular, have a very strong affinity to their region and tend to avoid travelling for their studies unless the course that they wish to study is not available in the region. Galway and Limerick are the main cities on the west coast of Ireland and a similar impetus is revealed by this component.

Also of note with regard to the geographical effect is the frequent distinction between sets of applicants who apply for degrees of similar discipline but are deemed separate based on whether or not the institutions to which they apply are located in Dublin (the capital of Ireland). Of Ireland's 3.92 million population, 1.12 million reside in County Dublin, and 2.11 million in the province of Leinster (the area around Dublin). Dublin is the center of Irish governmental, financial and business dealings. Therefore some applicants are drawn to living there, while others prefer to stay away to avoid living in a large city. This goes some way in explaining why applicants view courses of a similar type as different, based on whether the location of the institution is in Dublin or not. This effect is clear on the groups of applicants applying for computer science courses, and to a lesser extent on the applicants for business, marketing and commerce degrees.

4.7.2 The 'Points Race'

On the surface the components determined by the model-based clustering verify the arguments of the supporters of the CAO system. Detractors insist that applicants are influenced by media hype and by the perceived social standing of some courses (revealed through their high points requirements). Examination of the reported components and their associated parameters provides deeper insight into the behaviour of the CAO applicants.

Two approaches are taken to examine this phenomenon. Courses are examined according to the probability of the course being chosen within a component, that is using the $\mathbf{P}\{\text{Course } j|\text{Component } k\} = p_{kj}$ values (estimated by $\hat{p}_{kj}$). Also examined are the posterior probabilities of belonging to a component given that a particular course is chosen, that is using the $\mathbf{P}\{\text{Component } k|\text{Course } j\} \propto \pi_k p_{kj}$ values (estimated by $\hat{\pi}_k \hat{p}_{kj}$).

To demonstrate that there may actually be a points race, the results for the health sciences component are examined using the two approaches described above. The results of this deeper analysis is given in Chapters 4.7.3-4.7.4.

4.7.3 Examination of Component Parameters

Table 4.2 identifies 30 courses with the highest probability of selection (listed in decreasing order) given that an applicant belongs to the health sciences component (see Table 4.1). The support parameters for these courses and their approximate standard errors are also given. As with the standard errors reported in Table 4.1, the diagonal empirical information matrix was inverted due to computing constraints. Table 4.2 provides an illustration of how components were assigned a summarizing label – from a glance it is clear that applicants belonging to this component have high probability of choosing courses leading to a degree in the health sciences sector. Many health science degree programs have high entry requirements, due to demand, a limited supply of places and the fact that these courses attract highly achieving second level students. Medicine, pharmacy, dentistry and veterinary medicine are annually reported as degree programs with higher points requirements than other courses and the resulting careers are highly esteemed within Irish society. They also are vocationally driven careers, and thus it would be expected that applicants would have a tendency to apply for many courses within a discipline for which they feel that they have a vocation.

Within the top 30 courses in Table 4.2 four have been highlighted. Arts as offered by University College Dublin, law as offered by University College Dublin and Trinity College Dublin and engineering as offered by University College Dublin. While the probabilities of ranking these courses given that an applicant belongs to the health sciences component are small, in relative terms applicants are almost equally likely to rank medicinal chemistry, law or therapeutic radiography. While some would, perhaps correctly, argue that a career in law is also a vocation, it could also be argued that equally so are careers such as those in the education sector. The difference between law and education degrees, in Ireland at least, is their points requirements. Law would be considered a consistently high requirement degree, whereas an education degree would have lower points requirements. There is little evidence of health science applicants choosing education programs with high probability. Therefore, some weight has been added to the assertions of CAO detractors that the CAO system influences applicants to apply for courses that are prestigious (in terms of points). Another explanation is that the applicants are attracted to courses that tend to lead to high salaried professions. In any case, this implies that courses are being chosen by their status in society rather than by the discipline. How otherwise would health science applicants be as likely to choose law

Table 4.2: The thirty most probable courses to be ranked on an application form, given that an applicant belongs to the health sciences component. Clearly health science degrees dominate, but the presence of high status law degrees adds some weight to the argument that applicants are influenced by the 'prestige' of some courses' points requirements. Approximate standard errors of the estimated support parameters are given in parentheses.

INSTITUTION	COURSE	PROBABILITY	
UCD	Medicine	0.4723	(0.003)
TCD	Medicine	0.2413	(0.001)
UCG	Medicine	0.2004	(0.001)
UCC	Medicine	0.1219	(0.001)
RCSI	Medicine	0.0610	(<0.001)
UCD	Science	0.0351	(<0.001)
TCD	Science	0.0297	(<0.001)
TCD	Pharmacy	0.0280	(<0.001)
TCD	Dentistry	0.0280	(<0.001)
UCD	Physiotherapy	0.0260	(<0.001)
TCD	Physiotherapy	0.0241	(<0.001)
UCC	Dentistry	0.0233	(<0.001)
UCD	Veterinary Medicine	0.0163	(<0.001)
RCSI	Medicine with Leaving Certificate Scholarship	0.0153	(<0.001)
UCG	Science	0.0140	(<0.001)
UCC	Biological & Chemical Sciences	0.0125	(<0.001)
TCD	Human Genetics	0.0121	(<0.001)
UCG	Biomedical Science	0.0116	(<0.001)
DIT	Optometry	0.0104	(<0.001)
UCD	Radiography	0.0101	(<0.001)
TCD	Medicinal Chemistry	0.0099	(<0.001)
UCD	Arts	0.0092	(<0.001)
UCD	Law	0.0091	(<0.001)
TCD	Law	0.0085	(<0.001)
UCD	Engineering	0.0083	(<0.001)
TCD	Therapeutic Radiography	0.0081	(<0.001)
TCD	Psychology	0.0074	(<0.001)
RCSI	Physiotherapy	0.0069	(<0.001)
RCSI	Medicine with RCSI Scholarships	0.0065	(<0.001)
UCD	Psychology	0.0059	(<0.001)

as therapeutic radiography?

The four courses highlighted in Table 4.2 also have the common characteristic of being based in institutions in Dublin. It is clear from other components that the geographical location of a course influences applicants. Within the health sciences component it appears both geography and course status affect the way in which applicants rank courses. However, it is difficult to fully distinguish between these influences.

Also of note is the high probability of choosing the arts degree (in University College Dublin) and engineering (in University College Dublin). As the name partially suggests, in an arts degree students study one or two subjects from a range of arts and humanities subjects. Thus the arts degree is a very general degree that provides a broad basis from which many different career paths may emerge. In fact, it is the most frequently ranked degree program amongst all CAO applicants and has relatively achievable points requirements. Its popularity, or perhaps its reputation as a 'fail safe' third level choice, are possible explanations of its high choice probability within the health science component.

The inclusion of engineering as a high probability course could also be due to applicants including a 'fail safe' alternative. The required points for engineering in University College Dublin were much lower than the health science degrees in Table 4.2, so the points status would not appear to be a contributing factor. It is clear that health science applicants select a general science degree with high probability and perhaps are then also attracted to the general scientific aspects of an engineering degree. More of note perhaps is that in 2000 engineering as offered by University College Dublin was a general entry degree where students did not choose a specific vein of engineering until later in their degree. While both Trinity College Dublin and University College Galway ran a similar style program, the required points that year were considerably higher than those required for entry to University College Dublin's degree. Thus, similarly to the arts degree, engineering may have been viewed as the 'fail safe' science-based option.

4.7.4 Examination of Posterior Component Membership Probabilities

An alternative approach can be taken in the analysis of the parameter estimates by examining the posterior probability of belonging to component k given that course j was selected, that is, $\mathbf{P}\{\text{Component } k | \text{Course } j\}$. Table 4.3 illustrates the twenty-five courses whose selection gives highest posterior probability of belonging to the health sciences component. The reported standard errors of the posterior probability estimates are estimated in the manner described in Chapter 4.4.2. Due to computing constraints however only the diagonal empirical information matrix was formed and inverted; it follows that the reported standard errors are approximate.

Examination of the mixture model in this way further highlights the subtle effect the points race may have on some applicants' choices. Within the top twenty-five courses that suggest high probability of belonging to the health sciences component are Mathematics and Latin and Mathematics and Psychology, both offered by Trinity College Dublin. It appears strange to have high probability of belonging to a component dominated by health sciences courses due to the selection of either of these courses. Both are part of Trinity College's version of the general arts degree – the Two Subject Moderatorship (TSM) program. In the TSM program, students choose two modules from a range of arts and humanities subjects and study them simultaneously. However, each combination is viewed as a separate course by the CAO and due to the wide range of subjects, and therefore combinations, their choice is usually quite rare leading to sparse data. Strange results emerged when initially analyzing the CAO data due to the rarity of some course selections within the TSM program. However the inclusion here of only two of the wide range of TSM courses suggests a contributing factor other than data sparsity. These two TSM courses both include mathematics; in that particular year points requirements for TSM courses involving mathematics were at a similar level to many of the listed health science programs in Table 4.3. Therefore, deeper investigation of the posterior probabilities highlights again the possibility of a subtle effect that a course's points requirement may have on CAO applicants.

Why the focus on law programs and mathematics programs as examples of the points race? Other high points courses such as Actuarial and Financial Studies (in

Table 4.3: Twenty-five courses whose selection on a CAO application form gives highest probability of belonging to the health sciences component. Associated standard errors are all less than 2.4×10^{-3}.

INSTITUTION	COURSE	PROBABILITY
UCD	Medicine	0.9440
TCD	Medicine	0.9392
RCSI	Medicine	0.8886
RCSI	Medicine with Leaving Certificate Scholarship	0.8813
UCG	Medicine	0.8724
RCSI	Medicine with RCSI Scholarships	0.8010
UCC	Medicine	0.7633
TCD	Dentistry	0.5391
UCC	Dentistry	0.3674
TCD	Pharmacy	0.3110
TCD	Medicinal Chemistry	0.2725
TCD	Therapeutic Radiography	0.2662
TCD	Human Genetics	0.2579
UCD	Radiography	0.2230
TCD	Physiotherapy	0.2141
RCSI	Physiotherapy	0.2035
TCD	Mathematics/Latin	0.2013
DIT	Optometry	0.1989
UCD	Physiotherapy	0.1986
UCD	Veterinary Medicine	0.1796
TCD	Mathematics/Psychology	0.1642
UCG	Biomedical Science	0.1517
UCG	Biomedical Engineering	0.1093
TCD	Science	0.1053
TCD	Occupational therapy	0.0992

University College Dublin) also appear within the top 50 programs in both views of the model; again, this course appears to be a strange course to appear amongst a component dominated by the "vocational" health science sector.

Why the focus on the health sciences component only? It seems natural to also consider the law component that is also deemed as high points and high status. The points race effect is also apparent here – psychology in both University College Dublin and Trinity College Dublin, which had high entry requirements that year, have high probability of being selected given that an applicant belongs to the law component. While law and psychology have some similarities, they would not be deemed as members of the same discipline suggesting some element of the points race is present. However, examination of the posterior probabilities for the law component gives less of an indication of the presence of a points race. It seems the points status of courses has more of an effect in the health science component than in the other components in the mixture model.

4.7.5 The Gender Effect

The only covariate available was the gender of the CAO applicants in 2000; of the 53757 applicants, 24419 were male. The data was partitioned according to applicant gender and mixtures of Plackett-Luce models were fitted to the two resulting data sets. Examination of the support parameter estimates, $\hat{p}_{kj}$, led to the summarizing component labels as outlined in Table 4.4. Approximate standard errors of the mixing proportions are also reported. Due to computing constraints which arise when attempting to compute the full covariance matrix, the square root of the inverse of the diagonal elements of the empirical information matrix (Chapter 4.4) are used to provide an approximation of the standard errors.

The resulting mixtures fitted to the partitioned data provide good insight into the different choice behaviour of the male and female applicants. The predominant aspect of the component labels is subject discipline, thus enhancing the supporting view of the CAO that applicants are inclined to follow their vocational interests. The geographical effect discussed in Chapter 4.7.1 is again apparent, but it is more apparent in the male results. In particular, some male components reveal a common discipline but at different geographical locations; this occurs more so than in the

Table 4.4: The resulting 16 components from analysis of the female applicants, and the resulting 17 components from analysis of the male applicants. Approximate standard errors of the mixing proportions are given in parentheses.

FEMALE RESULTS			**MALE RESULTS**		
Component Label	**Proportion**		**Component Label**	**Proportion**	
Hospitality Management	0.11	(0.001)	Construction Studies	0.09	(0.002)
Social Science	0.11	(0.002)	Communications & Journalism	0.09	(0.002)
Business & Marketing (Dublin)	0.09	(0.002)	Business & Marketing (Dublin)	0.09	(0.002)
Biological Sciences	0.08	(0.001)	Computer Science (Ex-Dublin)	0.08	(0.002)
Cork Based Courses	0.08	(0.002)	Hospitality Management	0.07	(0.002)
Applied Computing (Ex-Dublin)	0.07	(0.003)	Computer Science (Dublin)	0.07	(0.002)
Communications & Journalism	0.07	(0.002)	Arts/Humanities	0.06	(0.002)
Business & Commerce (Ex-Dublin)	0.07	(0.001)	Engineering (Ex-Dublin)	0.06	(0.002)
Law & Psychology	0.06	(0.002)	Business & Commerce	0.06	(0.002)
Galway & Limerick Based Courses	0.06	(0.002)	Cork Based Courses	0.06	(0.002)
Education	0.05	(0.002)	Law & Business	0.06	(0.002)
Engineering & Computer Science	0.04	(0.001)	Engineering (Dublin)	0.05	(0.002)
Art/Design & Music	0.04	(0.002)	Sports Science & Education	0.05	(0.002)
Business & Languages	0.04	(0.002)	Science	0.04	(0.001)
Health Sciences	0.04	(0.003)	Limerick Based Courses	0.04	(0.001)
Noise Component	0.003	(0.002)	Health Sciences	0.03	(0.001)
			Noise Component	0.004	(0.001)

female components. For example, the male engineering applicants are partitioned by the location of the institution in Dublin, as are the computer science applicants.

Stereotypical differences between the two genders are very apparent in the resulting components – there appears to be distinct components for females in social science, art and design, music and education whereas female applicants with an interest in engineering and computer science are grouped together. Not only are the male engineering and computer science components separate, they are further divided within these disciplines by geography. Further, the largest component (with mixing proportion 0.09) in the male results involves construction studies courses whereas this does not appear as a distinct component in the female results.

Other results of interest are the popularity of biological sciences amongst females whereas males have a general science component in their results, both genders have education components but the male education component also has a sports aspect.

In addition, in close similarity to the results for all applicants (see Chapter 5.3.3), the male health sciences component contains three law degrees in the top thirty most probable courses. Similar results are revealed for the females, but the probability of selection of the law courses is lower within the health sciences component.

4.7.6 Clustering of Applicants

A major advantage of fitting mixture models via the EM algorithm, as detailed by Fraley and Raftery (1998), is that the value $\hat{z}_{ik}$ at convergence is an estimate of the conditional probability that observation i belongs to component k; these values can be used to cluster observations into groups. A clustering of the set of applicants is simply achieved by examining

$$\max_k \mathbf{P}\{\text{Component } k | \text{Application } i\}$$

$\forall\ i$ and assigning applicants to the group for which the maximum is achieved.

The clustering of applicants can be scrutinized in different ways. As suggested by Bensmail et al. (1997), the uncertainty associated with an applicant's component membership can be measured by $U_i = \min_{k=1,\ldots,K}(1 - \mathbf{P}\{\text{Component } k|\text{Application } i\})$. When i is very strongly associated with group k then $\mathbf{P}\{\text{Component } k|\text{Application } i\}$ will be large and so U_i will be small. Figure 4.2 illustrates the uncertainty associated with the clustering of the male and female applicants.

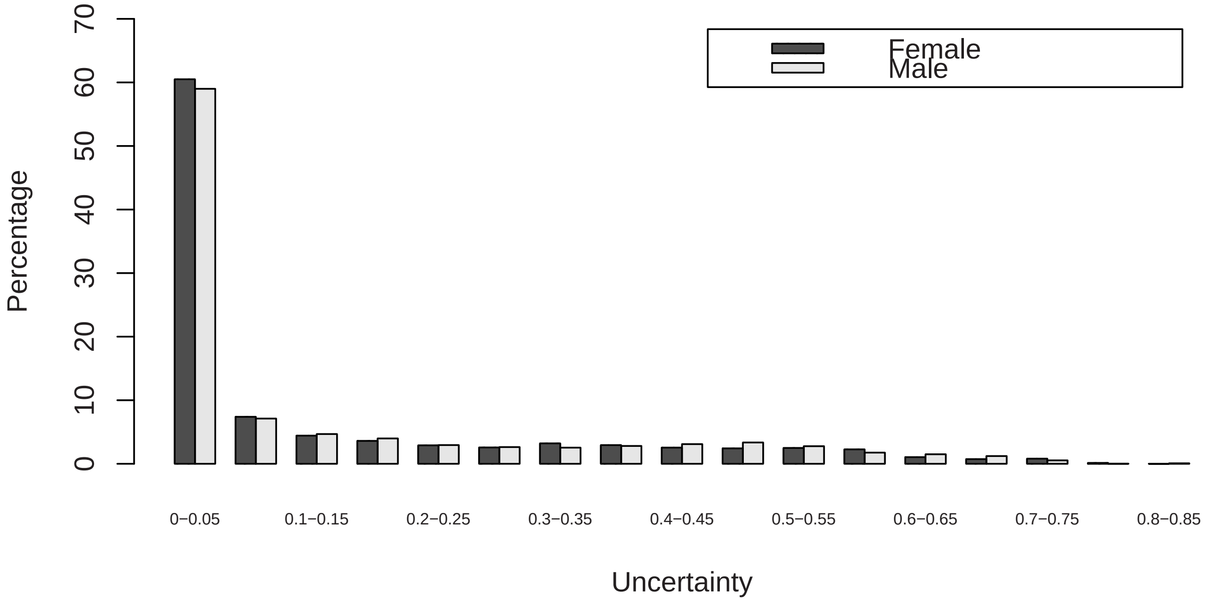


Fig. 4.2: Uncertainty in the clustering of female and male applicants.

Clearly, the clustering uncertainty values tend to be very small, with 61% of females and 59% of males classified with an uncertainty of less than 0.05. Summary statistics for the uncertainty values further demonstrate how well the model allocates applicants to components; these are given in Table 4.5.

Table 4.5: Summary statistics associated with the clustering uncertainty of male and female applicants.

	1st Quartile	**Mean**	**3rd Quartile**
FEMALE	0.0002	0.1228	0.1866
MALE	0.0002	0.1301	0.2043

4.8 Conclusions

This chapter presents a model-based statistical analysis of degree level applicants to Irish institutions of third level education. The methods seek to find groups of similar applicants, and to draw conclusions about the merits and failures of the centralized applications system from the defining characteristics of these groups.

A top level view of the groups of applicants suggested by the analysis verifies a supporting view of the CAO system — applicants appear to follow their vocational interests and rank their third level course choices in a manner which reflects this. The analysis suggests that the majority of CAO applicants use the system as it is intended and rank courses in view of their genuine preferences and/or career choice. However, it is apparent that more subtle influences also contribute to course choice and a detailed examination of the mixture components indicates the faint presence of the reported 'points race'. It appears there are those who choose courses on the points levels of previous years and therefore on the prestige attached to some of these courses.

While most discussions of the CAO system in Irish education circles focus on the influence of the 'points race' this work highlights other factors which have an influence on an applicant's course choice. The geographical location of the institution to which an applicant applies has a clear affect on the choice process. Whether this is due to a vocational desire to study a particular course in a specific institution, the desire to live in a certain area or because of financial viability, it is a striking feature of the groups of applicants. A course's geographical location appears to be almost as important as vocational interest in an applicant's choice process. Whether this feature is a benefit of the CAO system or not remains to be researched.

Further to the effects of vocation, geography and the points race, the gender of the applicant also affects course choice. Geography and the points race may have a larger effect on male applicants than on females. Stereotypical gender differences are also apparent — only 4% of female applicants are 'classified' as engineering and computer science students compared to 26% of the population of male applicants. Further differences (see Chapter 4.7.5) indicate that males and females need to be targeted in different manners, with regard to third level education, and this should be of interest to third level institutions and to governmental education departments.

In terms of the model employed within components, the Plackett-Luce model performs well when modelling the rankings of the preferred third level choices of the CAO applicants. While the model does suffer from independence from irrelevant alternatives (IIA) (see Chapter 3.1) in this application it appears to provide a realistic representation of the course choice process.

The only covariate available for this analysis is the gender of the applicant – relationships between course choice and other covariates are very likely to be present. Expanding the analysis to include other covariates would also be desirable, but further covariates were not available for this study.

Chapter 5

Mixtures of Benter Models

Mixtures of Benter models are proposed as a tool for modelling heterogeneous populations who generate rank data. Two Irish elections are used to demonstrate the applicability of such models: voters in the 1997 presidential election and voters in the Dublin West constituency in the 2002 general election are modelled. The work presented in this chapter follows work reported in Gormley and Murphy (2005). While the mixture of Plackett-Luce models fitted in Chapter 4 is a special case of a mixture of Benter models (see Chapter 3.2) the methodology developed in the previous chapter is more amenable to cases which involve large choice sets.

5.1 The Benter Model

The Plackett-Luce model for rank data suffers from the property that the probability of a candidate with a low support parameter being ranked highly is too small. Thus the Benter model (Chapter 3.2) is used to model election data. The Benter model has two parameters: the support parameter $\underline{p} = (p_1, p_2, \ldots, p_N)$ where $\sum_{j=1}^{N} p_j = 1$ and the dampening parameter

$$\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N).$$

The support parameter p_j represents the probability of candidate j being given a first preference and the dampening parameters model the way in which lower preferences may be chosen less carefully than higher preferences by a voter. Under the Benter model, the probability of voter i's ballot $\underline{x}_i$ is:

$$\mathbf{P}\{\underline{x}_i|\underline{p},\underline{\alpha}\} \;=\; \prod_{t=1}^{n_i} \frac{p_{c(i,t)}^{\alpha_t}}{\sum_{s=t}^{N} p_{c(i,s)}^{\alpha_t}}.$$

Within the context of voting it follows that $n_i \leq N$ where N is the number of candidates running in the election.

5.2 Homogeneous Benter Models

As an initial analysis of the Irish electorate a single component Benter model was fitted, within a maximum likelihood framework, to each of the eight opinion polls from the 1997 presidential election campaign and to the 2002 general election data from the Dublin West constituency (see Chapter 2). A single component Benter model where the dampening parameters were constrained such that $\underline{\alpha} = (1, \ldots, 1)$ (i.e. a Plackett-Luce model) was also fitted. The estimated model parameters and their associated standard errors are reported in Figure 5.1, Table 5.1 and Table 5.2. The approximate standard errors reported throughout this chapter are derived within the EM algorithm as proposed by McLachlan and Krishnan (1997) and McLachlan and Peel (2000). The derivation of these standard errors is outlined in Chapter 5.4.

Figure 5.1 demonstrates the support parameter associated with each presidential candidate under both the Plackett-Luce and Benter models inferred from each opinion poll. A general popularity ordering of McAleese, Banotti, Scallon, Roche and then Nally emerges under both models. The most striking feature of the plots is perhaps the rapid decline in support for Adi Roche. Roche began as favourite for the presidential seat but, after criticism from co-workers about her style of work and claims that she was an unsuitable person to be president, her campaign never recovered. McAleese and Banotti maintained first and second place across the polls while Rosemary Scallon's position improved. Early criticisms of the conservative candidate fizzled out as the campaign developed and as her professional presentation skills became more evident she finished in a respectable third place.

The model parameter estimates differ in the final polls where the Plackett-Luce estimates seem to shrink together but the Benter estimates become more dispersed. This can, in part, be explained by the fact that in the 30/10 poll people were

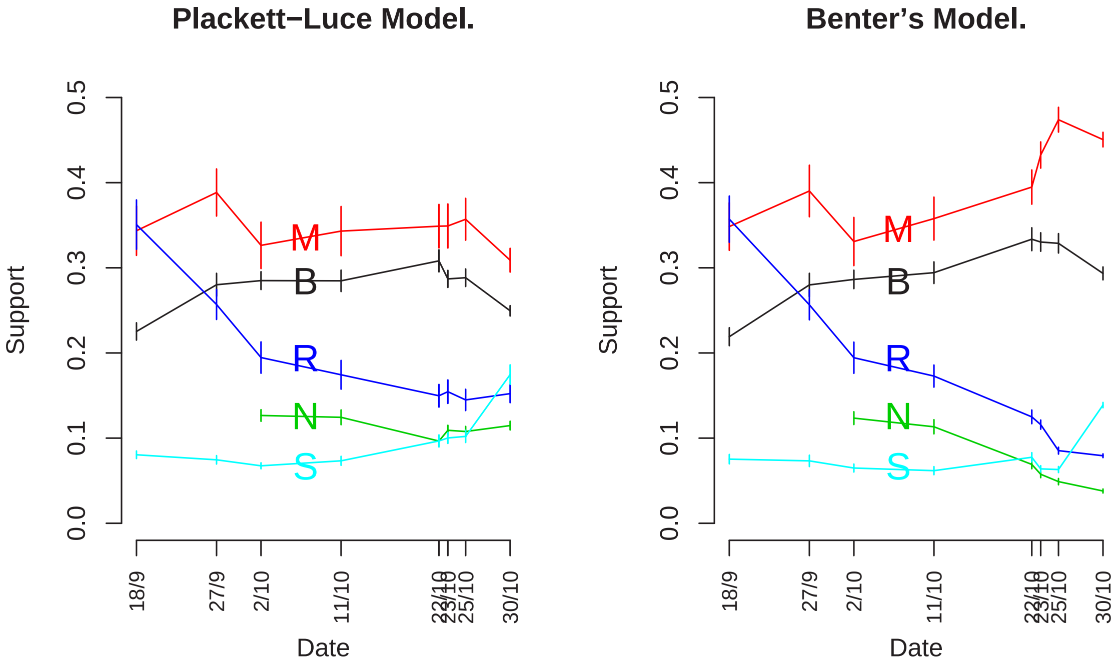


Fig. 5.1: A graphical representation of the maximum likelihood estimates of the Plackett-Luce support parameter and the Benter support parameter for each of the eight polls from the 1997 presidential election campaign. Each of the five candidates are denoted by their surname initial. Note that Nally was not a candidate when the first two polls were taken. Two standard errors either side of each estimate are also illustrated.

Table 5.1: The values of the Benter dampening parameter for each of the polls from the 1997 presidential election campaign. The fourth value of the dampening parameter is not computed for the first two polls as there were only four candidates when the poll was taken. Standard errors of the estimates are given in parentheses.

		Dampening Parameter			
Date	**Poll**	α_1	α_2	α_3	α_4
18/9	IMS	1.00	0.80 (0.07)	1.00 (0.08)	—
27/9	MRBI	1.00	0.94 (0.00)	1.00 (0.09)	—
2/10	IMS	1.00	1.00 (0.08)	1.00 (0.08)	1.00 (0.09)
11/10	MRBI	1.00	1.00 (0.08)	1.00 (0.08)	1.00 (0.10)
22/10	MRBI	1.00	1.00 (0.05)	0.95 (0.07)	1.00 (0.10)
23/10	IMS	1.00	0.98 (0.04)	0.80 (0.05)	0.99 (0.07)
25/10	IMS	1.00	0.92 (0.04)	0.73 (0.05)	0.63 (0.07)
30/10	Lansdowne	1.00	0.73 (0.03)	0.18 (0.03)	0.00 (0.04)

Table 5.2: Maximum likelihood estimates of the Plackett-Luce and Benter support parameters for the 2002 Dublin West constituency election data; the proportion of the total first preference votes for each candidate is included for comparison purposes.

Candidate	**Party**	**First Preference**	**Plackett-Luce Estimate**	**Benter Estimate**
Bonnie, R.	GP	0.03	0.07	0.06
Burton, J.	Lab	0.13	0.16	0.16
Doherty-Ryan, D.	FF	0.08	0.11	0.11
Higgins, J.	SP	0.22	0.16	0.17
Lenihan, B.	FF	0.27	0.18	0.20
McDonald, M.	SF	0.08	0.06	0.06
Morrissey, T.	PD	0.08	0.12	0.11
Smyth, J.	CSP	0.00	0.02	0.01
Terry, S.	FG	0.12	0.12	0.12

encouraged to express all preferences. Interestingly, the support parameter estimates under the Benter model in the 30/10 poll (the exit poll) are very similar to the first preference proportions for each candidate. The dampening parameters for this poll (Table 5.1) complement these estimates as it is clear that lower place preferences are strongly dampened in this poll, thus giving a lot of priority to higher preferences.

The dampening parameters associated with the exit poll data give a good demonstration of the value of estimating such parameters — here they provide an illustration of how many preference levels have an effect on estimating the support parameters of the model. The third level dampening parameter of 0.18 suggests that the third place preferences are only made with around one fifth of the certainty that the first place preferences are. Also $\alpha_4 = 0$ suggests that voters select the remaining candidates with equal probability.

Similar types of effects are apparent when examining the estimated support parameters for the voting data in the Dublin West constituency (Table 5.2). Again the Plackett-Luce parameters seem to shrink towards the mean — lower support parameters are pulled up (e.g. Bonnie from 0.03 to 0.07) and larger support parameters are pulled down (e.g. Higgins from 0.22 to 0.16). The shrinkage of the Benter estimates towards the first preference proportions is less extreme but again the dampening parameter values go some way in explaining this. The largest standard error of the support parameters under either model was 2×10^{-6}.

The Benter dampening parameter estimate for the Dublin West data was

$$\underline{\hat{\alpha}} = (1.00, 0.92, 0.66, 0.44, 0.89, 0.94, 0.94, 0.00, 0.00)$$

with the associated standard errors less than 1×10^{-4}. These dampening parameters suggest that lower preferences should be taken into account when modelling such data: $\alpha_6 = \alpha_7 = 0.94$ shows that the sixth and seventh level preferences are almost as influential as first place preferences.

One exception to the pattern of shrinkage of the support parameters is McDonald — she got 8% of the first preference votes and estimated support parameters of 0.06 under both the Plackett-Luce and Benter models. McDonald was a Sinn Féin candidate in the Dublin West election. Sinn Féin have a well defined body of support in Ireland in that they are “the only party committed to achieving a democratic socialist republic and the end of British rule in Ireland”. Voters would tend to rank

Sinn Féin candidates first or else not at all thus explaining why McDonald's support parameters are close to the first preference proportions. This in turn suggests there is a group of such voters present in the Irish electorate among potentially many others.

While the above analyses provide evidence to suggest the types of ranking models introduced are both applicable and necessary, the grouping structure within the electorate is not exposed. The next section details the exploration of the Irish electorate using mixtures of Benter models. This approach provides an easily interpretable model for the heterogeneity in the electorate.

5.3 Mixtures of Benter Models

Similar to the mixtures of Plackett-Luce models framework, suppose that the population of voters consists of K sub-populations where voters belong to sub-population k with probability π_k. Given that a voter is in sub-population k their vote follows an $f(\underline{x}_i|\underline{\theta}_k)$ density. Then the probability of each vote is

$$\mathbf{P}\{\underline{x}_i\} = \sum_{k=1}^{K} \pi_k f(\underline{x}_i|\underline{\theta}_k)$$

which is a finite mixture model. It is assumed that $\{f(\underline{x}_i|\underline{\theta}_k) : \underline{\theta}_k \in \Theta\}$ is a parametric family of Benter model densities where $\underline{\theta}_k = (\underline{p}_k, \underline{\alpha})$. Thus $p_{kc(i,t)}$ now denotes the probability of the candidate chosen in tth position by voter i being ranked first, given that voter i is a member of sub-population k.

5.3.1 Mixture Constraints

The proposed mixture models allow the parameters in the different components to be constrained in different ways; this offers modelling flexibility.

The Plackett-Luce model is a special case of the Benter model with dampening parameter $\underline{\alpha} = \underline{1} = (1, 1, \ldots, 1)$. Therefore, as in the homogeneous Benter model in Chapter 5.2, the option of constraining the $\underline{\alpha}$ value to be identically $\underline{1}$ or leaving it unconstrained is investigated.

The option of forcing one component in the mixture to be a uniform component is also examined; that is a component with $\underline{p}_k = (1/N, 1/N, \ldots, 1/N)$. This uniform component can “soak up” any outlying data values and allows for better modelling of the remaining data. This use of a noise component is analogous to that used in Chapter 4.

Four different types of model were fitted to the data:

1. a mixture of Plackett-Luce models,
2. a mixture of Plackett-Luce models constrained such that one component is fixed to be a noise component,
3. a mixture of Benter models and
4. a mixture of Benter models constrained such that one component is fixed to be noise component.

In both of the Plackett-Luce mixture models $\underline{\alpha}$ is, by definition, constrained to be $\underline{1}$ whereas in both the Benter mixture models $\underline{\alpha}$ is to be estimated. Hence $f(\mathbf{x}|\mathbf{p},\underline{\alpha})$ is used as notation for the Benter model and $f(\mathbf{x}|\mathbf{p},\underline{1})$ for the Plackett-Luce model.

5.3.2 Fitting Mixtures of Benter Models

The mixture models were fitted using maximum likelihood methods; that is, the likelihood

$$L(\underline{\pi},\mathbf{p},\underline{\alpha}|\mathbf{x}) \;=\; f(\mathbf{x}|\underline{\pi},\mathbf{p},\underline{\alpha}) = \prod_{i=1}^{M}\left[\sum_{k=1}^{K}\pi_k f(\underline{x}_i|\underline{p}_k,\underline{\alpha})\right],$$

is maximized with respect to $\underline{\pi} = (\pi_1,\pi_2,\cdots,\pi_K)$, $\mathbf{p} = (\underline{p}_1,\underline{p}_2,\cdots,\underline{p}_K)$ and $\underline{\alpha} = (\alpha_1,\alpha_2,\cdots,\alpha_N)$, in the case where $\underline{\alpha}$ is not constrained to be $\underline{1}$.

The fitting of mixtures of Benter models using maximum likelihood (similar to mixtures of Plackett-Luce models in Chapter 4) can be implemented using the EM algorithm (Dempster et al., 1977). To use the EM algorithm, a membership label is introduced for each voter such that $z_{ik} = 1$ if voter i belongs to component k and $z_{ik} = 0$ otherwise. The likelihood of the observed data and the unobserved labels is called the complete data likelihood,

$$L_C(\underline{\pi},\mathbf{p},\underline{\alpha}|\mathbf{x},\mathbf{z}) \;=\; \prod_{i=1}^{M}\prod_{k=1}^{K}\left[\pi_k\left(\prod_{t=1}^{n_i}\frac{p_{kc(i,t)}^{\alpha_t}}{\sum_{s=t}^{N}p_{kc(i,s)}^{\alpha_t}}\right)\right]^{z_{ik}}.$$

The EM algorithm involves an E step that replaces the missing data $\mathbf{z}$ with its expected value given the current parameter estimates and an M step that maximizes the complete data log likelihood computed with the estimates of $\mathbf{z}$. Specifically, the EM algorithm proceeds as follows:

0. **Initialize:** Let $l = 0$ and choose initial parameter estimates $\underline{\pi}^{(0)}$, $\mathbf{p}^{(0)}$ and $\underline{\alpha}^{(0)}$.

1. **E Step:** Compute the quantities

$$z_{ik}^{(l+1)} = \frac{\pi_k^{(l)} f(\underline{x}_i | \underline{p}_k^{(l)}, \underline{\alpha}^{(l)})}{\sum_{k'=1}^{K} \pi_{k'}^{(l)} f(\underline{x}_i | \underline{p}_{k'}^{(l)}, \underline{\alpha}^{(l)})}.$$

   for $i = 1, \ldots, M$ and $k = 1, \ldots, K$.

2. **M step:** Maximize the expected complete data log likelihood:

$$Q \quad = \quad \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \left\{ \log \pi_k + \sum_{t=1}^{n_i} \alpha_t \log p_{kc(i,t)} - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t} \right\} \quad (5.1)$$

   with respect to $\underline{\pi} = (\pi_1, \ldots, \pi_K)$, $\mathbf{p} = (\underline{p}_1, \ldots, \underline{p}_K)$ and $\underline{\alpha} = (\alpha_1, \ldots, \alpha_N)$ (if required). Call the maximizing values $\underline{\pi}^{(l+1)}$, $\mathbf{p}^{(l+1)}$ and $\underline{\alpha}^{(l+1)}$.

3. If converged, then stop. Otherwise, increment l and return to Step 1.

Convergence of the EM algorithm was assessed using the Aitken acceleration estimate of the final maximized likelihood (see Chapter 4.3). The algorithm is considered to be converged when the current likelihood value is within a tolerance of the Aitken estimate (Böhning et al., 1994; Lindsay, 1995; McLachlan and Peel, 2000).

The ECM algorithm (Meng and Rubin, 1993) proved to be useful when fitting mixtures of Benter models. This algorithm replaces maximization in the M step with a series of easier conditional maximization steps. In this case, the conditional maximizations are with respect to $\pi_1, \ldots, \pi_k$, $\underline{p}_1, \ldots, \underline{p}_K$ and $\alpha_2, \ldots, \alpha_{N-1}$. Maximizing (5.1) with respect to π_k as detailed in Chapter 4.3 provides the iterative estimate

$$\pi_k^{(l+1)} \quad = \quad \frac{\sum_{i=1}^{M} \hat{z}_{ik}^{(l+1)}}{M}$$

for $k = 1, \ldots, K$.

Estimating $p_{kj}^{(l+1)}$ and $\alpha_t^{(l+1)}$ is difficult in practice. Thus the conditional maximizations with respect to the support and dampening parameters in the M step are implemented using the MM algorithm (Lange et al., 2000; Hunter and Lange, 2004). This algorithm works by first constructing a function that minorizes the objective function and then maximizing the minorizing function. This process is iterated leading to a sequence of parameter estimates giving increasing objective function values (see Chapter 4.3.1).

5.3.3 Maximization with Respect to Support Parameters.

Consider the complete data log likelihood which is to be maximized:

$$Q = \sum_{i=1}^{M}\sum_{k=1}^{K} \hat{z}_{ik}\{\log \pi_k + \sum_{t=1}^{n_i} \alpha_t \log p_{kc(i,t)} - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}\}. \tag{5.2}$$

The term $-\log \sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}$ is problematic when trying to optimize this expression. In this case α_t is treated as a fixed constant $\bar{\alpha}_t$. Let $\bar{\alpha}_t$ be the value of α_t at the previous iteration of the MM algorithm.

By the supporting hyperplane property of a convex function (see (4.4)), the strict convexity of the $-\log(\theta)$ function implies that

$$-\log(\theta) \quad \geq \quad -\log(\theta^n) + 1 - \frac{\theta}{\theta^n}$$

for some value θ^n. Thus,

$$-\log\sum_{s=t}^{N} p_{kc(i,s)}^{\bar{\alpha}_t} \quad \geq \quad -\log\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \ + 1 \ - \frac{\sum_{s=t}^{N} p_{kc(i,s)}^{\bar{\alpha}_t}}{\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t}}.$$

where $\bar{p}_{kj}$ is a constant and in practice is the estimate of p_{kj} from the previous iteration of the MM algorithm. It follows from (5.2) that, up to a constant,

$$Q(p_{kj}) \quad \geq \quad q_1(p_{kj}) \quad = \quad \sum_{i=1}^{M}\sum_{k=1}^{K}\sum_{t=1}^{n_i} \hat{z}_{ik} \left\{ \bar{\alpha}_t \log p_{kc(i,t)} - \left(\frac{\sum_{s=t}^{N} p_{kc(i,s)}^{\bar{\alpha}_t}}{\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t}} \right) \right\}.$$

Again maximizing the function q_1 with respect to p_{kj} poses maximization problems. By the supporting hyperplane property of convex functions (4.4) the convex function $f(p) = -p^{\bar{\alpha}}$ becomes

$$-p^{\bar{\alpha}} \ \geq \ -\bar{p}^{\bar{\alpha}} \ - \ \bar{\alpha}\, \bar{p}^{\bar{\alpha}-1}\,(p - \bar{p})$$

which provides the surrogate function

$$q_1(p_{kj}) \geq q_2(p_{kj}) = \sum_{i=1}^{M}\sum_{k=1}^{K}\sum_{t=1}^{n_i} \hat{z}_{ik} \left[\bar{\alpha}_t \log p_{kc(i,t)} \quad - \quad \left\{ \sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \right\}^{-1} \left\{ \sum_{s=t}^{N} \bar{\alpha}_t \bar{p}_{kc(i,s)}^{\bar{\alpha}_t - 1} p_{kc(i,s)} \right\} \right]$$

up to a constant. By iterative maximization of the surrogate function q_2 we produce a sequence of p_{kj} values which converge to the maximum of Q with respect to p_{kj}. Thus differentiation of $q_2(p_{kj})$ with respect to p_{kj} gives

$$\frac{\partial q_2}{\partial p_{kj}} = \sum_{i=1}^{M}\sum_{t=1}^{n_i} \hat{z}_{ik} \left\{ \frac{\bar{\alpha}_t}{p_{kc(i,t)}} \mathbf{1}_{\{j=c(i,t)\}} - \left(\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \right)^{-1} \left(\sum_{s=t}^{N} \bar{\alpha}_t \bar{p}_{kc(i,s)}^{\bar{\alpha}_t - 1} \mathbf{1}_{\{j=c(i,s)\}} \right) \right\} \quad (5.3)$$

where $\mathbf{1}_{\{j=c(i,s)\}}$ is an indicator function such that

$$\mathbf{1}_{\{j=c(i,s)\}} \quad = \quad \begin{cases} 1 & \text{if } j = c(i,s) \\ 0 & \text{otherwise.} \end{cases}$$

Denoting

$$\omega_{kj} \quad = \quad \sum_{i=1}^{M}\sum_{t=1}^{n_i} \hat{z}_{ik} \bar{\alpha}_t \mathbf{1}_{\{j=c(i,t)\}}$$

and

$$\delta_{ijs} \quad = \quad \begin{cases} 1 & \text{if } j = c(i,s) \text{ and } 1 \leq s \leq n_i \\ 1 & \text{if } j \neq c(i,l) \text{ for } 1 \leq l \leq n_i \text{ and } s = N+1 \\ 0 & \text{otherwise} \end{cases} \quad (5.4)$$

and equating (5.3) to zero gives

$$\frac{\omega_{kj}}{p_{kj}} \quad = \quad \sum_{i=1}^{M}\sum_{t=1}^{n_i} \hat{z}_{ik} \left\{ \sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \right\}^{-1} \left\{ \sum_{s=t}^{(N+1)} \bar{\alpha}_t \bar{p}_{kj}^{\bar{\alpha}_t - 1} \delta_{ijs} \right\}$$

which implies that

$$\hat{p}_{kj} \quad = \quad \frac{\omega_{kj}}{\sum_{i=1}^{M}\sum_{t=1}^{n_i} \hat{z}_{ik} \left\{ \sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \right\}^{-1} \left\{ \sum_{s=t}^{(N+1)} \bar{\alpha}_t \bar{p}_{kj}^{\bar{\alpha}_t - 1} \delta_{ijs} \right\}}.$$

5.3.4 Maximization with Respect to Dampening Parameters.

In this case the original complete data log likelihood function (5.2) is treated as a function of α_t. p_{kj} is treated as a constant with $\bar{p}_{kj}$ denoting the estimate of p_{kj} from the previous iteration of the MM algorithm. Applying (4.4) to the problematic term $-\log \sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\alpha_t}$ gives

$$-\log \sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\alpha_t} \;\; \geq \;\; -\log \sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \;\; + \;\; 1 \;\; - \;\; \frac{\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\alpha_t}}{\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t}}.$$

It therefore follows, up to a constant,

$$Q(\underline{\alpha}) \;\; \geq \;\; q_1(\underline{\alpha}) \;\; = \;\; \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{t=1}^{n_i} \hat{z}_{ik} \left\{ \alpha_t \log \bar{p}_{kc(i,t)} \;\; + \;\; \left(\frac{-\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\alpha_t}}{\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t}} \right) \right\}.$$

Similar to the maximization with respect to p_{kj}, this surrogate function is still difficult to optimize. Also, the function $f(\alpha) = -\bar{p}^{\alpha}$ is a concave function and minorization by a linear surrogate function is not possible. Bounding a concave function $f(\theta)$ around θ^n using a quadratic gives

$$f(\theta) \;\; \geq \;\; f(\theta^n) + \left[f'(\theta^n)\right]^T (\theta - \theta^n) + \frac{1}{2}(\theta - \theta^n)^T B(\theta - \theta^n)$$

where B is a negative definite matrix, $H(\theta^n) - B > 0$ and $H(\theta^n)$ is the Hessian $d^2 f/d(\theta^n)^2$. Thus for $f(\alpha) = -\bar{p}^{\alpha}$

$$-\bar{p}^{\alpha} \;\; \geq \;\; -\bar{p}^{\bar{\alpha}} - (\log \bar{p})\bar{p}^{\bar{\alpha}}(\alpha - \bar{\alpha}) - 1/2(\alpha - \bar{\alpha})^2 (\log \bar{p})^2$$

because $H(\bar{\alpha}) > B = -(\log \bar{p})^2$. Hence the surrogate function becomes

$$\begin{aligned} q_1(\underline{\alpha}) \;\; &\geq \;\; q_2(\underline{\alpha}) \\ &= \;\; \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{t=1}^{n_i} \hat{z}_{ik} \left[\alpha_t \log \bar{p}_{kc(i,t)} + \left(\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \right)^{-1} \left\{ \sum_{s=t}^{N} \left(-\log \bar{p}_{kc(i,s)} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} (\alpha_t - \bar{\alpha}_t) \right. \right. \right. \\ &\qquad \left. \left. \left. -1/2(\alpha_t - \bar{\alpha}_t)^2 (\log \bar{p}_{kc(i,s)})^2 \right) \right\} \right] \end{aligned}$$

up to a constant. Iterative maximization of this surrogate function with respect to α_t leads to a sequence of $\hat{\alpha}_t$ values that converge to a local maximum of Q. Thus

$$
\begin{aligned}
\frac{\partial q_2(\underline{\alpha})}{\partial \alpha_t} \;=\; & \sum_{i=1}^{M}\left\{\sum_{k=1}^{K}\hat{z}_{ik}\left[\log \bar{p}_{kc(i,t)} + \left(\sum_{s=t}^{N}\bar{p}_{kc(i,s)}^{\bar{\alpha}_t}\right)^{-1}\left\{\sum_{s=t}^{N}\Big(-\log \bar{p}_{kc(i,s)}\bar{p}_{kc(i,s)}^{\bar{\alpha}_t}\right.\right.\right. \\
& \left.\left.\left. -(\alpha_t-\bar{\alpha}_t)(\log \bar{p}_{kc(i,s)})^2\Big)\right\}\right]\right\}.\mathbf{1}_{\{t\le n_i\}}
\end{aligned}
$$

which implies that

$$
\hat{\alpha}_t = \frac{\displaystyle\sum_{i=1}^{M}\left\{\sum_{k=1}^{K}\hat{z}_{ik}\left[\log \bar{p}_{kc(i,t)} - \left(\sum_{s=t}^{N}\bar{p}_{kc(i,s)}^{\bar{\alpha}_t}\right)^{-1}\left\{\sum_{s=t}^{N}\log \bar{p}_{kc(i,s)}\bar{p}_{kc(i,s)}^{\bar{\alpha}_t} - \bar{\alpha}_t(\log \bar{p}_{kc(i,s)})^2\right\}\right]\right\}.\mathbf{1}_{\{t\le n_i\}}}{\displaystyle\sum_{i=1}^{M}\left\{\sum_{k=1}^{K}\hat{z}_{ik}\left(\sum_{s=t}^{N}\bar{p}_{kc(i,s)}^{\bar{\alpha}_t}\right)^{-1}\sum_{s=t}^{N}(\log \bar{p}_{kc(i,s)})^2\right\}.\mathbf{1}_{\{t\le n_i\}}}.
$$

5.3.5 The EM/MM Algorithm.

In summary, when fitting mixtures of Benter models the EM algorithm (incorporating the MM algorithm) reduces to the following steps:

0. Let $l = 0$ and choose initial parameter estimates $\underline{\pi}^{(0)}$, $\mathbf{p}^{(0)}$ and $\underline{\alpha}^{(0)}$.

1. **E step:** Compute the quantities

$$
z_{ik}^{(l+1)} = \frac{\pi_k^{(l)} f(\underline{x}_i|\underline{p}_k^{(l)}, \underline{\alpha}^{(l)})}{\sum_{k'=1}^{K}\pi_{k'}^{(l)} f(\underline{x}_i|\underline{p}_{k'}^{(l)}, \underline{\alpha}^{(l)})}
$$

2. **M step:** Compute

$$
\pi_k^{(l+1)} = \frac{\sum_{i=1}^{M} z_{ik}^{(l+1)}}{M}
$$

$$
p_{kj}^{(l+1)} \;=\; \frac{\omega_{kj}}{\displaystyle\sum_{i=1}^{M}\sum_{t=1}^{n_i} z_{ik}^{(l+1)}\left\{\sum_{s=t}^{N}\bar{p}_{kc(i,s)}^{\bar{\alpha}_t}\right\}^{-1}\left\{\sum_{s=t}^{(N+1)}\bar{\alpha}_t\bar{p}_{kj}^{\bar{\alpha}_t-1}\delta_{ijs}\right\}}
$$

(where $\bar{p}_{kj}$ and $\bar{\alpha}_t$ denote p_{kj}^{l} and α_t^{l} respectively. ω_{kj} and δ_{ijs} are as defined in Chapter 5.3.3.)

$$\alpha_t^{(l+1)} =$$

$$\frac{\sum_{i=1}^{M}\left\{\sum_{k=1}^{K} z_{ik}^{(l+1)}\left[\log \bar{p}_{kc(i,t)} + \left(\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t}\right)^{-1}\left\{\sum_{s=t}^{N} - \log \bar{p}_{kc(i,s)} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} + \bar{\alpha}_t (\log \bar{p}_{kc(i,s)})^2\right\}\right]\right\} .\mathbf{1}_{\{t \leq n_i\}}}{\sum_{i=1}^{M}\left\{\sum_{k=1}^{K} z_{ik}^{(l+1)}\left(\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t}\right)^{-1} \sum_{s=t}^{N} (\log \bar{p}_{kc(i,s)})^2\right\} .\mathbf{1}_{\{t \leq n_i\}}}.$$

(where $\bar{p}_{kj}$ and $\bar{\alpha}_t$ denote p_{kj}^l and α_t^l respectively.)

3. If converged, then stop. Otherwise, increment h and return to Step 1.

5.4 Estimation of Standard Errors

As detailed in Chapter 4.4, the covariance matrix of the estimated model parameters can be approximated by the empirical observed information matrix $\mathcal{I}_e$ which can be expressed in terms of the score function of the complete data log likelihood. Specifically for parameter estimates θ

$$\mathcal{I}_e \;=\; \sum_{i=1}^{M} s(\underline{x}_i|\theta) s^T(\underline{x}_i|\theta).$$

where $s(\underline{x}_i|\theta) = \frac{\partial \log L_i(\theta)}{\partial \theta}$.

From the complete data log likelihood:

$$Q(\theta) = \sum_{i=1}^{M}\sum_{k=1}^{K} \hat{z}_{ik}\{\log \pi_k + \sum_{t=1}^{n_i} \alpha_t \log p_{kc(i,t)} - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}\}.$$

the score function with respect to π_k for $k = 1, \ldots, K$ is

$$\begin{aligned} s(\underline{x}_i|\pi_k) &= \frac{\partial Q}{\partial \pi_k} \\ &= \frac{\hat{z}_{ik}}{\pi_k}. \end{aligned}$$

Similarly the score function with respect to p_{kj} for $k = 1, \ldots, K$ and $j = 1, \ldots, N$ is

$$
\begin{aligned}
s(\underline{x}_i|p_{kj}) &= \frac{\partial Q}{\partial p_{kj}} \\
&= \hat{z}_{ik}\left[\sum_{t=1}^{n_i}\left\{\frac{\alpha_t \mathbf{1}_{\{j=c(i,t)\}}}{p_{kj}} - \frac{\alpha_t p_{kj}^{\alpha_t-1}\mathbf{1}_{\{j=c(i,s)\text{for } s=t,\ldots,N.\}}}{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}}\right\}\right] \\
&= \hat{z}_{ik}\left[\sum_{t=1}^{n_i}\left\{\frac{\alpha_t \mathbf{1}_{\{j=c(i,t)\}}}{p_{kj}} - \frac{\sum_{s=t}^{N+1}\alpha_t p_{kj}^{\alpha_t-1}\delta_{ijs}}{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}}\right\}\right].
\end{aligned}
$$

where δ_{ijs} is defined by (5.4).

Finally the score function with respect to α_t for $t = 2, \ldots, N-1$ is calculated as

$$
\begin{aligned}
s(\underline{x}_i|\alpha_t) &= \frac{\partial Q}{\partial \alpha_t} \\
&= \sum_{k=1}^{K}\hat{z}_{ik}\left[\mathbf{1}_{t\leq n_i}\left\{\log p_{kc(i,t)} - \frac{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}\log p_{kc(i,s)}}{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}}\right\}\right].
\end{aligned}
$$

Formulating a matrix S which contains the score function for each voter evaluated for each parameter on convergence of the EM algorithm, setting the empirical observation matrix $I_e = S^T S$ and taking the square root of the diagonal of I_e^{-1}, the approximate standard errors of the parameters of a mixture of Benter models emerges. In Chapter 4 due to computing constraints the inverse of the expected information matrix was approximated by the inverse of the diagonal terms — this is not the case here as inverting the full expected information matrix is computationally feasible.

In the maximization step of the EM algorithm the Newton-Rhapson optimization technique could be employed using the approximated information matrix detailed in this section. The Newton-Rhapson technique involves inverting the information matrix at each iteration; the implementation of the MM algorithm at the M step of the EM algorithm as detailed neatly avoids the inversion of such a $[(K-1) + K(N-1) + (N-2)]^2$ information matrix.

5.5 Model Comparison

Many different mixture models were fitted to the election data sets by varying the component models (i.e. the Benter model or the Plackett-Luce model) and the number of components. A criterion is required for comparing the fitted models.

A range of model selection techniques are detailed in Chapter 3. The BIC was used as the main model fitting tool in this analysis. The cross-validated likelihood method suggested the maximum number of groups fitted as the best model which was deemed to be a case of overfitting. The BIC consistently returned the most parsimonious and interpretable models and was used throughout as the main model selection tool. The BIC is defined to be

$$BIC = 2(\text{maximized likelihood}) - (\text{number of parameters})\log(M)$$

which can be viewed as a criterion which rewards model fit, but penalizes model complexity. Table 5.3 details the parameters to be estimated within each type of model considered.

Table 5.3: The number of parameters in the various types of mixture models proposed for modelling Irish election data.

Model	**Proportions**	**Support**	**Dampening**
Plackett-Luce	$K-1$	$K(N-1)$	0
Plackett-Luce (with Noise)	$K-1$	$(K-1)(N-1)$	0
Benter	$K-1$	$K(N-1)$	$N-2$
Benter (with Noise)	$K-1$	$(K-1)(N-1)$	$N-2$

5.6 Analysis of the Irish Electorate

The proposed mixture model approach for exploring heterogeneity within the Irish electorate is demonstrated on Irish presidential and general election data. The analysis of the electorates of these elections using this approach establishes that there are homogeneous sub-populations of voters in the electorate and the form of these sub-populations is revealed.

5.6.1 The 1997 Presidential Election

Mixtures of Plackett-Luce models and mixtures of Benter models, with up to 10 components, were fitted to the 1997 presidential election data sets. The BIC was used as the model selection criterion. For all polls (with the exception of two) the

Plackett-Luce model with varying numbers of components was selected. In some of these polls a mixture of Plackett-Luce models which included a noise component was deemed the best model. For the two polls on 23/10 and 30/10 mixtures of Benter models were selected but the difference in BIC between the Benter and Plackett-Luce mixtures was small. Thus mixtures of Plackett-Luce models are reported for all polls (Table 5.4) for ease of comparison.

Examination of Table 5.4 shows that the Irish electorate began the 1997 presidential campaign as a single group which then partitioned over the course of the campaign.

At the beginning of the campaign, as demonstrated by the first two polls, the electorate appeared to be composed of a single component which had larger levels of support for the three most high profile candidates — Banotti, McAleese and Roche. However Roche's support dropped by almost 10% between the polls taken on 18/9 and 27/9. As mentioned in Chapter 5.2, shortly after the initial nominations of candidates Adi Roche, who up until then had been the bookies favourite, was publicly criticized by fellow workers and her popularity dropped off significantly. This drop in support for Roche continued throughout all the polls detailed.

A month before polling day, demonstrated by the 2/10 poll, 40% of the electorate were best modelled as noise. The electorate appears to have become partitioned into a noise group and the original group who supported the high profile candidates of Banotti, McAleese and Roche. Perhaps Roche's drop off in popularity left some undecided voters.

By 11/10, the future pattern of the presidential race became clear. The Banotti and McAleese camps emerged strongly with the group weighted towards McAleese making up the larger 73% of the electorate. Notably, the group who favored McAleese also appear to have a good level of support for Banotti.

Between the polls conducted on 11/10 and the 22/10 a great deal of controversy arose in the presidential campaign. It was reported that Mary McAleese had sympathies with the republican party Sinn Féin which would have had a detrimental effect on her support. Further fuel was added to these allegations when the president of the Sinn Féin party gave McAleese the party's backing. Throughout this period McAleese consistently denied the claims and after defending her position well in a

Table 5.4: Parameter estimates when mixtures of Plackett-Luce models were fitted to each of the eight presidential election poll data sets are reported. Standard errors associated with these estimates are given in parentheses.

Date	**Banotti**	**McAleese**	**Nally**	**Roche**	**Scallon**	$\hat{\pi}_k$
18/9	0.23 (0.005)	0.34 (0.014)	-	0.35 (0.014)	0.08 (0.002)	*1.00*
27/9	0.28 (0.007)	0.39 (0.014)	-	0.26 (0.009)	0.07 (0.002)	*1.00*
2/10	0.32 (0.010)	0.42 (0.029)	0.07 (0.004)	0.16 (0.011)	0.02 (0.002)	*0.60* (0.090)
	0.20	0.20	0.20	0.20	0.20	*0.40* (0.036)
11/10	0.56 (0.059)	0.09 (0.010)	0.11 (0.011)	0.20 (0.027)	0.05 (0.005)	*0.27* (0.062)
	0.20 (0.004)	0.50 (0.010)	0.10 (0.004)	0.14 (0.006)	0.07 (0.003)	*0.73* (0.073)
22/10	0.57 (0.123)	0.03 (0.007)	0.10 (0.011)	0.09 (0.019)	0.03 (0.005)	*0.14* (0.059)
	0.28 (0.003)	0.53 (0.034)	0.05 (0.034)	0.10 (0.006)	0.04 (0.004)	*0.55* (0.102)
	0.18 (0.007)	0.32 (0.015)	0.10 (0.009)	0.16 (0.016)	0.24 (0.037)	*0.31* (0.137)
23/10	0.92 (0.001)	0.02 (0.012)	0.03 (< 0.001)	0.02 (0.002)	0.02 (0.005)	*0.16* (0.047)
	0.02 (0.025)	0.92 (0.002)	0.01 (0.003)	0.02 (0.002)	0.04 (0.002)	*0.20* (0.047)
	0.33 (0.002)	0.47 (0.008)	0.05 (0.004)	0.12 (0.005)	0.03 (0.003)	*0.44* (0.089)
	0.20	0.20	0.20	0.20	0.20	*0.20* (0.030)
25/10	0.96 (0.164)	0.01 (< 0.001)	0.01 (< 0.001)	0.02 (0.002)	0.01 (0.001)	*0.16* (0.042)
	0.00 (< 0.001)	1.00 (0.010)	0.00 (< 0.001)	0.00 (< 0.001)	0.00 (< 0.001)	*0.14* (0.037)
	0.25 (< 0.001)	0.61 (< 0.001)	0.04 (< 0.001)	0.07 (< 0.001)	0.03 (0.002)	*0.49* (0.057)
	0.20	0.20	0.20	0.20	0.20	*0.21* (0.026)
30/10	0.81 (0.054)	0.01 (0.001)	0.06 (0.001)	0.07 (0.004)	0.05 (0.003)	*0.23* (0.028)
	0.01 (0.001)	0.83 (0.004)	0.02 (< 0.001)	0.03 (0.001)	0.11 (0.003)	*0.27* (0.028)
	0.25 (0.011)	0.58 (0.011)	0.04 (0.017)	0.08 (0.025)	0.04 (0.042)	*0.36* (0.036)
	0.12 (0.017)	0.09 (0.041)	0.15 (0.002)	0.17 (0.004)	0.47 (0.003)	*0.14* (0.078)

nationally broadcast current affairs program on October 20th she re-established herself. In fact, the false allegations had a larger detrimental effect on her presidential competitors, some of whom had publicly castigated her about the allegations.

These events are mirrored by the results of the polls taken on 22/10 and 23/10. On 22/10 the electorate is composed of three components. Again the strongly Banotti group was present, the strongly McAleese group (with some Banotti support) was the largest group making up 55% of the electorate and 31% of the electorate seemed to be nearly a noise component with a conservative flavour. The larger support in this third group was for the two conservative candidates McAleese and Scallon. Scallon's performance in the campaign was beginning to win her votes.

The results of the 23/10 poll indicate how well McAleese recovered and gained from the Sinn Féin controversy. The electorate really partitions at this stage into a group of Banotti supporters, a group of McAleese supporters and a group of voters who support the high profile candidates McAleese, Banotti and Roche; one fifth of the electorate are still modelled as noise. The results of the poll taken on 25/10 are very similar — the main theme of the four components remains the same, with the probability of belonging to each group altering slightly. The group with support for the candidates with the higher profiles (and supported by the larger parties) makes up almost half of the electorate.

The changes in the composition of the electorate between 25/10 and polling day pay tribute to Scallon's performance throughout her campaign — again the theme of each of the four sections of the electorate are similar but the estimated support parameters for each candidate drop in nearly every group, with the exception of Scallon. Her support parameters in each of the four groups are significantly higher than they were in the 25/10 poll. Figure 5.2 provides a graphical representation of the estimated model parameters of the 30/10 exit poll.

In summary, the mixture model finds groups of voters which appear logical in the context of this presidential election. One possible explanation for the predominant choice of the Plackett-Luce model over the Benter model is that there were only five candidates in the election. Thus the electorate was very familiar with all of the candidates explaining why lower preferences were made with as much certainty as higher preferences.

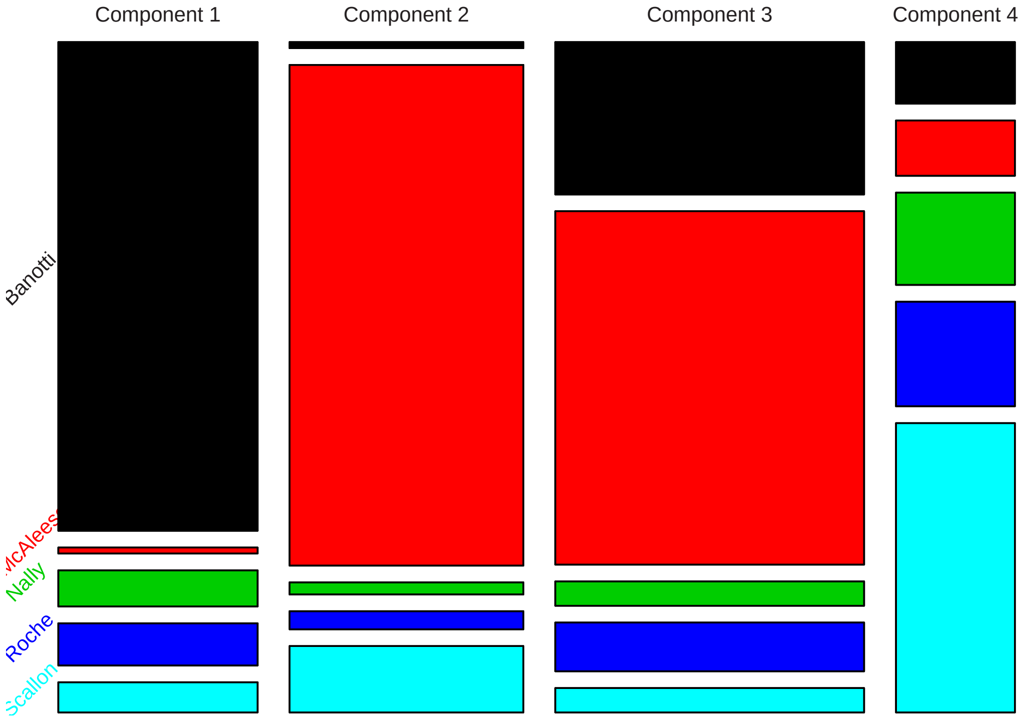


Fig. 5.2: A mosaic plot representation of the mixture of Plackett-Luce models fitted to the Lansdowne exit poll data (conducted on 30/10) for the 1997 presidential election. The column widths represent the mixture proportions and the columns are divided into sections representing the support parameter estimates for each candidate within the component.

5.6.2 The 2002 General Election

Mixtures of Plackett-Luce and Benter models were fitted to the data from the Dublin West constituency (see Chapter 2). The mixture with the highest BIC value was a fifteen component Benter mixture and is reported in Table 5.5 and Figures 5.3 and 5.4.

The support parameter estimates reported in Table 5.5 all have standard errors less than 5×10^{-3} with the exception of two — Lenihan's support parameter in component 6 has a standard error of 8×10^{-3} and McDonald's support parameter in component 10 has an associated standard error of 1×10^{-2}. The final row of the table gives the mixture component probabilities whose standard errors were all less than 8×10^{-3}. Benter dampening parameter estimates for the Dublin West data were

$$\underline{\hat{\alpha}} = (1.00,\ 1.00,\ 0.95,\ 0.74,\ 0.57,\ 0.41,\ 0.28,\ 0.15,\ 0.00) \tag{5.5}$$

with the associated standard errors all less than 1×10^{-2}.

Table 5.5: Fifteen component mixture of Benter models fitted to the Dublin West constituency data.

Candidate	**Components**														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Bon	0.01	0.02	0.00	0.03	0.01	0.01	0.13	0.04	0.13	0.06	0.03	0.01	0.00	0.04	0.01
Bur	0.01	0.24	0.00	0.17	0.13	0.09	0.17	0.25	0.22	0.39	0.01	0.03	0.05	0.04	0.01
Do-Ry	0.23	0.03	0.19	0.01	0.11	0.08	0.09	0.00	0.00	0.00	0.00	0.08	0.11	0.07	0.07
Hig	0.02	0.03	0.00	0.62	0.05	0.37	0.11	0.05	0.40	0.36	0.52	0.54	0.00	0.03	0.00
Len	0.70	0.11	0.80	0.03	0.45	0.36	0.13	0.00	0.00	0.02	0.00	0.20	0.55	0.07	0.22
McD	0.02	0.00	0.00	0.07	0.05	0.00	0.06	0.01	0.09	0.00	0.42	0.13	0.00	0.72	0.00
Mor	0.02	0.21	0.01	0.02	0.08	0.05	0.15	0.06	0.04	0.05	0.00	0.01	0.20	0.02	0.68
Smy	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.00
Ter	0.00	0.37	0.00	0.05	0.11	0.04	0.14	0.59	0.09	0.11	0.00	0.00	0.09	0.01	0.01
$\underline{\pi}$	0.10	0.09	0.09	0.09	0.08	0.08	0.07	0.07	0.06	0.06	0.06	0.05	0.05	0.03	0.02

The mixture model reveals some interesting features in the electorate. The components can be summarized as follows:

1. This component gives almost all of its support to Fianna Fáil. Lenihan gets more support than his running mate, Doherty-Ryan. This component is similar

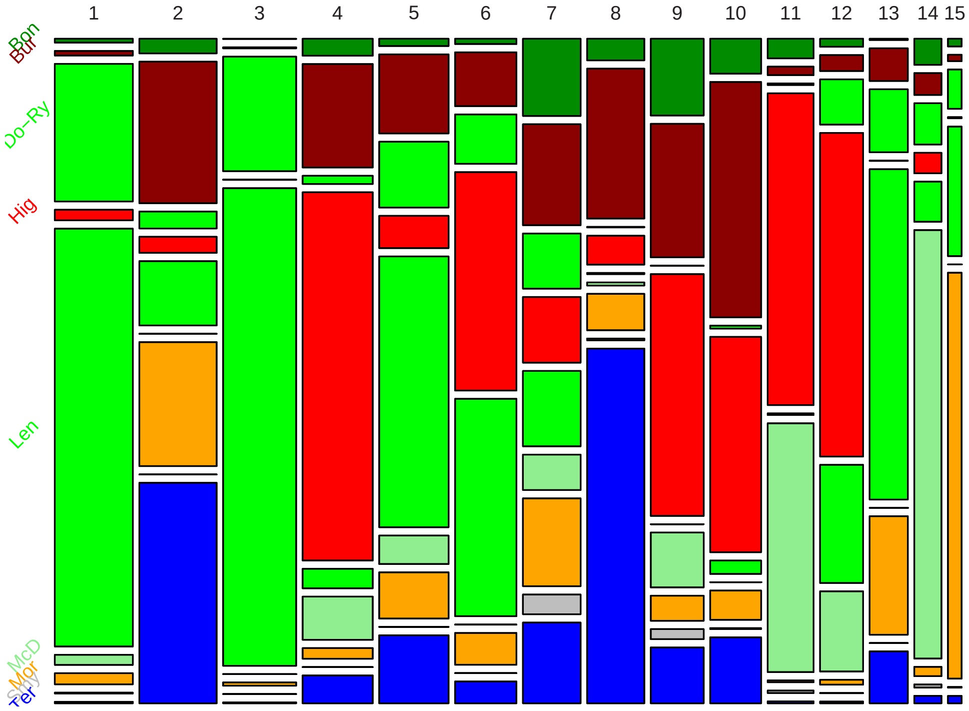


Fig. 5.3: A mosaic plot representation of the 15 group Benter mixture model fitted to the Dublin West data. The width of each column illustrates the mixing proportion for each group and the sections within each column represent the support parameter for each candidate within that group.

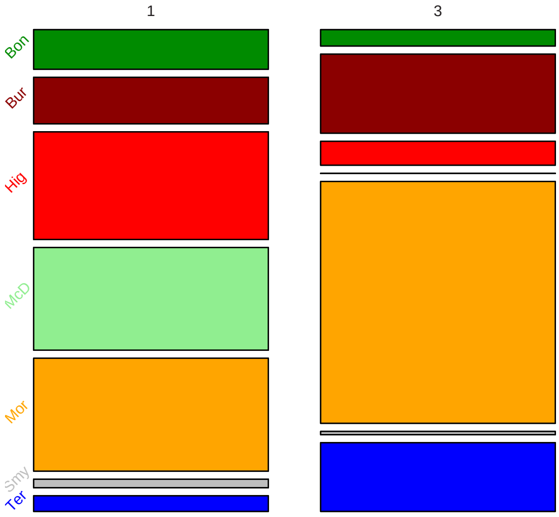


Fig. 5.4: A mosaic plot representing the parameter estimates for groups 1 and 3 of the Benter mixture model fitted to the Dublin West data. The two Fianna Fáil candidates (Doherty-Ryan and Lenihan) have been removed to determine the subtle differences between the groups.

to Component 3, but there are subtle differences (see Figure 5.4 and Component 3).

2. The support is divided between Fine Gael, Labour and the Progressive Democrats. These are the three parties that are next largest after Fianna Fáil in terms of the number of seats held in government.

3. Almost all of the support is for Fianna Fáil. Again, Lenihan gets more support than his running mate, Doherty Ryan. This component differs from Component 1, in that the candidate support conditional on having ranked the Fianna Fáil candidates Lenihan and Doherty-Ryan in first and second place (in either order) is strongly for the Progressive Democrats candidate Morrissey (Figure 5.4). Interestingly, the Progressive Democrats were in coalition government with Fianna Fáil prior to the election. Thus this component would appear to be the voters in the electorate who were in favor of a Fianna Fáil/Progressive Democrats coalition government.

4. The support is mainly for Joe Higgins of the Socialist party, but most of the remaining support is divided between the Labour and Sinn Féin candidates; historically, these would have been seen as left wing parties.

5. This component gives a lot of support to Lenihan, but divides its support quite evenly between Burton, Doherty-Ryan, Terry and Morrissey after that. This component appears to be predominantly candidate centered on Lenihan.

6. The support here is primarily for Higgins and Lenihan. These candidates are from very different parties, but both candidates have a very high profile within the constituency. There is reason to believe that this component may be geographically based.

7. This component shows almost uniform support for most of the major candidates in this constituency. Smyth and McDonald receive considerably less support than the other candidates.

8. The support is divided between the Fine Gael and Labour candidates. These two parties encourage voters to transfer their lower preferences between these parties. These parties are former coalition government parties (1994–1997).

9. Higgins and Burton get most of the support. Higgins and Burton are high profile candidates in the constituency. They are both from left-wing parties. The support for Bonnie, McDonald and Terry could be explained on similar party or idealistic grounds. This component has extremely low support for Fianna Fáil.

10. Burton and Higgins get most of the support with Terry receiving a moderate amount of support. Terry's party is closely linked to Burton's party (see Component 8).

11. This component shows support for the Socialist and Sinn Féin candidates. These are the two most left-wing candidates in the constituency.

12. Higgins, Lenihan and McDonald have the majority of the support. The candidates are from parties that are quite different. These candidates are all high profile. The relationship between these candidates is difficult to explain.

13. This component has strong support for the candidates from the two government coalition parties (Fianna Fáil and Progressive Democrats).

14. McDonald of Sinn Féin receives the majority of the support in this component.

15. The support mainly goes to the Progressive Democrats candidate. The remaining support is for the two Fianna Fáil candidates. All most all of the support is for the previous government coalition parties.

The fifteen component mixture of Benter models that was selected using BIC gives clear and meaningful groups. The groups confirm the idea that Irish elections are influenced by both candidate and party politics (Bowler and Farrell, 1991; Marsh, 2000). The mixture model found in this analysis provides strong support for this description of how Irish elections work in practice.

The estimated dampening parameter $\underline{\hat{\alpha}}$ (5.5) in the Benter mixture is also of interest. The parameter estimate shows that the first two preferences are very carefully chosen ($\alpha_1 = \alpha_2 = 1$) and that later preferences become more random (higher entropy) as α_t decreases with t. The parameter estimates also suggest that the choice of candidate at the last two choice levels is essentially uniform (maximum entropy). This is interesting, because one could postulate that the high and low preferences are made very carefully and that the middle preferences are very random. However, the fitted estimate indicates that choices get more random as a ballot is completed. Laver (2004) noted that the median (and modal) number of preferences expressed by voters was three in this constituency. His findings may also support the idea that voters give a few top preferences carefully and after that they either don't select candidates or they select them in a more random manner.

5.7 Conclusions

The proposed mixtures of Plackett-Luce and Benter models provide an interpretable model for PR-STV election data. The models can be used to discover and model any heterogeneity present in voting behaviour of the electorate.

The use of a noise component in the mixture models was found to be advantageous. The component accounted for small groups of voters who didn't fit into the main groups in the mixture. This is in agreement with previous uses of noise components in mixtures.

The model fitting by maximum likelihood using the EM and MM algorithms provides an efficient method for fitting these models.

In the general election context, no covariate information was available for the voters. However, in the case of opinion polls, including covariate information could provide insight into the form which groups in the electorate take. Mixtures-of-experts models which avail of such covariates are detailed in Chapter 7.

Chapter 6

A Grade of Membership Model for Rank Data

A major advantage of fitting mixture models via the EM algorithm, as detailed by Fraley and Raftery (1998), is that the value of the imputed missing membership labels at convergence is an estimate of the conditional probability that an object belongs to each component of the mixture. These values can be used to cluster observations into groups. The grade of membership or mixed membership model (Erosheva, 2003) provides similar group membership probabilities but in this case the probabilities are direct parameters of the model. The grade of membership model allows objects to have partial membership of each of the homogeneous subgroups which constitute the population. In this chapter the parameters of the grade of membership model when fitted to data from the 1997 Irish presidential election are estimated within a Bayesian framework; the uncertainty associated with the model parameters can therefore be quantified.

6.1 Model Specification

Irish voting data (see Chapter 2) possess some unique properties which require careful statistical modelling. The grade of membership (GoM) model is used to model the heterogeneity within the electorate alongside the Plackett-Luce model which models the ranked nature of the preferences expressed by the voters.

6.1.1 The Plackett-Luce Model

It is assumed that the electorate is a heterogeneous population composed of K homogeneous 'extreme profiles' of which each voter may have partial membership. Each of the K extreme profiles is characterized by a specific parameterization of a probability density. In the case of modelling Irish election data the characterizing densities are Plackett-Luce densities with different parameterizations (see Chapter 3.1). Given that voter i expressed n_i preferences and is a complete member of extreme profile k, under the Plackett-Luce model the probability of voter i's ballot $\underline{x}_i$ is

$$\begin{aligned}\mathbf{P}\{\underline{x}_i|\underline{p}_k\} &= \prod_{t=1}^{n_i} \frac{p_{kc(i,t)}}{p_{kc(i,t)} + p_{kc(i,t+1)} + \cdots + p_{kc(i,N)}} \\ &= \prod_{t=1}^{n_i} q_{kc(i,t)} \qquad (6.1)\end{aligned}$$

where $c_{(i,t)}$ denotes the candidate in position t in vote i.

6.1.2 The Grade of Membership Model

Mixture models provide a flexible suite of modelling tools which model a population as a finite collection of homogeneous sub-groups, each of which is characterized by a specific parameterization of a probability density. While based on a similar concept, GoM models allow each member of the population have partial membership of each of the homogeneous sub-groups (or 'extreme profiles') which constitute the population. Thus a soft clustering of the population members is achievable.

The GoM model originally appeared in the 1970s where it was employed in the context of medical diagnosis problems. Manton et al. (1994) provide a full description. Early estimation methods for the GoM model were maximum likelihood based. Erosheva (2002) reformulated the GoM model as a hierarchical Bayesian mixed-membership model — Airoldi et al. (2006) discuss model choice within such a framework. Erosheva (2003) estimated the GoM model for multivariate categorical data within a Bayesian framework; a similar approach is taken here to estimate the GoM model for rank data.

Under the GoM model each voter $i = 1, \ldots, M$ has an associated *GoM score* or *mixed membership vector*

$$\underline{\pi}_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{iK})$$

where $\sum_{k=1}^{K} \pi_{ik} = 1$ and $\pi_{ik} > 0$ for $k = 1, \dots, K$. The GoM score $\underline{\pi}_i$ describes the probability of voter i's membership of each of the K extreme profiles within the electorate. Given that each extreme profile is characterized by a Plackett-Luce density, the likelihood of the votes cast $\mathbf{x}$ is

$$\mathbf{P}\{\mathbf{x}|\pi, \theta\} = \prod_{i=1}^{M} \prod_{t=1}^{n_i} \left[\sum_{k=1}^{K} \pi_{ik} q_{kc(i,t)} \right]$$

where $q_{kc(i,t)}$ is given by (6.1).

As in Erosheva (2003) a latent class representation of the GoM model is considered which can provide insight into any unobservable underlying phenomenon. Latent class models involve augmenting the data with categorical latent variables — these define the latent classes. For each voter i, n_i binary vectors $\underline{z}_{it}$ of length K are imputed where

$$\underline{z}_{it} = (0, \dots, 1, \dots, 0)$$

and

$$z_{itk} = \begin{cases} 1 & \text{with probability } \pi_{ik} \\ 0 & \text{otherwise.} \end{cases}$$

It follows that under the GoM model the 'complete data likelihood' of all the data and the latent variables is therefore:

$$\mathbf{P}\{\mathbf{x}, \mathbf{z}|\pi, \theta\} = \prod_{i=1}^{M} \prod_{k=1}^{K} \prod_{t=1}^{n_i} \left\{ \pi_{ik} q_{kc(i,t)} \right\}^{z_{itk}}.$$

A soft clustering of the electorate can be inferred by modelling the ranked preferences using the GoM model and incorporating the Plackett-Luce model.

6.1.3 Prior and Posterior Distributions

A Bayesian approach (see Chapter 3) is used to estimate the GoM model and thus the specification of prior distributions for the parameters of the model is required. It is assumed that the mixed membership variables for each voter follow a Dirichlet($\underline{\alpha}$)

distribution and that the support parameters within each extreme profile follow a Dirichlet($\underline{\beta}$) distribution i.e.

$$\begin{aligned}\underline{\pi}_i &\sim \text{Dirichlet}\left\{\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)\right\}\\ \underline{p}_k &\sim \text{Dirichlet}\left\{\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)\right\}.\end{aligned}$$

Definition 8 *If* $\underline{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})$ *has a* ***Dirichlet distribution*** *with* K *categories and parameters* $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ *then the density function of* $\underline{\pi}_i$ *is*

$$\text{Dir}(\underline{\pi}_i|\underline{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\pi_{ik}^{\alpha_k-1}$$

where $\alpha_k > 0$, $\alpha_0 = \sum_{k=1}^{K}\alpha_k$ *and* $\sum_{k=1}^{K}\pi_{ik} = 1$. ▮

The Dirichlet distribution is easily obtained from the basis of independent, equally scaled, gamma distributed components. Also, since

$$\text{corr}(\pi_{ik}, \pi_{il}) = -(\alpha_k\alpha_l)^{1/2}\left\{(\alpha_0 - \alpha_k)(\alpha_0 - \alpha_l)\right\}^{-1/2}$$

for $k, l, = 1, \ldots, K$ the correlation between components is always negative. Thus the strong structure imposed by using a Dirichlet prior may not be suitable in cases where positive dependence is exhibited between components.

Erosheva (2003) employed Dirichlet priors for reasons of conjugacy with the multinomial distribution. Dirichlet priors are also employed in this application. Here the prior parameters are fixed as $\underline{\alpha} = (0.5, \ldots, 0.5)$ and $\underline{\beta} = (0.5, \ldots, 0.5)$ which is the neutral Jeffreys prior for the multinomial distribution (see O'Hagan and Forster, 2004, Chapter 5.35). Further work on specifications of the prior parameters is discussed in Chapter 6.5 and Chapter 9.

Given the prior distributions and the complete data likelihood for the GoM model, where each extreme profile is deemed to be characterized by some parameterization of the Plackett-Luce density, the posterior distribution of all the votes cast is

$$\mathbf{P}\{\mathbf{X}, \mathbf{z}, \pi, \mathbf{p}\} = \left[\prod_{i=1}^{M}\prod_{k=1}^{K}\prod_{t=1}^{n_i}\{\pi_{ik}q_{kc(i,t)}\}^{z_{itk}}\right]\left[\prod_{i=1}^{M}\prod_{k=1}^{K}\pi_{ik}^{\alpha_k-1}\right]\left[\prod_{k=1}^{K}\prod_{j=1}^{N}p_{kj}^{\beta_j-1}\right]. \quad (6.2)$$

6.2 Parameter Estimation

Markov chain Monte Carlo (MCMC) methods can be employed to produce realizations of the model parameters by sampling from the relevant posterior distribution. In particular, the MCMC technique known as the Gibbs sampler can be employed if the complete conditional distributions for all model parameters are available for sampling. Chapter 3 provides full details of some MCMC simulation methods.

6.2.1 The Gibbs Sampler

To implement the Gibbs sampler the complete conditional distributions of the model parameters are required. Often the complete conditional distribution is recognizable as a standard distribution and ways of drawing random samples from such common distributions have been extensively studied. One such standard distribution which arises in this GoM context is the Multinomial distribution.

Definition 9 *If a set of random variables $\underline{x} = (x_1, \ldots, x_M)$ have a probability function*

$$\mathbf{P}\{x_1, \ldots, x_M\} = \frac{M!}{\prod_{i=1}^{M} x_i!} \prod_{i=1}^{M} \theta_i^{x_i}$$

where x_i are nonnegative integers such that $\sum_{i=1}^{M} x_i = M$ and θ_i are constants such that $\theta_i > 0$ and $\sum_{i=1}^{M} \theta_i = 1$ then the joint distribution of $\underline{x} = (x_1, \ldots, x_M)$ is a ***Multinomial distribution****.* ▮

When implementing the Gibbs sampler for the GoM model the complete conditional distributions for the latent variables $\underline{z}_{it}$ and the GoM scores $\underline{\pi}_i$ are readily available. From the posterior distribution (6.2) the complete conditional distribution for the latent variables is

$$\mathbf{P}\{\underline{z}_{it} | \underline{x}_i, \underline{\pi}_i, \mathbf{p}\} \quad \propto \quad \prod_{k=1}^{K} \left\{\pi_{ik} q_{kc(i,t)}\right\}^{z_{itk}}$$

$$\Rightarrow \underline{z}_{it} \sim \text{Multinomial}\left(1, \frac{\pi_{ik} q_{kc(i,t)}}{\sum_{k'=1}^{K} \pi_{ik'} q_{k'c(i,t)}}\right)$$

where the probabilities of the Multinomial distribution have been normalized. Also, for the GoM scores

$$\begin{aligned}\mathbf{P}\{\underline{\pi}_i|\mathbf{z}_i, \underline{x}_i, \mathbf{p}\} &\propto \prod_{k=1}^{K}\prod_{t=1}^{n_i} \pi_{ik}^{z_{itk}+\alpha_k-1} \\ &= \prod_{k=1}^{K} \pi_{ik}^{\sum_{t=1}^{n_i} z_{itk} + \alpha_k - 1}\end{aligned}$$

$$\Rightarrow \quad \underline{\pi}_i \sim \text{Dirichlet}(\alpha_1 + \sum_{t=1}^{n_i} z_{it1}, \ldots, \alpha_K + \sum_{t=1}^{n_i} z_{itK})$$

by Definition 8. In the case of the Plackett-Luce support parameters the complete conditional distribution is

$$\mathbf{P}\{\underline{p}_k|\pi, \mathbf{z}, \mathbf{x}\} \propto \left[\prod_{i=1}^{M}\prod_{t=1}^{n_i}\left\{\frac{\pi_{ik}p_{kc(i,t)}}{\sum_{s=t}^{N} p_{kc(i,s)}}\right\}^{z_{itk}}\right]\left[\prod_{j=1}^{N} p_{kj}^{\beta_j - 1}\right]. \qquad (6.3)$$

Due to the intricate form of the Plackett-Luce density, the full conditional distribution of the support parameters is not easily recognizable as a standard distribution and a straight forward Gibbs sampler algorithm cannot be fully implemented. Thus a hybrid algorithm is employed as detailed in the following section.

6.2.2 The Metropolis Within Gibbs Sampler

Different MCMC algorithms may be combined to draw on and accumulate their individual strengths. The Metropolis within Gibbs (or the 'variable-at-a-time Metropolis algorithm' (O'Hagan and Forster, 2004)) algorithm imbeds T Metropolis steps within an outer Gibbs sampling algorithm. Generally $T = 1$ is used which in effect simply substitutes a Metropolis step for a Gibbs step. Carlin and Louis (2000) discuss convergence issues associated with such a hybrid algorithm.

When the conditional distributions required for the Gibbs sampler are not in standard form (and techniques such as rejection sampling (O'Hagan and Forster, 2004) are inefficient) it is often better to resort to sampling from an alternative proposal distribution in a Metropolis-Hastings style step. To implement a Metropolis-Hastings step to sample Plackett-Luce support parameter values a proposal distribution which approximates the full conditional (6.3) is required.

One possibility examined was to approximate the complete conditional distribution using Rosén's approximation (Rosén, 1972) of the Plackett-Luce model (see Chapter 3.1)

$$\mathbf{P}\{\underline{x}_i|\underline{p}_k\} \approx p_{kc(i,1)}p_{kc(i,2)}\cdots p_{kc(i,n_i)}.$$

Under this approximation the conditional posterior distribution of $\underline{p}_k$ is recognizable as a Dirichlet distribution which would provide a proposal distribution which is straight forward to sample from. However, in electoral data sets N and n_i are typically small thus the conditions required for Rosén's approximation are generally not satisfied. An alternative method, motivated by the MM algorithm, for constructing suitable proposal distributions is detailed in the following section.

6.2.3 Surrogate Proposal Distributions

The MM algorithm (Lange et al., 2000; Hunter and Lange, 2004) is an optimization tool which operates by iteratively optimizing a surrogate function for a problematic objective function. Implementation of the MM algorithm is discussed in Chapter 4.3.1. Here the technique of constructing a surrogate function, and iteratively updating it, is borrowed to form a workable proposal distribution.

Taking logs of (6.3), the complete conditional of $\underline{p}_k$, gives

$$\log\mathbf{P}\{\underline{p}_k|\mathbf{z},\mathbf{x},\pi\} - C = \sum_{i=1}^{M}\sum_{t=1}^{n_i} z_{itk}\left\{\log p_{kc(i,t)} - \log\sum_{s=t}^{N} p_{kc(i,s)}\right\} + \sum_{j=1}^{N}(\beta_j - 1)\log p_{kj} \quad (6.4)$$

where C is a constant. The function $-\log(\theta)$ is a convex function and thus the supporting hyperplane property of convex functions (see (4.4)) can be applied to the problematic term $-\log\sum_{s=t}^{N} p_{kc(i,s)}$ in (6.4). This provides a linear minorizing surrogate function for the log of the complete conditional of $\underline{p}_k$. Thus by (4.4)

$$-\log\sum_{s=t}^{N} p_{kc(i,s)} \geq -\log\sum_{s=t}^{N} \bar{p}_{kc(i,s)} - \frac{\sum_{s=t}^{N} p_{kc(i,s)}}{\sum_{s=t}^{N} \bar{p}_{kc(i,s)}} + 1$$

where $\bar{p}_{kc(i,s)}$ is a constant value of the respective support parameter. Denoting

$$\delta_{kj} = \sum_{i=1}^{M}\sum_{t=1}^{n_i} z_{itk}\mathbf{1}_{\{c(i,t)=j\}} \qquad \text{where}$$

$$\mathbf{1}_{\{c(i,t)=j\}} = \begin{cases} 1 & \text{if } c(i,t) = j \\ 0 & \text{otherwise} \end{cases} \qquad \text{and}$$

$$\psi_{ijt} \;=\; \begin{cases} 1 & \text{if } t = 1 \\ 1 & \text{if } t > 1 \text{ and } c(i,1), \ldots, c(i,t-1) \neq j \\ 0 & \text{otherwise} \end{cases}$$

then (6.4) becomes

$$\begin{aligned}
&\log \mathbf{P}\{\underline{p}_k | \mathbf{z}, \mathbf{x}, \pi\} - C \\
&\geq \sum_{j=1}^{N} \delta_{kj} \log p_{kj} - \sum_{i=1}^{M} \sum_{t=1}^{n_i} \left\{ z_{itk} \left(\sum_{s=t}^{N} \bar{p}_{kc(i,s)} \right)^{-1} \left(\sum_{s=t}^{N} p_{kc(i,s)} \right) \right\} + \sum_{j=1}^{N} (\beta_j - 1) \log p_{kj} \\
&= \sum_{j=1}^{N} (\beta_j + \delta_{kj} - 1) \log p_{kj} - \sum_{i=1}^{M} \sum_{t=1}^{n_i} \left(\sum_{s=t}^{N} \bar{p}_{kc(i,s)} \right)^{-1} \left(\sum_{j=1}^{N} z_{itk} p_{kj} \psi_{ijt} \right). \qquad (6.5)
\end{aligned}$$

A definition of the standard Gamma distribution is necessary to recognize this surrogate proposal distribution.

Definition 10 *A **Gamma distributed** random variable x has density*

$$\mathbf{P}\{x\} = \frac{1}{\beta^{\alpha} \Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta)$$

where α is known as a shape parameter and β as a scale parameter. ▮

Thus by examining (6.5) and the log of the Gamma distribution it is clear that the Plackett-Luce support parameters are approximately Gamma distributed i.e.

$$p_{kj} \;\sim\; \text{Gamma} \left(\beta_j + \delta_{kj} \;, \quad \left[\sum_{i=1}^{M} \sum_{t=1}^{n_i} \left\{ \sum_{s=t}^{N} \bar{p}_{kc(i,s)} \right\}^{-1} z_{itk} \psi_{ijt} \right]^{-1} \right).$$

This Gamma function becomes computationally unstable and difficult to sample from as the shape parameter $\beta_j + \delta_{kj}$ is generally large — usually the number of voters M in an electoral data set is large which by definition increases the value of δ_{kj}. However, since

$$\text{Gamma}(r, \lambda) \;\rightarrow\; \text{Normal}(r\lambda,\; r\lambda^2) \;\; \text{as} \;\; r \rightarrow \infty$$

(see Casella and Berger (1990)) a Normal$(\mu_{kj}, \sigma^2_{kj})$ distribution is used as a workable proposal distribution when sampling the support parameters p_{kj} where

$$\mu_{kj} \;=\; \frac{\beta_j + \delta_{kj}}{\sum_{i=1}^{M} \sum_{t=1}^{n_i} \left\{ \sum_{s=t}^{N} \bar{p}_{kc(i,s)} \right\}^{-1} z_{itk} \psi_{ijt}} \qquad \sigma^2_{kj} = \frac{\beta_j + \delta_{kj}}{\left[\sum_{i=1}^{M} \sum_{t=1}^{n_i} \left\{ \sum_{s=t}^{N} \bar{p}_{kc(i,s)} \right\}^{-1} z_{itk} \psi_{ijt} \right]^2}$$

for $j = 1, \ldots, N$ and $k = 1, \ldots, K$.

Since the support parameters to be sampled are constrained such that $0 \leq p_{kj} \leq 1$ and $\sum_{j=1}^{N} p_{kj} = 1$ normalization must be performed during the Metropolis-Hastings step within the Gibbs sampler. Let $\underline{pu}_k$ denote the vector of un-normalized support parameters for extreme profile k; similarly let $\underline{pn}_k$ denote the vector of normalized support parameters within extreme profile k. Subsequent to choosing suitable starting values for the support parameters, the Metropolis-Hastings step of the Metropolis within Gibbs algorithm proceeds as follows for each $k = 1, \ldots, K$:

1. Generate N pu_{kj} values, where $\underline{pu}_k = (pu_{k1}, \ldots, pu_{kN}) \sim \mathrm{N}(\underline{\mu}_k, \underline{\sigma}_k^2)$.

2. Set $\underline{pn}_k = \underline{pu}_k/S = (pu_{k1}/S, \ldots, pu_{kN}/S)$ where $S = \displaystyle\sum_{j=1}^{N} pu_{kj}$.

3. Let $\underline{\bar{p}}_k$ denote the value of $\underline{p}_k$ from the previous iteration. Calculate the log of the acceptance ratio α for support parameters $\underline{p}_k$ where

$$\begin{aligned}
\log(\alpha) &= \min\left[\log\left\{\frac{\mathbf{P}(\underline{pn}_k|...)q(\underline{\bar{p}}_k|...)}{\mathbf{P}(\underline{\bar{p}}_k|...)q(\underline{pn}_k|...)}, 1\right\}\right] \\
&= \min\Bigg[\sum_{i=1}^{M}\sum_{t=1}^{n_i}\Bigg[z_{itk}\Bigg[\log pn_{kc(i,t)} - \log\Bigg\{\sum_{s=t}^{N} pn_{kc(i,s)}\Bigg\} - \log \bar{p}_{kc(i,t)} \\
&\quad +\log\Bigg\{\sum_{s=t}^{N} \bar{p}_{kc(i,s)}\Bigg\}\Bigg]\Bigg]\Bigg] + \sum_{j=1}^{N}\left[(\beta_j - 1)\left\{\log(pn_{kj}) - \log(\bar{p}_{kj})\right\}\right. \\
&\quad \left. + \frac{(pn_{kj} - \mu_{kj}/S)^2 - (\bar{p}_{kj} - \mu_{kj}/S)^2}{2\sigma^2_{kj}/S^2}\right], 0\Bigg]
\end{aligned}$$

where ... represents all other parameters and $q(.)$ the Normal proposal distribution.

4. Generate a uniform random variable $u \sim U(0, 1)$.

5. If $\log(u) \leq \log(\alpha)$ define $\underline{p}_k = \underline{pn}_k$ otherwise $\underline{p}_k = \underline{p}_k$.

Thus each time the Metropolis-Hastings step occurs the proposal distribution is updated to depend on the previous estimate of the support parameter which therefore provides a good approximation of the full conditional distribution.

6.3 Model Features

When sampling the parameters via MCMC algorithms some special features of the GoM model require attention. A fundamental issue in the fitting of any mixture based model within a Bayesian framework is that of label switching. Another obvious issue is inferring the correct number of extreme profiles present.

6.3.1 Label Switching

The GoM model likelihood is invariant under relabelling of the extreme profiles of the population. This phenomenon leads to posterior distributions which are multimodal or symmetric and thus estimating parameters by their posterior mean is inappropriate. Several approaches to alleviate this problem are detailed in the literature — Richardson and Green (1997) suggest minimizing label switching by imposing artificial identifiability constraints such as ordering the mixing proportions or other model parameters. The selection of the parameters on which to base the ordering and indeed selecting the ordering itself is somewhat ad hoc however. Relabelling strategies using a decision theoretic approach as proposed by Celeux et al. (2000) and Stephens (2000) are implemented here.

A decision theoretic approach involves defining a loss function $\mathcal{L}(\,\hat{\mathbf{p}};\mathbf{p})$ which quantifies the loss incurred by choosing $\hat{\mathbf{p}}$ when the true parameter value is $\mathbf{p}$. The aim is thus to minimize the posterior expected loss $\mathbf{E}\{\mathcal{L}(\,\hat{\mathbf{p}};\mathbf{p})|\mathbf{x}\}$. Due to the nature of the label switching problem it is intuitive to employ a loss function that is invariant under permutations of the parameters.

The support parameters $\mathbf{p}$ of the Plackett-Luce model are used to rectify the label switching issue since the data provides more information about the support parameters than about the GoM scores. Each mixed-membership vector is estimated only by a single vote whereas all the votes contribute to estimating the Plackett-Luce support parameters. The 'true' or reference value of the support parameters is set to be the *maximum a posteriori* estimate $\mathbf{p}^R$ obtained after a number of initial uphill only moves in the Metropolis-Hastings step of the algorithm. This MAP value is used as the reference to which each new estimate $\hat{\mathbf{p}}^t$ will be 'matched' to correct for any label switching that may occur during estimation. A sum of squares function is employed as the loss function to be minimized i.e.

$$\mathcal{L}(\ \hat{\mathbf{p}}^t; \mathbf{p}^R) \ = \ \sum_{k=1}^{K} \sum_{j=1}^{N} (\hat{p}^t_{kj} - p^R_{kj})^2.$$

Once the MAP estimate has been obtained and subsequent to a typical burn-in period of the Markov chain, the rows of the estimated matrix $\hat{\mathbf{p}}^t$ are permuted after each Metropolis-Hastings step until the loss function is minimized. An online algorithm (Stephens, 2000) which corrects for any label switching and calculates valid parameter posterior means then proceeds as follows:

1. Generate all $K!$ permutations ν_l for $l = 1, \ldots, K!$. Set $t = 0$.

2. After discarding the burn-in Metropolis-Hastings steps, denote the next estimated parameter set by $\hat{\mathbf{p}}^t$. Choose permutation ν_l for $l = 1, \ldots, K!$ that minimizes the loss function

$$\mathcal{L}(\hat{\mathbf{p}}^t_{\nu_l}; \mathbf{p}^R) = \sum_{k=1}^{K} \sum_{j=1}^{N} (\hat{p}^t_{\nu_l(k)j} - p^R_{kj})^2. \qquad (6.6)$$

3. Calculate the posterior mean support parameter as $p_{kj} = \frac{t}{t+1} p_{kj} + \frac{1}{t+1} \hat{p}^t_{\nu_l(k)j}$ for $k = 1, \ldots, K$, $j = 1, \ldots, N$ where ν_l is the permutation which when applied to the rows of the matrix $\hat{\mathbf{p}}^{\mathbf{t}}$ minimizes the loss function (6.6).

4. Similarly calculate the posterior mean GoM scores as $\pi_{ik} = \frac{t}{t+1} \pi_{ik} + \frac{1}{t+1} \hat{\pi}^t_{i\nu_l(k)}$ for $i = 1, \ldots, M$ and $k = 1, \ldots, K$. Set $t = t + 1$ and repeat steps $2 - 4$ subsequent to each Metropolis-Hastings step within the sampler.

It follows that label switching will be minimized ensuring the validity of posterior mean estimates.

6.3.2 Model Selection

Another feature of the GoM model is the selection of the value K, the number of extreme profiles present in the Irish electorate. Airoldi et al. (2006) discuss methods of model selection for GoM models. Erosheva (2002) fixed the value of K to be equal to two when fitting GoM models.

In this application of the GoM model two model selection criteria are examined. The Deviance Information Criterion (DIC) introduced by Spiegelhalter et al. (2002)

is used to provide a measure of model fit (see Chapter 3.5). It is defined as the Bayesian deviance evaluated at the parameter means plus twice the effective number of parameters

$$\mathrm{DIC} = D(\bar{\mathbf{p}}, \bar{\pi}) + 2p_D.$$

The DIC is a typical model selection criterion which rewards model fit but penalizes over parameterization. Models with smaller DIC values are preferable.

Pritchard et al. (2000) suggest a similar model selection criterion based on an approximation of the posterior distribution $\mathbf{P}\{D|\mathbf{x}\}$. It is computationally similar to the DIC but penalizes the mean of the Bayesian deviance by a quarter of it's variance as opposed to the effective number of parameters. Detailed discussion of these criteria is given in Chapter 3.5.

6.4 Application of the GoM Model to the Irish Electorate

The GoM model incorporating the Plackett-Luce model was applied to the preferences expressed in an exit poll conducted on the day of the 1997 presidential election (see Chapter 2). The Metropolis within Gibbs sampler was run over 50000 iterations, with a burn-in period of 10000 iterations, over the range $K = 1, \ldots, 5$ extreme profiles. Dirichlet priors with $\underline{\alpha} = (0.5, \ldots, 0.5)$ and $\underline{\beta} = (0.5, \ldots, 0.5)$ were imposed on the mixed membership variables and the support parameters respectively. A previous mixture modelling analysis of this electoral data set was reported in Chapter 5.

Figure 6.1 illustrates the model selection criterion values obtained when fitting the GoM models to the 1997 presidential exit poll data. Similar to results reported in Chapter 8 the DIC and Pritchard's criterion give different results for the number of extreme profiles present in the polled Irish electorate (the DIC suggests an electorate composed of four extreme profiles whereas Pritchard's criterion suggests three). Since a previous analysis of this data (see Chapter 5) suggested a four component mixture of Plackett-Luce models was appropriate for this data the four component GoM model is discussed. Also, as demonstrated in Figure 6.1 the values of Pritchard's criterion are relatively close for $K = 2, 3$ and 4.

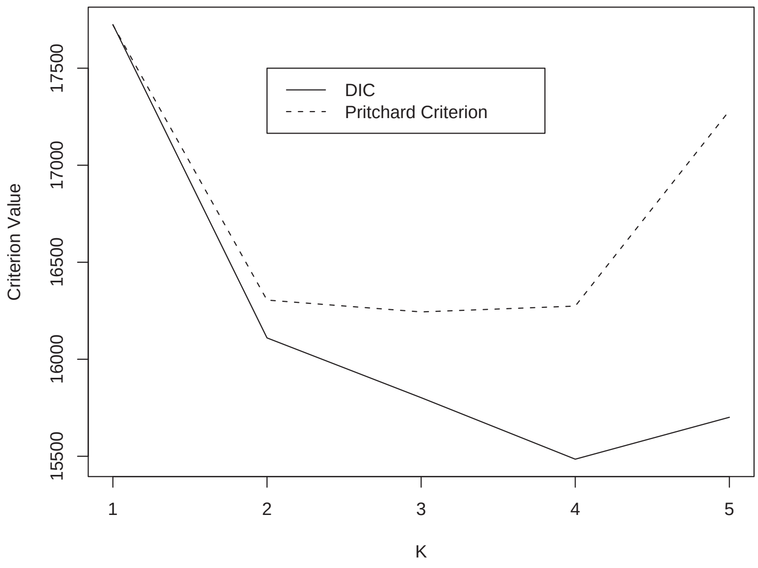


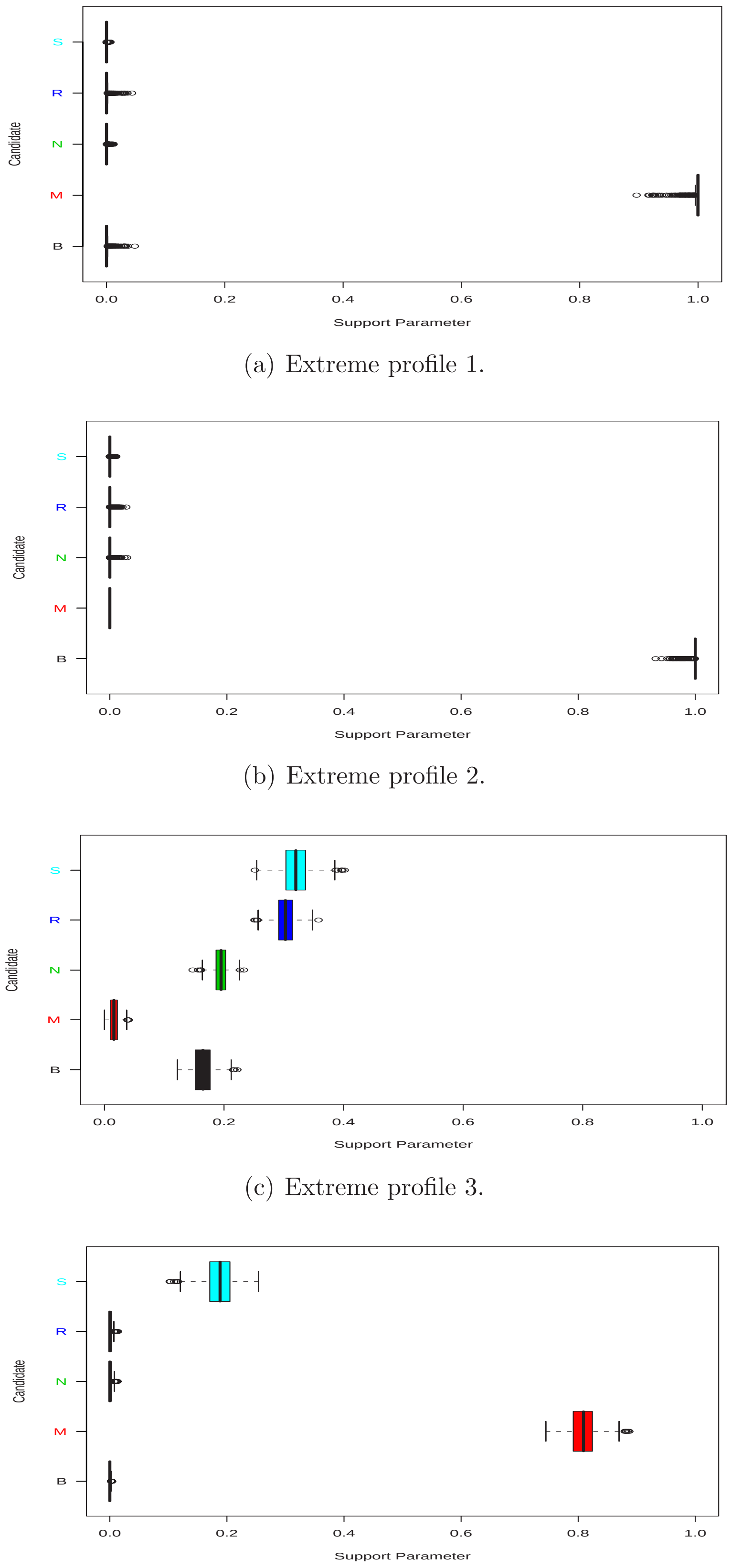
Fig. 6.1: Values of the DIC and Pritchard's criterion for the GoM model fitted over different values of the number of extreme profiles K to the 1997 exit poll data.

6.4.1 Support for the Presidential Candidates

The DIC suggests a four component model best fits the electorate polled (see Figure 6.1). Figure 6.2 illustrates the posterior mean support parameters and their associated uncertainty for each electoral candidate within the four extreme profiles. The five candidates were Banotti, McAleese, Nally, Roche and Scallon with McAleese winning the presidential seat. The four extreme profiles have distinct and intuitive interpretations within the context of the 1997 Irish presidential election. The uncertainty associated with the posterior means is relatively small throughout. Also, Figure 6.3 illustrates the trace plots for the support parameters estimated by the Markov chain.

Extreme profile one: pro-McAleese voters.

> (See Figure 6.2(a).) The posterior mean support parameter estimate for candidate McAleese within this group is 0.99 with small associated uncertainty. It follows therefore that within group one there is little or no support for the other candidates. This group models voters who strongly favor McAleese; Mary McAleese was elected as President of Ireland in the 1997 election. Banotti and Roche have the largest associated uncertainty of the other candidates and thus may have support parameters slightly larger than zero. Banotti was McAleese's closest challenger and although Roche was not a major challenger



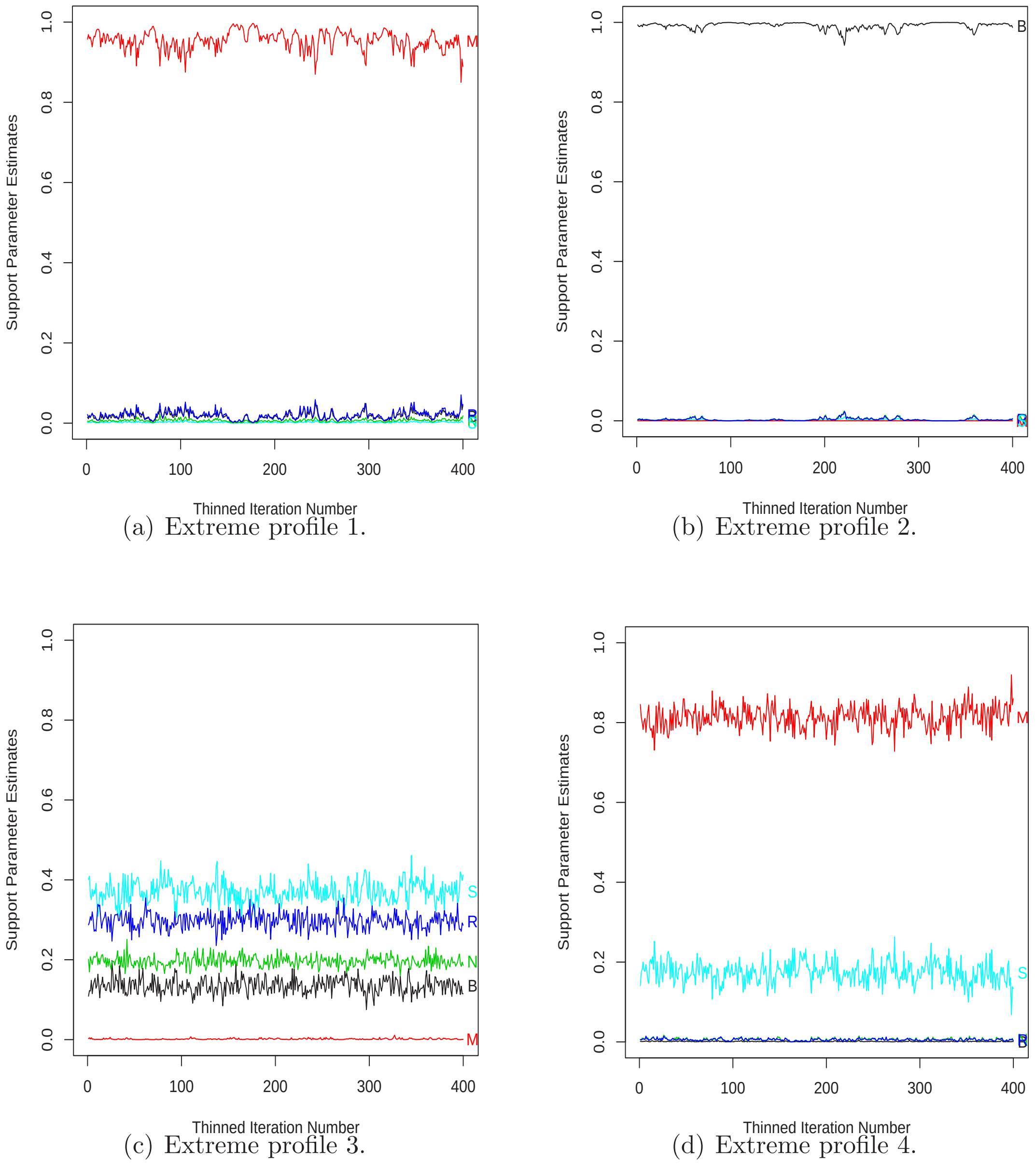
(a) Extreme profile 1.

(b) Extreme profile 2.

(c) Extreme profile 3.

(d) Extreme profile 4.

Fig. 6.2: Box and whisker plots of the posterior mean support parameter estimates, with their associated uncertainty, for each of the five electoral candidates within the four extreme profiles highlighted. Each candidate is denoted by their initial.



(a) Extreme profile 1.

(b) Extreme profile 2.

(c) Extreme profile 3.

(d) Extreme profile 4.

Fig. 6.3: Trace plots of samples of support parameters for the presidential candidates within each extreme profile obtained after convergence of the Markov chain. Each figure illustrates 40000 samples thinned every 100th iteration. The initial at the end of each trace indicates which candidate's support parameter is traced.

on polling day, she had maintained a large public profile throughout the campaign.

Extreme profile two: pro-Banotti voters.

(See Figure 6.2(b).) Banotti has high support in this group. While there is essentially zero support for McAleese the other candidates have some uncertainty around zero. Banotti supporters appear to dislike McAleese strongly, where McAleese supporters (extreme profile one) tend to be less extreme in their views of the other candidates.

Extreme profile three: anti-McAleese voters.

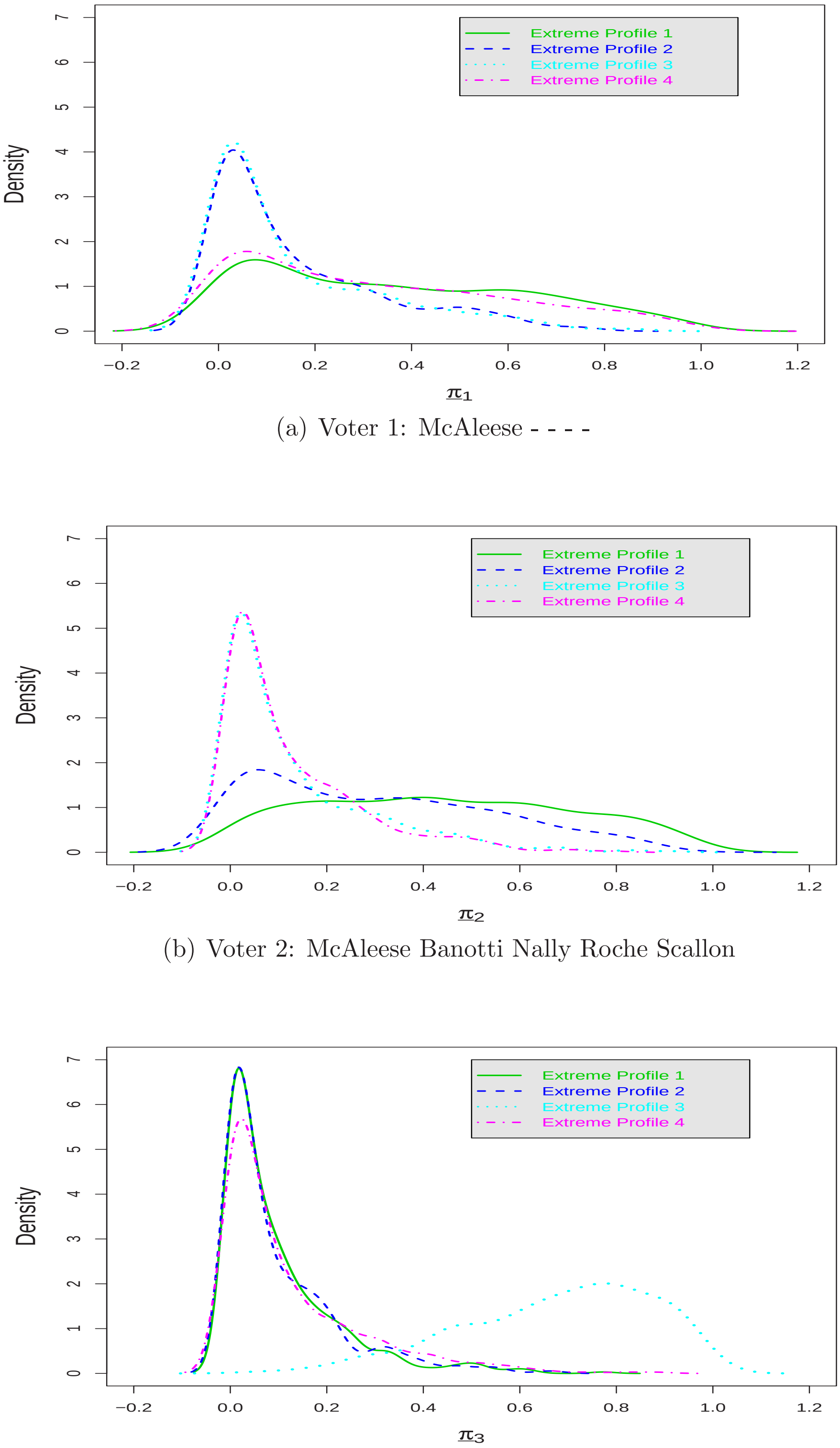
(See Figure 6.2(c).) With the exception of McAleese, each candidate has some level of support in this group. The candidates with the larger support parameters had smaller public profiles and were backed by smaller, if any, political parties. Chapter 2.2.1 details the political affiliations of the candidates. This extreme profile models voters who are generally in favor of any candidate except McAleese.

Extreme profile four: conservative voters.

(See Figure 6.2(d).) The final group models a conservative group of voters — McAleese and Scallon emerged as the more conservative candidates during the campaign and have the larger support parameters.

6.4.2 Mixed Membership Parameters for the Electorate

The unique feature of the GoM model is that the partial memberships of the extreme profiles for each voter are inferred directly when estimating the model. Figure 6.4 illustrates the kernel density estimates of the mixed membership realizations sampled during the Metropolis within Gibbs algorithm (subsequent to burn-in) for three randomly sampled voters. All have GoM scores which are interpretable within the context of the 1997 Irish presidential election.



(a) Voter 1: McAleese - - - -

(b) Voter 2: McAleese Banotti Nally Roche Scallon

(c) Voter 3: Scallon Nally Roche - -

Fig. 6.4: Density estimates of realizations (subsequent to burn-in) of the mixed membership parameter $\underline{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{i4})$ for three randomly sampled voters. The preferences expressed by each voter are detailed under each figure. The symbol - denotes the case where a voter chose not to express any further preferences. The four extreme profiles referred to are as reported in Figure 6.2.

Voter one.

(See Figure 6.4(a).) This voter, who only ranked McAleese, has larger probability of belonging to extreme profiles one (the pro-McAleese extreme profile) and four (the conservative extreme profile). The probability of this voter's membership of extreme profile three (the anti-McAleese extreme profile) is tightly distributed around zero. The posterior mean GoM score for voter one was $\underline{\pi}_1 = (0.36, 0.16, 0.16, 0.32)$. Thus 36% of this voter's voting behaviour can be characterized by extreme profile one and 32% of it by extreme profile four.

Voter two.

(See Figure 6.4(b).) This voter chose to express all five preferences and has larger probability of belonging to extreme profiles one (the pro-McAleese extreme profile) and extreme profile two (the pro-Banotti extreme profile). This is a natural assignment as the first two preferences expressed were McAleese and then Banotti. The probability of being assigned to either extreme profile three or four is small. This again makes intuitive sense as extreme profile three encapsulates the anti-McAleese voters, which clearly voter two is not. Also extreme profile four models the conservative voters who favor McAleese and Scallon. Since Scallon was ranked last by this voter it follows that their probability of belonging to the conservative extreme profile of voters should be small.

In the case of voter two the posterior mean GoM score was $\underline{\pi}_2 = (0.44, 0.31, 0.13, 0.12)$; 44% of voter two's behaviour can be characterized by the pro-McAleese extreme profile with 31% characterized by the pro-Banotti extreme profile. Also the uncertainty associated with voter two's membership of extreme profiles one and two is quite large which coincides with their ranking of McAleese first and Banotti second. Such high rankings of both candidates contradict membership of both extreme profiles hence introducing more uncertainty.

Voter three.

(See Figure 6.4(c).) Voter three chose not to rank either of the high profile candidates of McAleese or Banotti. Voter three has very high probability of belonging to the extreme profile of anti-McAleese voters and very small probability of membership of any of the alternative extreme profiles, all of which have support for McAleese and/or Banotti. The posterior mean GoM score of $\underline{\pi}_3 = (0.10, 0.10, 0.68, 0.12)$ further highlights how voter three is mainly characterized by the anti-McAleese extreme profile.

6.5 Conclusions

The GoM model incorporating the Plackett-Luce model for rank data is a suitable and necessary framework in which the structure of the Irish electorate may be examined. The GoM mixed membership parameters provide deeper insight to the mechanisms and opinions that drive each voter individually. Thus the loss of information which results from a hard clustering is reduced by the provision of a soft or fuzzy clustering of the heterogeneous electorate.

Run times for the implemented methodology were small. The use of a surrogate proposal distribution and the subsequent generalization of Gamma$(r, \lambda) \Rightarrow$ Normal$(r\lambda, r\lambda^2)$ appears to have worked well. Good mixing of the Markov chain has also been illustrated.

Further model accuracy could be attained by imposing a hierarchical framework — a hyperprior could be introduced for the Dirichlet parameters $\underline{\alpha}$ and $\underline{\beta}$ of the mixed membership and support parameter priors respectively. Erosheva (2003) employed such hierarchical priors.

Chapter 7

A Mixtures-of-Experts Model for Rank Data

Mixtures-of-expert (ME) models (Jacobs et al., 1991) combine the ideas of both mixture models (see Chapter 4.1) and generalized linear models (see Chapter 7.2). Complex problems can be formulated as a mixture model where generalized linear model theory provides the statistical structure within the mixture.

ME models build further on the structure implemented by mixture models by taking account of both the observations and associated covariates when modelling a heterogeneous population. Specifically in this chapter the heterogeneous population is a subset of the Irish electorate; both the votes cast and the covariates associated with the voters are modelled. In similar vein to Chapter 5 the aim is to perform an exploratory analysis of the Irish electorate to determine which social factors influence voting patterns and what the induced voting patterns are. In particular the IMS presidential opinion poll conducted on October 2nd 1997 (see Chapter 2) is analyzed.

7.1 Mixtures-of-Experts Models

Mixtures-of-experts models model the relationship between a set of response and covariate variables where they assume that the conditional distribution of the responses given the covariates is a finite mixture distribution. The components of the finite mixture distribution are known as the ‘expert networks’.

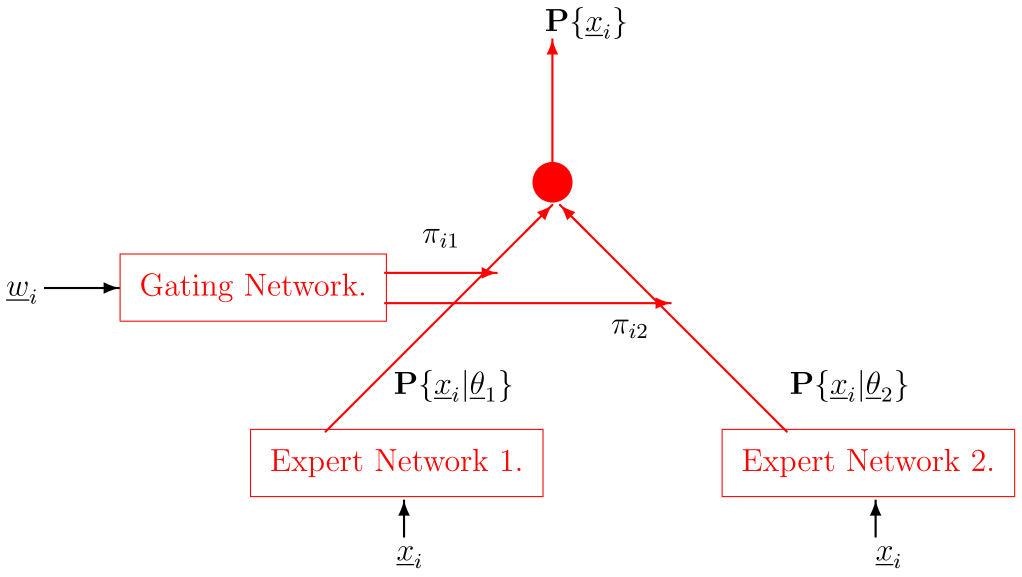

$\mathbf{P}\{\underline{x}_i\}$
π_{i1}
$\underline{w}_i$
Gating Network.
π_{i2}
$\mathbf{P}\{\underline{x}_i|\underline{\theta}_1\}$
$\mathbf{P}\{\underline{x}_i|\underline{\theta}_2\}$
Expert Network 1.
Expert Network 2.
$\underline{x}_i$
$\underline{x}_i$

Fig. 7.1: Tree like structure of a single layer mixtures-of-experts model with two expert networks.

ME models have a tree like structure where the expert networks are the leaves and 'gating networks' form the non-terminal nodes of the tree (see Figure 7.1). A hierarchical mixtures-of-experts model (HME) has the same structure but has multiple layers of expert and gating networks. The gating network parameter π_{ik} represents the probability of voter i being a complete member of expert network k given voter i's associated covariates $\underline{w}_i$. The gating network parameters are weighting probabilities constrained such that they are nonnegative and sum to one for each voter. The probability of voter i's ballot $\mathbf{P}\{\underline{x}_i|\underline{\theta}_k\}$ according to the expert networks in the mixture model are then blended by the gating network parameters to produce an overall probability of voter i's ballot. Thus $\mathbf{P}\{\underline{x}_i\}$ is a convex combination of the outputs from the expert networks.

Jordan and Jacobs (1994) assume each component of the mixture model (i.e. each expert network) produces its output as a generalized linear function of input predictor variables. Within in the context of rank Irish voting data it is assumed each of K expert networks follows a Benter model distribution (see Chapter 3.2) with different parameterizations; here the parameterization is constant with respect to the covariates. It is possible to allow covariates contribute to the expert networks (see Jordan and Jacobs (1994) and Peng et al. (1996)) but this case is not examined here.

Benter's model for expert network k is parameterized by $\underline{\theta}_k = (\underline{p}_k, \underline{\alpha})$ where $\underline{p}_k = (p_{k1}, \ldots, p_{kN})$ is the set of support parameters for each candidate $j = 1, \ldots, N$ and $\underline{\alpha} = (\alpha_1, \ldots, \alpha_N)$ is the vector of dampening parameters associated with the N choice levels. As usual $\sum_{j=1}^{N} p_{kj} = 1$ for $k = 1, \ldots, K$. The dampening parameters are constrained such that $0 \leq \alpha_t \leq 1$ for $t = 1, \ldots, N$ and $\alpha_1 = 1$ for identifiability reasons. Under Benter's model the probability of vote $\underline{x}_i$ given that voter i is completely characterized by expert network k is

$$\mathbf{P}\{\underline{x}_i|\underline{p}_k, \underline{\alpha}\} = \prod_{t=1}^{n_i} \frac{p_{kc(i,t)}^{\alpha_t}}{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}}$$

where n_i is the number of preferences expressed by voter i and $c(i,t)$ denotes the candidate ranked at the tth choice level by voter i. Under the ME model then

$$\mathbf{P}\{\underline{x}_i|\mathbf{p}, \underline{\alpha}, \underline{w}_i\} = \sum_{k=1}^{K} \underset{\substack{\uparrow \\ k\text{th gating network.}}}{\pi_{ik}(\underline{w}_i)} \underset{\substack{\uparrow \\ k\text{th expert network.}}}{\mathbf{P}\{\underline{x}_i|\underline{p}_k, \underline{\alpha}\}} \tag{7.1}$$

where $\mathbf{p} = (\underline{p}_1, \ldots, \underline{p}_K)$.

In case (7.1) as in Figure 7.1 the covariates of voter i, $\underline{w}_i$, are involved only in the gating networks through the use of generalized linear models; this methodology is discussed in Chapter 7.2.

7.2 Generalized Linear Models

Generalized linear models (GLMs) model the relationship between the mean of a response variable x and an independent or predictor variable w. Dobson (2002) and McCullagh and Nelder (1983) deal comprehensively with GLMs. An integral element of generalized linear models is the idea of the family of exponential distributions.

Definition 11 *A probability density function $f(\underline{x}|\underline{\theta})$ given a parameter $\underline{\theta}$ is said to belong to the* ***exponential family of distributions*** *if it can be written in the form*

$$\exp\{a(\underline{x}) + b(\underline{\theta}) + c(\underline{x})d(\underline{\theta}^T)\}$$

where a, b, c, d are known functions. If $c(\underline{x}) = \underline{x}$, then the distribution is said to be in ***canonical form****. When the distribution is in canonical form, the function $d(\underline{\theta}^T)$ is called a* ***natural parameter****.* ▮

The Gaussian and Poisson distributions are typical continuous distributions which are members of the exponential family — the binomial and geometric distributions are examples of discrete members.

A GLM has three constituent parts:

1. **the random component.** The response variable x forms the 'random' component of a GLM. It is assumed that the response variables are independent and that their distributions come from the exponential family. The response variables need not be identically distributed but they do have a distribution from the same family.

2. **the systematic component.** The function f of the predictor variable w which is linear in the parameters and is related to the mean of the response variable.

3. **the link function.** This links the random and systematic components by defining $g(\mu) = f(w)$ where $\mu = \mathbf{E}[x]$. The identity function, the logit function and the complementary log log function are all typical link functions (see McCullagh and Nelder (1983)).

The well known logistic regression model is a GLM — the responses x are independently distributed and $x \sim \text{Bernoulli}(\pi)$. Under a logistic regression model $\pi = \mathbf{P}\{x = 1\}$ is related to the predictor variable w by a logit link function i.e.

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta w.$$

The model assumes the log-odds of a 'success' for x is a linear function of the predictors w. Thus α is the log-odds of success at $w = 0$ and β is the change in the log-odds corresponding to a one unit increase in w.

Essentially the gating network probabilities in the mixtures-of-experts model are the success probabilities from a multinomial logistic regression where the probability of belonging to each of $K-1$ expert networks compared to a 'baseline category' is a function of the covariates. Voter i's gating network probabilities $\underline{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$ are modelled by a logistic function of their L covariates $\underline{w}_i = (w_{i1}, \ldots, w_{iL})$ i.e.

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \beta_{k0} + \beta_{k1}w_{i1} + \beta_{k2}w_{i2} + \cdots + \beta_{kL}w_{iL} \tag{7.2}$$

where expert network 1 is used as the baseline category and β_{k0} is an intercept term.

7.3 Fitting a ME Model.

Under the ME model (7.1) the likelihood of the covariates $\mathbf{w} = (\underline{w}_1, \ldots, \underline{w}_M)$ and the votes $\mathbf{x} = (\underline{x}_1, \ldots, \underline{x}_M)$ is

$$\mathcal{L}(\beta, \mathbf{p}, \underline{\alpha}|\mathbf{w}, \mathbf{x}) = \prod_{i=1}^{M}\sum_{k=1}^{K} \pi_{ik}(\underline{w}_i)\mathbf{P}\{\underline{x}_i|\underline{p}_k, \underline{\alpha}\} \tag{7.3}$$

where $\beta = (\underline{\beta}_1, \ldots, \underline{\beta}_K)$.

Similar to Jordan and Jacobs (1994) a maximum likelihood approach (see Chapter 3) via the EM algorithm is taken when estimating this model. Peng et al. (1996) introduce a method of estimating the ME model within a Bayesian framework where Gibbs sampling and the EM algorithm are used as training methods.

The EM algorithm is often used to produce parameter estimates when missing data is a feature of the problem or when optimization of the likelihood would be simplified if an additional set of variables were known. As it is difficult to directly maximize the likelihood (7.3) in this case the data is augmented by imputing latent variables which record the membership of the expert networks for each voter i.e. for each voter $i = 1, \ldots, M$

$$z_{ik} = \begin{cases} 1 & \text{if } i \text{ is characterized by expert network } k \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, \ldots, K$ expert networks. Thus the complete data likelihood is

$$\begin{aligned} \mathcal{L}_c(\beta, \mathbf{p}, \underline{\alpha}|\mathbf{w}, \mathbf{x}) &= \prod_{i=1}^{M}\prod_{k=1}^{K} \Big[\pi_{ik}(\underline{w}_i)\mathbf{P}\{\underline{x}_i|\underline{p}_k, \underline{\alpha}\}\Big]^{z_{ik}} \\ \Rightarrow \quad l_c(\beta, \mathbf{p}, \underline{\alpha}|\mathbf{w}, \mathbf{x}) &= \sum_{i=1}^{M}\sum_{k=1}^{K} z_{ik}\log\pi_{ik}(\underline{w}_i) + \sum_{i=1}^{M}\sum_{k=1}^{K} z_{ik}\log\mathbf{P}\{\underline{x}_i|\underline{p}_k, \underline{\alpha}\} \quad (7.4) \end{aligned}$$

where l_c denotes the log of the complete likelihood.

Maximum likelihood estimates of $\hat{\beta}$, $\hat{\mathbf{p}}$ and $\underline{\hat{\alpha}}$ are achieved via the EM algorithm. This is a two step algorithm consisting of an E (estimation) step and a M (maximization) step. The E step involves the estimation of the expected value of the latent variables and the M step involves the maximization of the likelihood which is updated subsequent to each E step. These steps are continuously iterated until convergence to stable parameter estimates or to a pre-specified criterion has been reached.

7.3.1 The E Step

The E step of the EM algorithm takes the expectation of the complete data log likelihood (7.4). Practically this translates to estimating the expected value of the missing variables. Since

$$\begin{aligned}
\mathbf{P}\{\underline{z}_i|\underline{w}_i, \underline{x}_i\} &= \frac{\mathbf{P}\{\underline{z}_i, \underline{x}_i|\underline{w}_i\}}{\mathbf{P}\{\underline{x}_i|\underline{w}_i\}} \\
&= \frac{\prod_{k=1}^{K} \left\{ \left(\prod_{t=1}^{n_i} \frac{p_{kc(i,t)}^{\alpha_t}}{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}} \right) \pi_{ik} \right\}^{z_{ik}}}{\sum_{k'=1}^{K} \left(\prod_{t=1}^{n_i} \frac{p_{k'c(i,t)}^{\alpha_t}}{\sum_{s=t}^{N} p_{k'c(i,s)}^{\alpha_t}} \right) \pi_{ik'}} \\
&= \prod_{k=1}^{K} \left\{ \frac{\left(\prod_{t=1}^{n_i} \frac{p_{kc(i,t)}^{\alpha_t}}{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}} \right) \pi_{ik}}{\sum_{k'=1}^{K} \left(\prod_{t=1}^{n_i} \frac{p_{k'c(i,t)}^{\alpha_t}}{\sum_{s=t}^{N} p_{k'c(i,s)}^{\alpha_t}} \right) \pi_{ik'}} \right\}^{z_{ik}} \\
&= \prod_{k=1}^{K} \phi_{ik}^{z_{ik}}
\end{aligned}$$

where $\sum_{k=1}^{K} \phi_{ik} = 1$ then $\underline{z}_i \sim \text{Multinomial}(1, \underline{\phi}_i)$ and $\mathbf{E}[z_{ik}] = \hat{z}_{ik} = \phi_{ik}$. Substituting these updated expected values into the complete data log likelihood forms the 'Q function'

$$Q = \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \log \pi_{ik}(\underline{w}_i) + \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \log \mathbf{P}\{\underline{x}_i|\underline{p}_k, \underline{\alpha}\} \qquad (7.5)$$

which is maximized with respect to the model parameters during the M step of the algorithm.

7.3.2 The M Step

From (7.2) it follows that the gating network probabilities are defined by

$$\pi_{ik}(\underline{w}_i) \quad = \quad \frac{\exp\left\{\underline{\beta}_k^T \underline{w}_i\right\}}{\sum_{k'=1}^{K} \exp\left\{\underline{\beta}_{k'}^T \underline{w}_i\right\}}$$

where $\underline{\beta}_k = (\beta_{k0}, \ldots, \beta_{kL})$. Explicitly the Q function becomes

$$\begin{aligned} Q \quad &= \quad \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \left[\underline{\beta}_k^T \underline{w}_i - \log \left\{ \sum_{k'=1}^{K} \exp\left(\underline{\beta}_{k'}^T \underline{w}_i \right) \right\} \right] \\ & \quad + \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{t=1}^{n_i} \hat{z}_{ik} \left[\alpha_t \log p_{kc(i,t)} - \log \sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t} \right]. \end{aligned} \tag{7.6}$$

Iterative maximization of Q provides MLE's of the model parameters. Since the gating network parameters (β) and the expert network model parameters ($\mathbf{p}$, $\underline{\alpha}$) influence the Q function through distinct terms the M step reduces to separate maximization problems for each parameter set. Moreover the EM algorithm performed is in fact an ECM (expectation and conditional maximization) algorithm where the M step is replaced by a conditional maximization step (Meng and Rubin, 1993). Thus maximizing (7.6) with respect to the Benter model parameters $\underline{p}_k$ and $\underline{\alpha}$ for each expert network makes use of the same theory introduced when fitting straight forward mixtures of Benter's models (Chapter 5.3.2) where the MM algorithm was employed to overcome some maximization issues. Thus

$$\hat{p}_{kj} \quad = \quad \frac{\omega_{kj}}{\displaystyle\sum_{i=1}^{M} \sum_{t=1}^{n_i} \hat{z}_{ik} \left\{ \sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \right\}^{-1} \left\{ \sum_{s=t}^{(N+1)} \bar{\alpha}_t \bar{p}_{kj}^{\bar{\alpha}_t - 1} \delta_{ijs} \right\}} \tag{7.7}$$

where ω_{kj}, $\bar{p}_{kj}$, $\bar{\alpha}_t$ and δ_{ijs} are are as defined in Chapter 5.3.3. Further, Chapter 5.3.4 details the steps involved in producing the the dampening parameter estimate

$$\hat{\alpha}_t = \frac{\displaystyle\sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} \hat{z}_{ik} \left[\log \bar{p}_{kc(i,t)} + \left(\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \right)^{-1} \left\{ \sum_{s=t}^{N} - \log \bar{p}_{kc(i,s)} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} + \bar{\alpha}_t (\log \bar{p}_{kc(i,s)})^2 \right\} \right] \right\} .\mathbf{1}_{\{t \leq n_i\}}}{\displaystyle\sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} \hat{z}_{ik} \left(\sum_{s=t}^{N} \bar{p}_{kc(i,s)}^{\bar{\alpha}_t} \right)^{-1} \sum_{s=t}^{N} (\log \bar{p}_{kc(i,s)})^2 \right\} .\mathbf{1}_{\{t \leq n_i\}}} . \tag{7.8}$$

7.3.3 Estimation of Gating Network Parameters

Maximization of the Q function (7.6) with respect to the gating network parameters β_{kl} for $k = 0, \ldots, K$ and $l = 0, \ldots, L$ is not straight forward. Theory from the MM algorithm again provides a tenable technique to maximize Q. Hunter and Lange (2004) outlined common inequalities used to construct majorizing or minorizing surrogate functions. As previously employed in Chapter 5.3.4, if $f(\theta)$ is a concave function which is twice differentiable and has bounded curvature then $f(\theta)$ can be minorized by a quadratic function with sufficiently high curvature and tangent to $f(\theta)$ at the point θ^n. For a negative definite matrix $\mathbf{B}$ such that $f''(\theta^n) - \mathbf{B} > 0$ then the inequality

$$f(\theta) \ \geq \ f(\theta^n) + f'(\theta^n)^T(\theta - \theta^n) + 1/2(\theta - \theta^n)^T\mathbf{B}(\theta - \theta^n)$$

provides a quadratic lower bound for $f(\theta)$. Hunter and Lange (2004) made use of this property within in a similar logistic regression framework.

For some constant C

$$\begin{aligned} q(\beta) = Q(\beta) + C \ &= \ \sum_{i=1}^{M}\sum_{k=1}^{K} \hat{z}_{ik}\left[\underline{\beta}_k^T\underline{w}_i - \log\left\{\sum_{k'=1}^{K}\exp\left(\underline{\beta}_{k'}^T\underline{w}_i\right)\right\}\right] \\ &= \ \sum_{i=1}^{M}\sum_{k=1}^{K} \hat{z}_{ik}(\underline{\beta}_k^T\underline{w}_i) - \sum_{i=1}^{M}\log\left\{\sum_{k'=1}^{K}\exp\left(\underline{\beta}_{k'}^T\underline{w}_i\right)\right\} \qquad (7.9) \end{aligned}$$

since, by definition, $\sum_{k=1}^{K} z_{ik} = 1$. Straight forward differentiation gives

$$S_{sr} = \frac{\partial q(\beta)}{\partial \beta_{sr}} \ = \ \sum_{i=1}^{M} w_{ir}(\hat{z}_{is} - \pi_{is}).$$

Further

$$\frac{\partial^2 q(\beta)}{\partial \beta_{sr}^2} \ = \ -\sum_{i=1}^{M} w_{ir}w_{ir}\pi_{is}\left[1 - \pi_{is}\right]$$

and

$$\frac{\partial^2 q(\beta)}{\partial \beta_{sr}\beta_{st}} \ = \ -\sum_{i=1}^{M} w_{ir}w_{it}\pi_{is}\left[1 - \pi_{is}\right].$$

Thus (7.9) is concave and since $\pi_{is}\left[1 - \pi_{is}\right]$ is bounded above by 1/4 a negative definite matrix $\mathbf{B}$ can be defined as $\mathbf{B} = -1/4\sum_{i=1}^{M}\underline{w}_i\underline{w}_i^T$. It follows that $q''(\beta^n) - \mathbf{B}$ is nonnegative definite and the quadratic function

$$g(\beta|\beta^n) = q(\beta^n) + \mathbf{S}(\beta^n)^T(\beta - \beta^n) + 1/2(\beta - \beta^n)^T\mathbf{B}(\beta - \beta^n)$$

minorizes $q(\beta)$ at the point β^n where $\mathbf{S} = (\underline{S}_1, \ldots, \underline{S}_K)$ and $\underline{S}_k = (S_{k0}, \ldots, S_{kL})$. Maximizing this with respect to the gating network parameters gives the iterative update formula

$$\beta^{n+1} = \beta^n - \mathbf{B}^{-1}\mathbf{S}(\beta^n) \tag{7.10}$$

which (as $\mathbf{B}$ is constant) therefore only requires the inversion of $\mathbf{B}$ once during the iterative algorithm.

This updating formula for the gating network parameters coincides with the widely known Newton-Raphson update

$$\beta^{n+1} \;=\; \beta^n - (\mathbf{H}^{-1})^n S^n.$$

where the traditional score functions S^n and Hessian $\mathbf{H}$ of the log-likelihood are replaced by the gradient vectors (evaluated at the current parameter estimates) and the Hessian matrix of the surrogate function. However the Newton-Raphson update requires matrix inversion at every iteration which is computationally expensive and is neatly avoided by using the MM update.

In the context of EM algorithms, Dempster et al. (1977) term an algorithm which increases the value of the complete data log likelihood without actually maximizing it a generalized EM (GEM) algorithm. Similarly, the technique detailed here to obtain maximum likelihood estimates of the gating network parameters is known as a gradient MM algorithm.

7.3.4 The EM/MM Algorithm

In summary, to obtain maximum likelihood estimates of the mixtures-of-experts model parameters the steps of the EM algorithm with the MM algorithm embedded at the M step stage proceed as follows:

0. **Initialize:** Choose starting values for $\mathbf{p}^{(0)} = (\underline{p}_1^{(0)}, \ldots, \underline{p}_K^{(0)})$, $\underline{\alpha}^{(0)} = (\alpha_1^{(0)}, \ldots, \alpha_N^{(0)})$ and $\beta^{(0)} = (\underline{\beta}_1^{(0)}, \ldots, \underline{\beta}_K^{(0)})$. Suggestions for good starting values are discussed in Chapter 7.5. Let $l = 0$.

1. **E step:** Compute the values

$$\hat{z}_{ik} = \frac{\pi_{ik}^{(l)} \prod_{t=1}^{n_i} \frac{p_{kc(i,t)}^{(l)}}{\sum_{s=t}^{N} p_{kc(i,s)}^{(l)}}}{\sum_{k'=1}^{K} \pi_{ik'}^{(l)} \prod_{t=1}^{n_i} \frac{p_{k'c(i,t)}^{(l)}}{\sum_{s=t}^{N} p_{k'c(i,s)}^{(l)}}}.$$

2. **M step:**

   From (7.7) calculate $p_{kj}^{(l+1)}$ for $k = 1, \ldots, K$ and $j = 1, \ldots, N$.

   From (7.8) calculate $\alpha_t^{(l+1)}$ for $t = 1, \ldots, N$.

   From (7.10) calculate $\beta_{kl}^{(l+1)}$ for $k = 2, \ldots, K$ and $l = 0, \ldots, L$.

   Increment l by 1.

3. **Convergence:** Repeat the E step and M step until convergence (as deemed by Aitken's acceleration criterion (see Chapter 3.3.1). The final parameter values are the maximum likelihood estimates $\hat{\mathbf{p}}$, $\underline{\hat{\alpha}}$ and $\hat{\beta}$.

7.4 Standard Errors for ME Parameters

Standard errors of the parameter estimates are not a natural by-product of the EM algorithm but they can be readily produced subsequent to convergence. Hunter and Lange (2004) discuss some approaches to obtaining standard errors within the context of the MM algorithm. The approach suggested by McLachlan and Krishnan (1997) and McLachlan and Peel (2000) as taken in Chapters 4.4 and 5.4 is taken here.

Following the theory outlined in Chapter 4.4 the covariance matrix of the estimated model parameters can be approximated by the inverse of the empirical information matrix $\mathcal{I}_e(\theta)$. Given the complete data log likelihood for voter i under the ME model

$$\begin{aligned} l_{ci}(\theta) &= \sum_{k=1}^{K} \hat{z}_{ik} \left[\underline{\beta}_k^T \underline{w}_i - \log \left\{ \sum_{k'=1}^{K} \exp\left(\underline{\beta}_{k'}^T \underline{w}_i \right) \right\} \right] \\ &\quad + \sum_{k=1}^{K} \sum_{t=1}^{n_i} \hat{z}_{ik} \left[\alpha_t \log p_{kc(i,t)} - \log \sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t} \right] \end{aligned}$$

the inverse empirical information matrix is

$$\mathcal{I}_e^{-1}(\theta) = \mathbf{S}^T \mathbf{S} = \sum_{i=1}^{M} s(\underline{x}_i|\theta) s^T(\underline{x}_i|\theta)$$

where

$$\begin{aligned} \theta &= \text{the model parameters } \hat{\mathbf{p}}, \underline{\hat{\alpha}}, \hat{\beta} \\ s(\underline{x}_i|\theta) &= \frac{\partial l_{ci}(\theta)}{\partial \theta}. \end{aligned}$$

Thus on convergence of the EM/MM algorithm approximate standard errors of the parameter estimates can be produced.

7.4.1 Score Function for Benter Support Parameters

Due to the distinct terms of parameters in the complete data log likelihood the standard errors for Benter's model parameters are as detailed in Chapter 5.3.3. The score function for the Benter support parameters for voter i is

$$\begin{aligned} s(\underline{x}_i|\mathbf{p}) &= \frac{\partial l_{ci}(\mathbf{p})}{\partial p_{kj}} \\ &= \hat{z}_{ik}\left[\sum_{t=1}^{n_i}\left\{\frac{\alpha_t \mathbf{1}_{\{j=c(i,t)\}}}{p_{kj}} - \frac{\sum_{s=t}^{N+1}\alpha_t p_{kj}^{\alpha_t-1}\delta_{ijs}}{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}}\right\}\right]. \end{aligned}$$

where $\mathbf{1}_{\{j=c(i,t)\}}$ and δ_{ijs} are as defined in Chapter 5.3.3.

7.4.2 Score Function for Dampening Parameters

Likewise for the Benter dampening parameters the relevant score function for voter i is

$$\begin{aligned} s(\underline{x}_i|\underline{\alpha}) &= \frac{\partial l_{ci}(\underline{\alpha})}{\partial \alpha_t} \\ &= \sum_{k=1}^{K}\hat{z}_{ik}\left[\mathbf{1}_{\{t\leq n_i\}}\left\{\log p_{kc(i,t)} - \frac{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}\log p_{kc(i,s)}}{\sum_{s=t}^{N} p_{kc(i,s)}^{\alpha_t}}\right\}\right]. \end{aligned}$$

7.4.3 Score Function for Gating Network Parameters

To estimate the standard errors associated with the gating network parameters the score function is

$$\begin{aligned} s(\underline{x}_i|\beta) &= \frac{\partial l_{ci}(\beta)}{\partial \beta_{kl}} \\ &= w_{il}(\hat{z}_{ik} - \pi_{ik}). \end{aligned}$$

Thus by calculating the square root of the diagonal elements of $\mathcal{I}_e^{-1}(\hat{\theta})$ formed from the relevant score functions the approximate standard errors for the ME model parameters can be provided.

7.5 Application of ME Models to Irish Presidential Opinion Poll Data

During the Irish presidential campaign of 1997 several opinion polls were conducted in which both the current preferences of the voters and some covariates were recorded. Chapter 2.2.1 provides details of these polls. The second opinion poll conducted by Irish Marketing Surveys (IMS) on October 2nd 1997 is analyzed here — the set of covariates recorded is the largest among the opinion polls conducted, the covariates recorded were deemed to be potentially the most informative and there was little missing data among the polled voters. Table 7.1 provides details of the covariates recorded by the IMS pollers. When fitting the ME model all covariates were standardized such that $0 \leq w_{il} \leq 1$ where w_{il} denotes the value of the lth covariate for voter i. Such standardization was employed to simplify interpretation of fitted parameters and to achieve numerical stability.

A single layer ME model rather than a hierarchical model was assumed to be applicable in this context. The nesting of choice levels and the partitioning of the available choices is an area of future research. Within a single layer ME model however the number of expert networks necessary to adequately summarize the data needs to be estimated. Jordan and Jacobs (1994) used a test set approach to model selection — training was stopped when the error on the test set reached a minimum. Peng et al. (1996) fixed the number of layers in their HME models as well as the number of expert networks. The Bayesian Information Criterion (BIC) (Chapter 3) is utilized here to select the optimal number of experts K. The BIC rewards well fitting models but penalizes over parameterization within them.

To estimate the ME model parameters for the IMS poll data a straight forward mixture of Benter models (see Chapter 5) was initially fitted to the preferences expressed by the polled voters. The EM/MM algorithm detailed in Chapter 5.3.5 was run for 500 iterations to provide good starting values for the Benter support parameters, dampening parameters and missing membership labels z_{ik} for $i = 1, \ldots, M$ and $k = 1, \ldots, K$. Good starting values for the gating network parameters β were obtained by then performing 1000 of the logistic regression style M steps (7.10) from the mixtures-of-experts EM/MM algorithm. The full EM/MM algorithm to

Table 7.1: The six covariates and their respective levels (if categorical variables) recorded in the October 2nd 1997 opinion poll conducted by Irish Marketing Surveys prior to the Irish presidential election.

Covariate	**Levels**
Age	—
Area	City
	Town
	Rural
Gender	Housewife
	Non-housewife
	Male
Government satisfaction level	Satisfied
	Dissatisfied
	Do not know/no opinion
Marital status	Married/living as married
	Single
	Widowed/divorced/separated
Social class	AB (Upper middle class & middle class)
	C1 (Lower middle class)
	C2 (Skilled working class)
	DE (Other working class & lowest level of subsistence)
	F50+ (Large farmers)
	F50– (Small farmers)

provide the maximum likelihood estimates of the ME model parameters was then iterated until convergence was achieved as deemed by Aitken's acceleration criterion (see Chapter 3.3.1). Subsequent to convergence approximate standard errors of the maximum likelihood estimates were calculated.

The ME model was fitted over the range $K = 2, \ldots, 5$ expert networks using a backward elimination style method to choose the informative covariates. Interaction terms were avoided. A model with all six covariates was initially fitted, then models with only five of the covariates. From this set of models the 'best' model as deemed by the BIC was selected and models with only four of the selected covariates were then fitted. This selection of the best subset of covariates and then backward elimination was continued until only one covariate was left in the model. The BIC values for all the models fitted were then compared. Table 7.2 details the five best fitting models as deemed by their BIC values. The covariates selected with each model are also detailed.

Table 7.2: The five best fitting ME models as deemed by the BIC. Larger BIC values indicate better fitting models. The number of expert networks K and the associated covariates of the best models are also reported.

BIC	K	**Covariates**
-8490.43	4	Age Government satisfaction
-8498.59	3	Age
-8507.33	3	Age Government satisfaction
-8511.37	3	Government satisfaction
-8512.62	5	Age Government satisfaction

The optimal model with $K = 4$ expert networks where age and government satisfaction are the influential covariates is discussed. Each of the five best fitting models considered age and/or government satisfaction as important covariates.

The IMS October 2nd presidential opinion poll was previously analyzed in Chapter 5.6.1 where a two component mixture of Plackett-Luce models was deemed the best model. Under the optimal ME model the Benter dampening parameter estimates are $\underline{\hat{\alpha}} = (1.00, 0.99, 0.97, 0.99, 1.00)$. Due to the constraint $\alpha_1 = 1$ and the non-identifiability of α_N only α_2, α_3 and α_4 are actually estimated. Their associated standard errors are $0.10, 0.12$ and 0.15 respectively. Since the Plackett-Luce model (Chapter 4) is a special case of Benter's model with $\underline{\alpha} = (1, \ldots, 1)$, the proximity of the dampening parameter estimates to 1 along with their relatively large standard errors suggest a Plackett-Luce model would be adequate for modelling this poll data.

Figure 7.2 illustrates the Benter support parameter estimates within each of the four expert networks in the optimal model. The relatively small approximate standard errors of the estimates are also given. Expert network 1 appears to favour the conservative candidates of McAleese and Scallon — at the early stage of the electoral campaign when this poll was conducted Scallon had not yet established herself as a true presidential contender. Thus the 31% support for Scallon in this network is the largest she obtains in this poll. Expert network 2 also reveals characteristics of the early stages of the presidential campaign. Adi Roche has large support in this expert network — at the start of the campaign Roche was a very popular candidate but her support quickly dropped when she became embroiled in difficulties and her campaign went into decline. The third and fourth expert networks have large support parameters for Mary McAleese and Mary Banotti respectively. McAleese (who was elected) and Banotti were two of the main candidates throughout the campaign. Of note is the low levels of support for Nally in any of the expert networks — Nally joined the electoral campaign later than the other candidates on September 29th and so had little time to win votes prior to this October 2nd poll.

According to the BIC, the ME model with four expert networks where age and 'government satisfaction' were influential covariates best models the IMS presidential opinion poll. Table 7.3 details the gating network parameter estimates, the associated odds ratios and the relevant 95% confidence intervals for the odds ratios. The gating network parameters associated with expert network 1 were used as the reference parameters i.e. $\underline{\beta}_1 = (\beta_{10}, \ldots, \beta_{1L}) = (0, \ldots, 0)$. Also, within the government satisfaction covariate the 'do not know/no opinion' level was used as the baseline category.

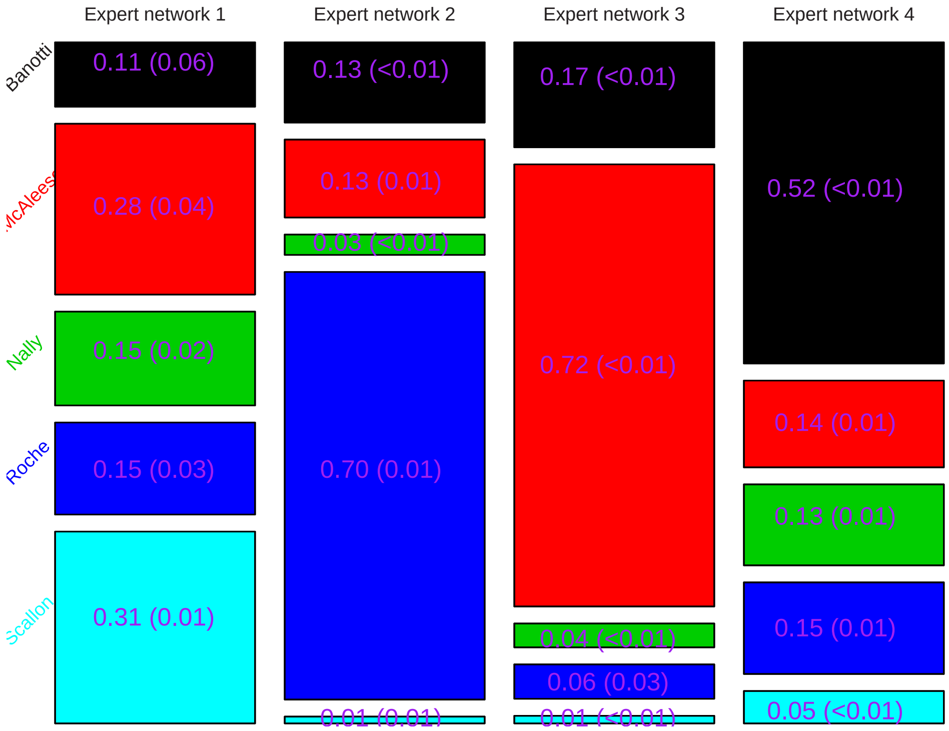


Fig. 7.2: A graphical representation of the maximum likelihood estimates of the Benter support parameters for the October 2nd 1997 presidential opinion poll. Each column of the mosaic represents an expert network — the segments within the columns represent the magnitude of the support parameters for the candidates within each expert network. The maximum likelihood estimate of each support parameter is detailed within each segment. Standard errors for all support parameter estimates are given in parentheses.

Table 7.3: Gating network parameter estimates $\underline{\hat{\beta}}_k$, the associated odds ratios and the 95% odds ratio confidence intervals under the ME model fitted to the October 2nd 1997 presidential opinion poll data. The covariates selected as informative were age and government satisfaction. 'Do not know/no opinion' was used as the reference level within the categorical government satisfaction covariate.

		Intercept	**Age**	**Satisfied**	**Not satisfied**
Expert	Log odds ($\underline{\hat{\beta}}_2$)	0.92	-5.16	0.13	1.03
network	Odds ratio [exp($\underline{\hat{\beta}}_2$)]	2.52	0.01	1.14	2.80
2	95% CI (Odds ratio)	[0.78, 8.16]	[0.00, 0.05]	[0.42, 3.11]	[0.77, 10.15]
Expert	Log odds ($\underline{\hat{\beta}}_3$)	-0.46	-0.05	1.14	1.33
network	Odds ratio [exp($\underline{\hat{\beta}}_3$)]	0.63	0.95	3.12	3.81
3	95% CI (Odds ratio)	[0.16, 2.49]	[0.32, 2.81]	[0.94, 10.31]	[0.90, 16.13]
Expert	Log odds ($\underline{\hat{\beta}}_4$)	0.54	0.44	-1.05	1.25
network	Odds ratio [exp($\underline{\hat{\beta}}_4$)]	1.71	1.56	0.35	3.50
4	95% CI (Odds ratio)	[0.52, 5.58]	[0.35, 6.91]	[0.12, 0.98]	[1.07, 11.43]

In terms of the gating network parameters which refer to expert network 2 (i.e. the pro Roche expert network), for every one unit increase in age the odds for being described best by expert network 2 are 0.01 times greater (or 100 times less) than the odds for being described by expert network 1 (i.e. the conservative expert network.) This would appear to be an intuitive characteristic of the Irish electorate — the more elderly generations in Ireland would generally be considered to be more conservatively minded. Note also the relatively small associated odds ratio confidence interval.

Also, if a voter was satisfied with the current government rather than having no opinion the odds for being described by expert network 2 are 1.14 times greater than the odds for being described by expert network 1. Similarly if a voter was not satisfied with the current government rather than having no opinion the odds for being described by expert network 2 are 2.80 times greater. However, both 95% confidence intervals for the government satisfaction covariate enclose 1 implying it is likely that the voters are no more likely to be described better by expert network 2 than 1. Thus younger voters (perhaps with political opinions) appear to be best described by expert network 2 and were more in favour of Adi Roche.

For every one unit increase in age the odds for being best described by expert network 3 is 0.95 times greater than being best described by expert network 1. Again, the confidence interval for this odds ratio includes 1 suggesting age is not a driving covariate within this expert network. The confidence intervals for the government satisfaction covariates also include 1 but only just. The odds of a voter being best described by expert network 3 are around 3 times greater than the odds for expert network 1 given that the voter has some political opinion. Thus voters with an interest in politics appear to favour Mary McAleese.

The gating parameters for expert network 4 indicate that voters with a dislike for the current government favored Mary Banotti. The confidence interval for the age covariate includes 1 suggesting it has little effect on the odds ratio for being described by expert network 4 over expert network 1. The odds of a voter who indicated a dislike for the 1997 government (a coalition government of Fianna Fáil and the Progressive Democrats) being best described by expert network 4 were 3.50 times greater than being described by expert network 1. In contrast, the odds of a voter in favor of the current government being best described by expert network 4 are 0.35 times greater than the odds for expert network 1. These results make intuitive sense within the context of the 1997 presidential election. Mary Banotti was endorsed by Fine Gael, the main opposition party to Fianna Fáil. Thus voters best described by expert network 4 appear to be Fine Gael supporters. Those voters in favour of the 1997 coalition government were less likely to be described by expert network 4.

7.6 Conclusions

In this chapter a single layer mixtures-of-experts model has been presented. A mixture of Benter's models for rank data has been fitted to the votes cast in an Irish presidential opinion poll, with the observations' covariates also utilized to determine the gating network parameters i.e. each voter's membership of the expert networks. The model was fitted via the EM algorithm with the MM algorithm successfully incorporated at the M step to estimate the model parameters.

An opinion poll conducted early in the 1997 Irish presidential electoral campaign was analyzed. Characteristics of the early stages of the competition were highlighted — one expert network had large support for Adi Roche whose campaign later wilted. Both political persuasion and age emerged as the influential covariates when estimating the gating network parameters.

Chapter 8

A Latent Space Model for Rank Data

Previous chapters have focussed on modelling and exploring the heterogeneous nature of a set of judges who generate rank data. The latent space model introduced in this chapter provides another tool for exploring such a population. The focus however is no longer on examining the heterogeneous nature of the judges but on estimating the relative locations of the judges (and the objects they rank) in a latent space.

A latent space model similar to that of Hoff et al. (2002) is proposed where both voters and candidates are located simultaneously in a D-dimensional latent space. The location of each candidate is inferred from the votes cast by the electorate — the Plackett-Luce model for rank data (Chapter 4) is employed to exploit the information incorporated in the ranked preferences contained in the votes. In turn, a voter's location is determined by their vote which demonstrates their support for each of the candidates. This model is fitted within the Bayesian paradigm (see Chapter 3); the Metropolis-Hastings algorithm is the primary model fitting tool. When fitting latent space models issues such as invariant configurations and choice of dimensionality arise; these are dealt with in Chapters 8.2.2 and 8.2.3 respectively.

The relative spatial locations of the candidates are suggestive of the type of relationships that may exist between the candidates, as viewed by the electorate. As coalition governments often occur in countries that use proportional representation election systems, interest lies in examining if candidates from different political parties are deemed alike. Which political parties are viewed as similar by the electorate? What characteristics do closely located candidates share? What mechanisms drive Irish general elections? Such questions will be answered by examining the relative locations of the candidates.

The latent space model presented in this chapter focuses on two Irish elections — the general election of 2002 and the Irish presidential election in 1997. The actual votes from the 2002 general election in the constituencies of Dublin North, Dublin West and Meath are analyzed. Eight opinion polls taken during the canvassing period prior to the 1997 presidential election are also examined. Details of these elections are outlined in Chapter 2.

Configurations of the candidates and electorate from the 2002 general election and from the 1997 Irish presidential election indicate that voter preferences are both politically and candidate driven. Mapping the spatial movement of the presidential candidates during canvassing provides an insight to how electoral opinions developed prior to the election.

The work presented in this chapter is reported in Gormley and Murphy (2006*b*).

8.1 Model Specification

A latent space model is combined with a model for rank data to provide a suitable tool for the modelling of PR-STV data.

8.1.1 The Latent Space Model

Hoff et al. (2002) proposed a model for social networks where the network actors are located in a latent space and the probability of a connection between two actors is determined by their proximity. In a similar vein to this work, a model is proposed for rank data where voters and candidates are located in the same D dimensional latent space $\mathbf{Z} \subseteq \Re^D$. It is assumed that each of M voters has latent location $\underline{z}_i \in \mathbf{Z}$ and each candidate j ($j = 1, \ldots, N$) has latent location $\underline{\zeta}_j \in \mathbf{Z}$. Hence, the N preferences of the M voters are described using $(N + M) \times D$ parameters. Let $d(\underline{z}_i, \underline{\zeta}_j)$ be the squared Euclidean distance between voter i and candidate j in the latent space $\mathbf{Z}$, that is

$$d(\underline{z}_i, \underline{\zeta}_j) = \frac{1}{D}\sum_{d=1}^{D}(z_{id} - \zeta_{jd})^2$$

for $1 = 1, \ldots, M$ and $j = 1, \ldots, N$. The squared Euclidean distance is invariant to rotations and translations. Many other distance measures are available as detailed by Mardia et al. (1979) and possible alternatives are discussed in Chapter 9.

The distance $d(\underline{z}_i, \underline{\zeta}_j)$ (for $j = 1 \ldots, N$) between voter i and the candidates describes the voter's electoral opinions. In a similar way the proximity of two candidates in the latent space quantitatively describes their relationship as deemed by the electorate.

By exploiting the information contained in the ranked preferences the latent locations of each voter and candidate can be inferred. Thus a latent space model is incorporated with a standard rank data model to spatially model Irish voting data.

8.1.2 The Plackett-Luce Model

In the Plackett-Luce model (see Chapter 3.1), a ranking is modelled as a sequential process in which each voter selects the next most preferred candidate. In the context of a latent space model, the Plackett-Luce model is parameterized by a 'support' parameter vector

$$\underline{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iN})$$

for each of $i = 1, \ldots, M$ voters where $\sum_{j=1}^{N} p_{ij} = 1$. The parameter p_{ij} can be interpreted as the probability of voter i selecting candidate j in first place on their ballot. The probability p_{ij} is a decreasing function of the distance between the voter and the candidate in the latent space. It is assumed that these probabilities take the form

$$p_{ij} = \frac{\exp\{-d(\underline{z}_i, \underline{\zeta}_j)\}}{\sum_{j'=1}^{N} \exp\{-d(\underline{z}_i, \underline{\zeta}_{j'})\}}$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$. Thus the position taken by each voter and candidate in the latent space is determined by the preferences expressed on the ballot forms.

Under a Plackett-Luce model with support parameters $\mathbf{p} = (\underline{p}_1, \ldots, \underline{p}_M)$ the probability of all votes $\mathbf{x}$ is

$$\mathbf{P}\{\mathbf{x}|\mathbf{p}\} = \prod_{i=1}^{M} \prod_{t=1}^{n_i} \frac{p_{ic(i,t)}}{\sum_{s=t}^{N} p_{ic(i,s)}}.$$

8.2 Model Fitting

The Plackett-Luce model combined with a latent space model allows modelling of the ranked nature of the PR-STV data and the spatial modelling of the electoral candidates. The parameters of this model and their related uncertainty are estimated within a Bayesian framework.

Prior densities for voter locations, $p_v(\underline{z}_i)$, and for candidate locations, $p_c(\underline{\zeta}_j)$ are assumed to be Normal and independent where $z_{id} \sim \mathrm{N}(\mu_{Pv}, \sigma^2_{Pv}) \sim \mathrm{N}(0, 3^2)$ and $\zeta_{jd} \sim \mathrm{N}(\mu_{Pc}, \sigma^2_{Pc}) \sim \mathrm{N}(0, 3^2)$ for $d = 1, \ldots, D$. The prior parameters were selected so that the prior was concentrated on a region around the origin without being overly informative. Thus the joint density $P\{\mathbf{X}, \mathbf{z}, \zeta\}$ of the votes cast, the voter locations and the candidate locations is

$$P\{\mathbf{X}, \mathbf{z}, \zeta\} \quad = \quad \left[\prod_{i=1}^{M} \prod_{t=1}^{n_i} \frac{\exp\{-d(\underline{z}_i, \underline{\zeta}_{c(i,t)})\}}{\sum_{s=t}^{N} \exp\{-d(\underline{z}_i, \underline{\zeta}_{c(i,s)})\}} \right] \left[\prod_{i=1}^{M} p_v(\underline{z}_i) \right] \left[\prod_{j=1}^{N} p_c(\underline{\zeta}_j) \right].$$

The location of each voter and each candidate in the latent space are to be estimated — samples from the posterior distribution $\mathbf{P}\{\mathbf{z}, \zeta|\mathbf{X}\}$ are generated using a Metropolis-Hastings algorithm. A random walk proposal density where each location was perturbed using normally distributed noise was employed — good acceptance rates (detailed in Chapter 8.3) were achieved in the estimation of both voter and candidate locations using this proposal.

8.2.1 Estimation of Voter and Candidate Latent Locations

The location of each voter $\underline{z}_i$ and each candidate $\underline{\zeta}_j$ within a D dimensional latent space is to be estimated. A random walk Metropolis-Hastings algorithm is used to sample from the joint density $\mathbf{P}\{\mathbf{z}, \zeta|\mathbf{X}\}$.

Estimates of $\underline{z}_i$ are generated from the posterior distribution via the following algorithm:

1. Generate a value ϵ from the symmetric proposal density $N(0, \sigma_v^2)$ and form the proposal point $z_{id}^* = z_{id} + \epsilon$ for $d = 1, \ldots, D$.

2. Compute the acceptance probability $\alpha(\underline{z}_i^*, \underline{z}_i)$ as follows

$$\begin{aligned}
\alpha(\underline{z}_i^*, \underline{z}_i) &= \min\left\{ \frac{P(\mathbf{z}^*, \zeta|\mathbf{X})}{P(\mathbf{z}, \zeta|\mathbf{X})} \ , \ 1 \right\} \\
&= \min\left[\frac{\left\{ \prod_{t=1}^{n_i} \frac{\exp\{-d(z_i^*, \zeta_{c(i,t)})\}}{\sum_{s=t}^{N} \exp\{-d(z_i^*, \zeta_{c(i,s)})\}} \right\} \exp\left\{ -\frac{(z_i^* - \mu_{Pv})^2}{2\sigma_{Pv}^2} \right\}}{\left\{ \prod_{t=1}^{n_i} \frac{\exp\{-d(z_i, \zeta_{c(i,t)})\}}{\sum_{s=t}^{N} \exp\{-d(z_i, \zeta_{c(i,s)})\}} \right\} \exp\left\{ -\frac{(z_i - \mu_{Pv})^2}{2\sigma_{Pv}^2} \right\}} \ , \ 1 \right]
\end{aligned}$$

where independence of voter locations and a symmetric random walk proposal distribution are assumed.

3. Generate a value $u \sim \text{Uniform}(0, 1)$.

4. If $u \leq \alpha(\underline{z}_i^*, \underline{z}_i)$ then define $\underline{z}_i = \underline{z}_i^*$, otherwise define $\underline{z}_i = \underline{z}_i$.

Similar methodology applies in the case of estimating candidate locations:

1. Generate a value $\epsilon \sim N(0, \sigma_c^2)$ and let $\zeta_{jd}^* = \zeta_{jd} + \epsilon$ for $d = 1, \ldots, D$.

2. Compute the acceptance probability $\alpha(\underline{\zeta}_j^*, \underline{\zeta}_j)$ as follows

$$\begin{aligned}
\alpha(\underline{\zeta}_j^*, \underline{\zeta}_j) &= \min\left\{ \frac{P\{\mathbf{z}, \zeta^*|\mathbf{X}\}}{P\{\mathbf{z}, \zeta|\mathbf{X}\}} \ , \ 1 \right\}. \\
&= \min\left[\frac{\left\{ \prod_{t=1}^{n_i} \frac{\exp\{-d(z_i, \zeta_{c(i,t)}^*)\}}{\sum_{s=t}^{N} \exp\{-d(z_i, \zeta_{c(i,s)}^*)\}} \right\} \exp\left\{ -\frac{(\zeta_j^* - \mu_{Pc})^2}{2\sigma_{Pc}^2} \right\}}{\left\{ \prod_{t=1}^{n_i} \frac{\exp\{-d(z_i, \zeta_{c(i,t)})\}}{\sum_{s=t}^{N} \exp\{-d(z_i, \zeta_{c(i,s)})\}} \right\} \exp\left\{ -\frac{(\zeta_j - \mu_{Pc})^2}{2\sigma_{Pc}^2} \right\}} \ , \ 1 \right]
\end{aligned}$$

where independence of candidates locations and a symmetric random walk proposal are assumed.

3. Generate a value $u \sim \text{Uniform}(0, 1)$.

4. If $u \leq \alpha(\underline{\zeta}_j^*, \underline{\zeta}_j)$ then define $\underline{\zeta}_j = \underline{\zeta}_j^*$, otherwise define $\underline{\zeta}_j = \underline{\zeta}_j$.

The algorithm sequentially estimates the voter locations and then the candidate locations until sufficient mixing of the Markov chain is achieved. Locations estimated subsequent to the burn-in period are considered when calculating final estimates.

8.2.2 Invariant Configurations

The measure of distance between voter and candidate locations in the latent space is quantified by the squared Euclidean distance. This distance is invariant to rotations and translations. As a result the model is not fully identifiable because the locations are only identified up to rotation and translation. Procrustean methods are used to eradicate this problem.

Procrustean methods (Krzanowski, 1988) match one configuration of points to another as well as possible in a least squares sense. Transformations such as dilation, rotation and translation are used to create the match. In this context only translations and rotations are applicable without altering the likelihood of the data due to the definition of the probabilities p_{ij}.

Assume $C^R = \zeta^R$ is a reference configuration of the candidate locations which is centered around the origin. As covariates associated with the voters are generally not publicly available the focus here is on the relative locations of the electoral candidates. Thus within this context when performing Procrustes techniques the reference configuration C^R refers to the configuration of candidates only. To match the estimated configuration $\hat{C}$ to the reference configuration C^R, $\hat{C}$ is first translated so that it is also centered around the origin. $\hat{C}$ is then rotated to provide the best match with C^R in a least squares sense.

To obtain Q, the optimal orthogonal rotation matrix, the sum

$$\begin{aligned} S &= \sum_{j=1}^{N}\sum_{d=1}^{D}(c_{id}^{R} - \hat{c}_{id})^2 \\ &= \sum_{j=1}^{N}\sum_{d=1}^{D}(\zeta_{jd}^{R} - \hat{\zeta}_{jd})^2 \\ &= \text{trace}\left\{C^R C^{R'} + \hat{C}\hat{C}' - 2C^R\hat{C}'\right\} \qquad (8.1) \end{aligned}$$

is minimized. The newly rotated configuration is denoted $\hat{C}Q$. Thus (8.1) becomes

$$S = \text{trace}\left\{C^R C^{R'} + \hat{C}\hat{C}' - 2C^R Q'\hat{C}'\right\}$$

and the minimization problem becomes the constrained maximization of $2C^RQ'\hat{C}'$. It follows that $Q = VU'$ where $U\Sigma V'$ is the singular value decomposition of $C^{R'}\hat{C}$. Thus by centering each estimated configuration around the origin and rotating the configuration using the rotation matrix Q the estimated configuration $\hat{C}$ is best matched with the reference configuration C^R.

Samples of the configuration $(\mathbf{z}, \zeta)$ are generated using the Metropolis-Hastings algorithm (Chapter 8.2.1). Initial iterations of the algorithm are constrained to only accept uphill moves (ie. moves when $\alpha(\underline{z}_i^*, \underline{z}_i) \geq 1$ and $\alpha(\underline{\zeta}_i^*, \underline{\zeta}_i) \geq 1$) to achieve an estimate of the *maximum a posteriori* (MAP) configuration of candidate locations. This MAP configuration is henceforth employed as C^R, the reference configuration, to which each subsequently estimated configuration $\hat{C}$ is matched. C^R is not assumed to be the correct configuration but is merely used as a standard to which others are matched. Locations estimated during the uphill only runs of the Metropolis-Hastings algorithm are not considered when calculating final estimates.

8.2.3 Dimensionality

The dimensionality D of the latent space is a further variable which requires estimation. Several techniques have been discussed in the literature (eg. Airoldi et al. (2006)) as potential methods for selecting the optimal dimensionality of a space. Selecting the optimal D can been viewed as a model selection process between models with different dimensions. Methods such as the deviance information criterion (DIC) and Pritchard et al.'s criterion (detailed in Chapter 3) are examined here.

A practical alternative to these two criteria is to apply principal components analysis Mardia et al. (1979) to the resulting configurations for each choice of dimension D. The principal components analysis (PCA) rotates the configuration of candidate locations so that the variance of the locations is concentrated in a subset of the dimensions: the first principal component dimension has maximal variance, the second has maximal variance subject to being orthogonal to the first dimension, etc.

PCA is applied to the configuration of candidates only as the predominant interest lies in the interpretation of the relative locations of the candidates. The variances of the resulting principal components are examined and the optimal number of dimensions D is selected to be the number of dimensions after which the addition of another dimension was not deemed to be beneficial; a threshold of 20% was used to determine if the addition of an extra dimension was worthwhile.

8.3 A Latent Space Model for Irish Voting Data

A latent space model for rank data was applied to votes from the 2002 Irish general election and to opinion polls conducted prior to the 1997 Irish presidential election.

8.3.1 The 2002 General Election: Dublin North Constituency

Twelve candidates campaigned for four parliamentary seats in the Dublin North constituency. Glennon (Fianna Fáil), Ryan (Labour), Sargent (Green Party) and Wright (Fianna Fáil) were elected to the Dáil (see Table 2.2). The latent space model incorporating the Plackett-Luce model was fitted to the 43942 Dublin North votes over the range of dimensions $D = 1, \ldots, 4$. A random walk proposal density was employed; for dimensions $D = 1$ and 2 the proposal parameters were fixed to be

$$\begin{aligned} N(0, \sigma_v^2) &= N(0, 3^2) \\ N(0, \sigma_c^2) &= N(0, 0.02^2). \end{aligned}$$

For dimensions $D = 3$ and 4 the proposal parameters were

$$\begin{aligned} N(0, \sigma_v^2) &= N(0, 10^2) \\ N(0, \sigma_c^2) &= N(0, 0.02^2). \end{aligned}$$

A range of model selection criteria were computed for each different dimension model fitted — the values obtained for the DIC and Pritchard et al.'s criterion are reported in Table 8.1. The dimensions of the best fitting models as deemed by these criterion appear to contradict each other somewhat and thus principal components analysis was employed as the method of selecting the optimal D.

Table 8.2 shows the variation captured by each principal component when different dimensions of latent space model were fitted to the data. Both dimensions $D = 1$ and $D = 2$ appear to summarize the data well in that each component explains more than one fifth of the variance of the candidate locations. The addition of further dimensions to the model only accounted for 12-15% of the variance of the data. Hence both the one dimensional and two dimensional configurations are analyzed to examine relationships between the Dublin North electoral candidates.

Table 8.1: Model selection criteria for latent space models of dimension $D = 1, 2, 3$ and 4 fitted to votes cast in the Dublin North constituency. The entries in bold font indicate the best fitting model according to each criterion.

Dimension	**DIC**	**Pritchard et al.**
1	873604	838510
2	837328	**829525**
3	**829522**	1057733
4	843512	2759114

Table 8.2: Proportion of the variance explained by each principal component when PCA is applied to the Dublin North candidate locations over the range of dimensions $D = 1, \ldots, 4$. Principal components analysis was applied to the average candidate configuration only as the main interest lies in the relative locations of the candidates. Entries in bold indicate the models deemed as good models using a 20% minimum variance criterion.

	Variances			
Dimension	σ_1^2	σ_2^2	σ_3^2	σ_4^2
1	**1**	-	-	-
2	**0.66**	**0.34**	-	-
3	0.59	0.29	0.12	-
4	0.59	0.26	0.12	0.03

Figure 8.1 illustrates the one dimensional configuration of the twelve Dublin North candidates. Each candidate is represented by an abbreviation of their surname and political affiliation as detailed in Table 2.2. Candidates with the same political affiliations are illustrated in the same colour. Figure 8.2 shows the two dimensional configuration of the twelve candidates.

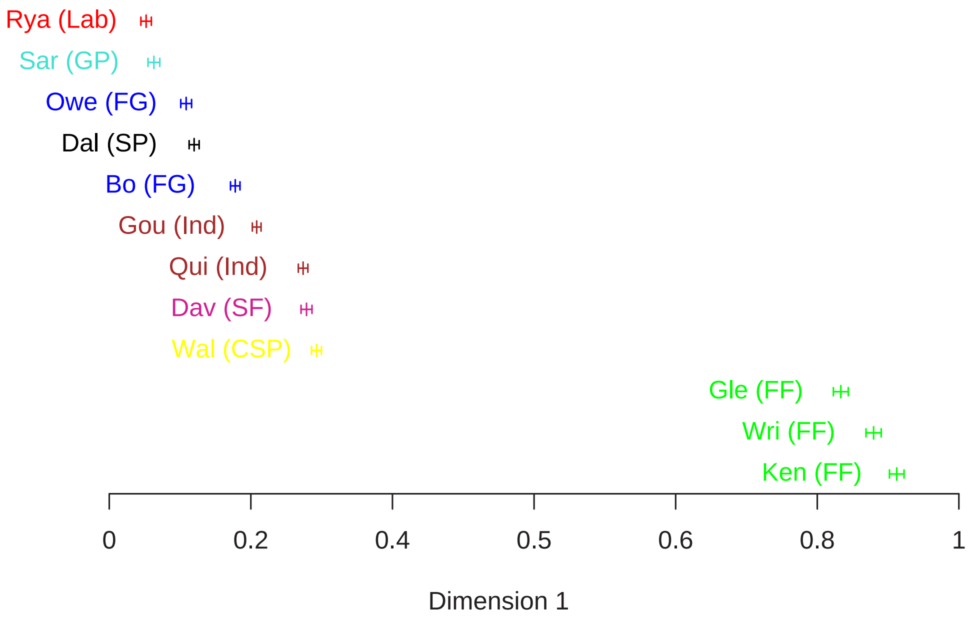


Fig. 8.1: The one dimensional configuration of the candidate means, averaged over a Metropolis-Hastings algorithm, and their associated uncertainty (indicated by ±2 standard deviation intervals). Each Dublin North candidate is denoted by an abbreviation of their surname and political affiliation (see Table 2.2). Candidates from different parties are plotted in different colours. The mean locations were estimated by 25000 Metropolis-Hastings iterations (post burn-in), thinned after every 100th iteration. The mean acceptance rate for the candidate locations was 12%.

Both configurations of the Dublin North candidates suggest party politics play an important role in the electorate's view of the candidates. Also of note in both configurations are the relatively small uncertainties associated with the estimated locations. The Fianna Fáil candidates are located on the far right of the first dimension in both configurations with all other candidates located on the opposite side of the dimension. Fianna Fáil are currently in power in Ireland and are the largest Irish political party.

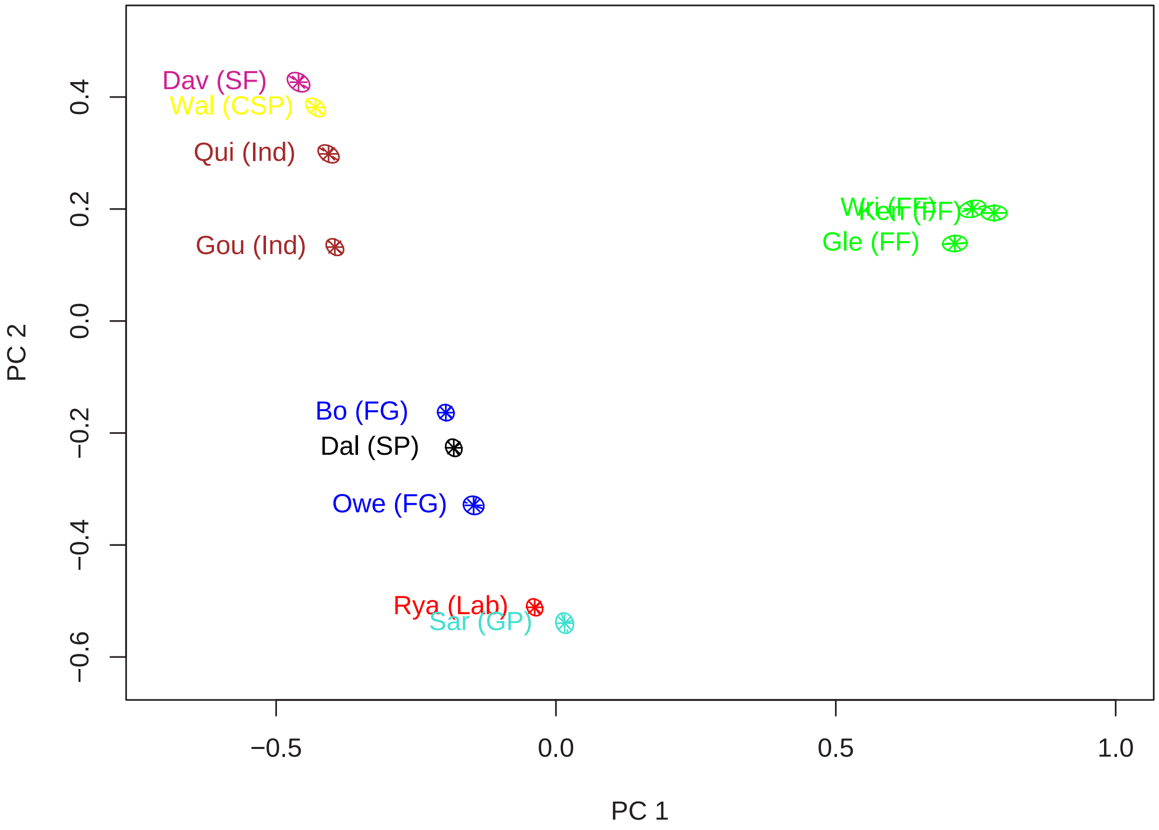


Fig. 8.2: The two dimensional configuration of the Dublin North candidate means with their associated uncertainty. The center of each ellipse indicates the posterior mean location of each candidate — the ellipses are approximate 95% posterior sets which indicate the uncertainty in the candidate positions. The position of each candidate and the ellipses are estimated by 25000 Metropolis-Hastings iterations (post burn-in), thinned after every 100th iteration. The mean acceptance rate for the candidate locations was 16%.

Interestingly the two candidates located farthest left in Figure 8.1 (Seán Ryan (Labour) and Trevor Sargent (Green Party)) were both elected as were the Fianna Fáil candidates Jim Glennon and G.V. Wright. There appears to be a wide range of political preferences within the Dublin North constituency.

In Figure 8.2 the first dimension demonstrates the 'Fianna Fáil versus the rest' characteristic of Irish elections. The second dimension however appears to be more candidate driven. The popular candidates of Ryan and Sargent are located towards the bottom of the second dimension. Candidates affiliated with Sinn Féin and the Christian Solidarity party would have a smaller more localized following and are positioned near the top of the second dimension. Also of interest is the location of Clare Daly of the Socialist Party amongst the two Fine Gael candidates of Nora Owen and Cathal Boland. While Owen and Boland were running mates, Owen and Daly were the only two female candidates in the constituency, perhaps giving further evidence of a candidate driven dimension.

8.3.2 The 2002 General Election: Dublin West Constituency

In 2002 three Dáil Éireann seats were to be filled in the constituency of Dublin West with nine candidates running for election. Burton (Labour), Higgins (Socialist Party) and Lenihan (Fianna Fáil) were the candidates elected. There was a total of 29988 valid votes cast to which a latent space model incorporating the Plackett-Luce model was fitted. Random walk proposal distributions were used within the Metropolis-Hastings algorithm. The proposal parameters used were as utilized in the Dublin North analysis (see Chapter 8.3.1).

As illustrated in Table 8.3 the optimal dimensional latent space required to represent the Dublin West electorate and candidates is unclear. Different model selection techniques provide different values for the optimal D. The multivariate analysis technique of principal component analysis was therefore applied to the configuration of candidates only as interest lies in the relative locations of the candidates. Table 8.4 details the proportion of the variance of the candidate locations accounted for by the principal components fitted to different dimensional latent configurations.

Dimensions $D = 1$, $D = 2$ and $D = 3$ appear to summarize the data well in that each component explains more than one fifth of the variance of the candidate locations. The addition of a fourth principal component only accounts for an extra 3% of the variance of candidate locations.

Table 8.3: Model selection criteria for latent space models of dimension $D = 1, \ldots, 4$ fitted to votes cast in the Dublin West constituency. The entries in bold font indicate the best fitting model according to each criterion.

Dimension	**DIC**	**Pritchard et al.**
1	465158	457374
2	457380	**442813**
3	**449307**	552498
4	456474	1536278

Table 8.4: Proportion of the variance explained by each principal component when PCA is applied to the Dublin West candidate locations over the range of dimensions $D = 1, \ldots, 4$. Principal components analysis was applied to the average candidate configuration only as the main interest lies in the relative locations of the candidates. Entries in bold indicate well fitting models determined by a 20% minimum variance criterion.

	Variances			
Dimension	σ_1^2	σ_2^2	σ_3^2	σ_4^2
1	**1**	-	-	-
2	**0.67**	**0.33**	-	-
3	**0.52**	**0.28**	**0.20**	-
4	0.50	0.27	0.20	0.03

Configurations of dimension $D = 1, 2$ and 3 are reported to examine the relationships which exist between the Dublin West candidates as deemed by the Irish electorate. Of note within each dimension is the relatively small uncertainty associated with each candidate's posterior mean location estimate.

The one dimensional plot of the Dublin West candidate configuration is similar to that produced when analyzing the Dublin North votes in Chapter 8.3.1. The Fianna

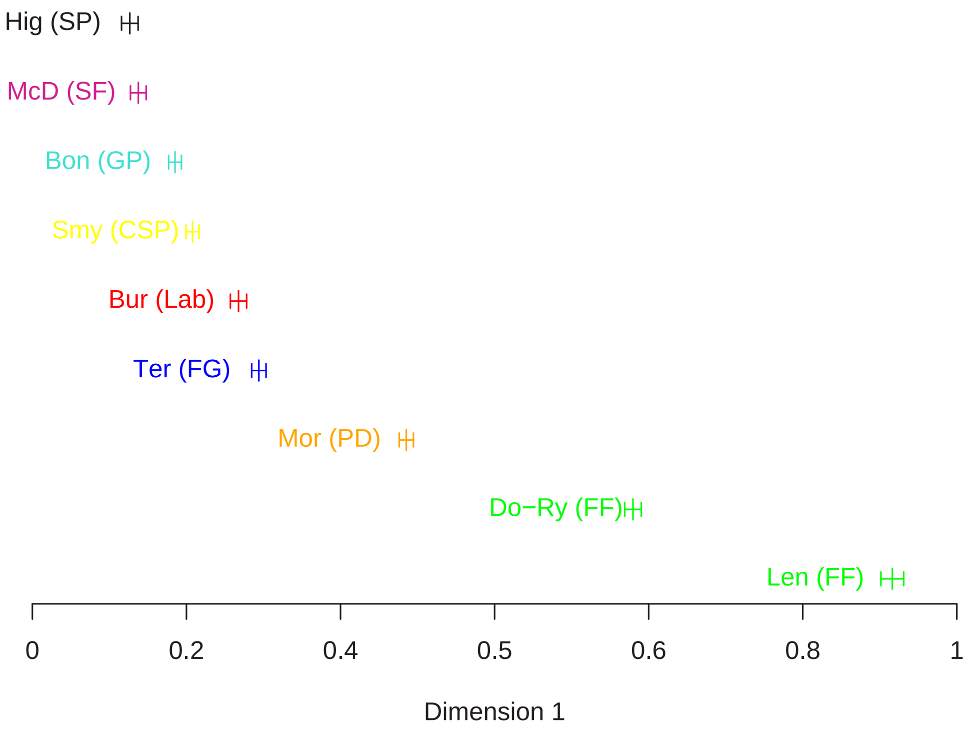


Fig. 8.3: The one dimensional configuration of the candidate means, averaged over a Metropolis-Hastings algorithm, and their associated uncertainty (indicated by ±2 standard deviation intervals). Each Dublin West candidate is denoted by abbreviations of their surname and political affiliation (see Table 2.3). Candidates from different parties are plotted in different colours. Candidate locations were estimated by 25000 iterations of the Metropolis-Hastings algorithm (post burn-in), thinned every 100th iteration. The mean acceptance rate for candidate locations was 14%.

Table 8.5: Numbering of the Dublin West candidates associated with Figure 8.5 and Figure 8.6.

1 = Bonnie (GP)	2 = Burton (Lab)	3 = Doherty-Ryan (FF)
4 = Higgins (SP)	5 = Lenihan (FF)	6 = McDonald (SF)
7 = Morrissey (PD)	8 = Smyth (CSP)	9 = Terry (FG)

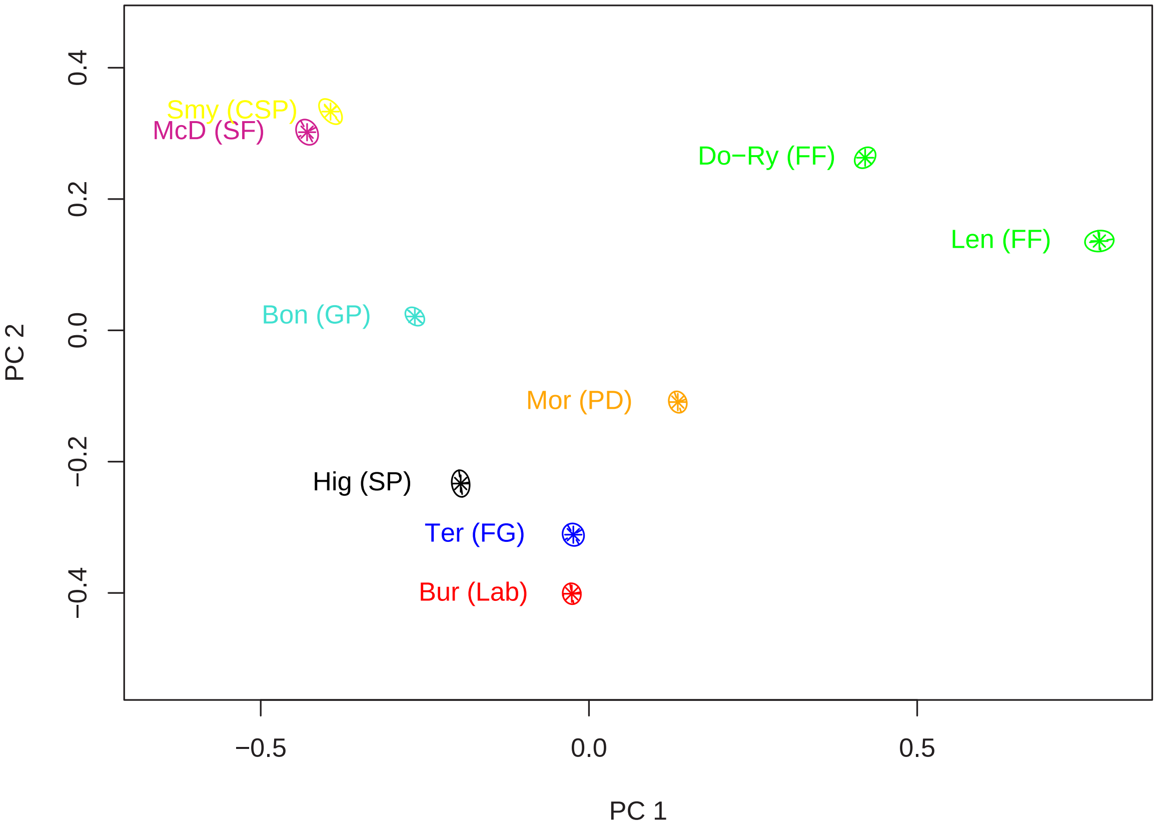


Fig. 8.4: The two dimensional configuration of the Dublin West candidate means with their associated uncertainty. The center of each ellipse indicates the posterior mean location of each candidate — the ellipses are approximate 95% posterior sets which indicate the uncertainty in the candidate positions. The position of each candidate and the ellipses are estimated by 25000 Metropolis-Hastings iterations (post burn-in), thinned after every 100th iteration. Of the proposed candidate locations 15% were accepted.

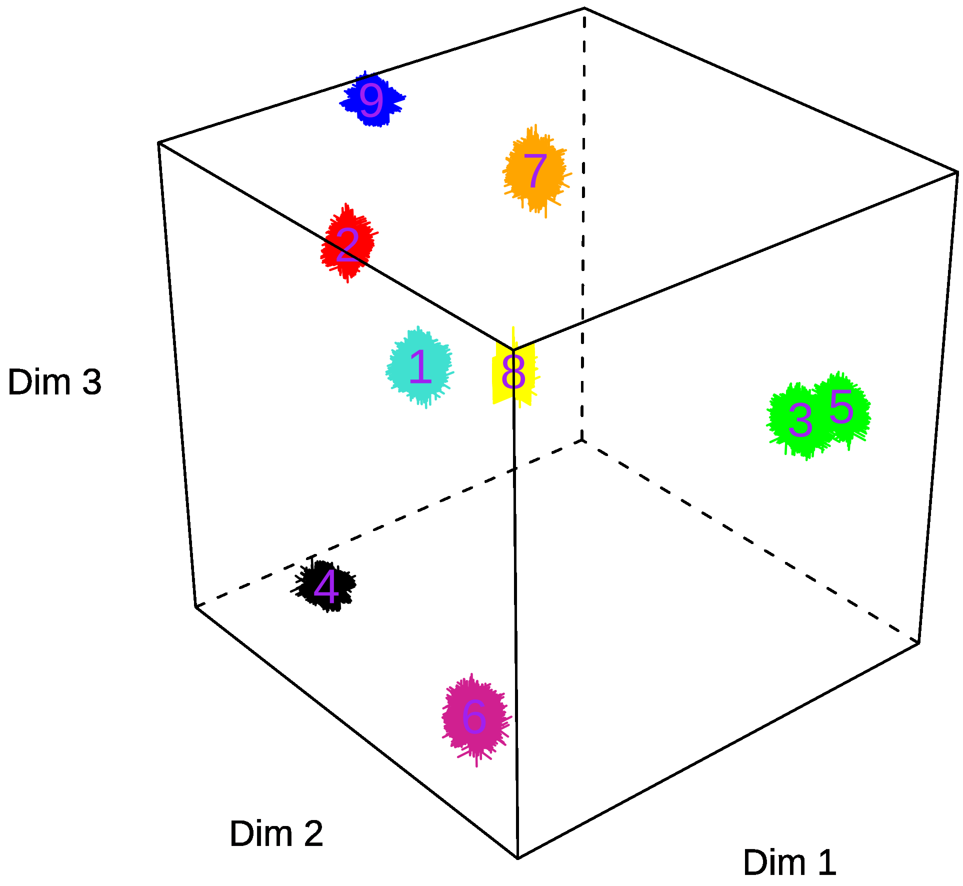


Fig. 8.5: The three dimensional configuration of the Dublin West candidate means. The candidates are numbered in alphabetical order as detailed in Table 8.5. Candidates with different political affiliations are coloured differently. Each '+' symbol denotes a realization of a candidate location sampled during 65000 Metropolis-Hastings iterations (post burn-in), thinned after every 100th iteration. 20% of proposed candidate locations were accepted.

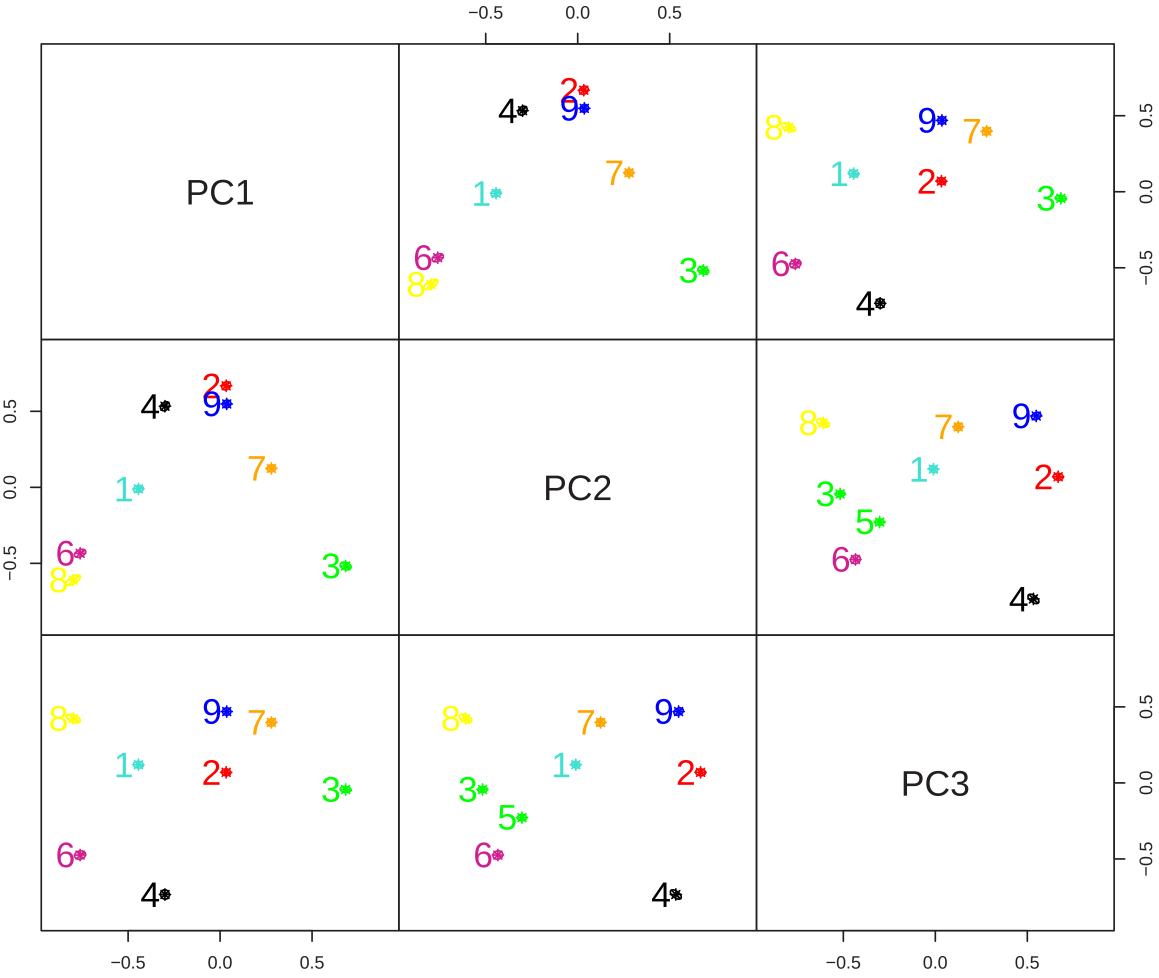


Fig. 8.6: A matrix of scatter plots of the three dimensional configuration of the Dublin West candidate means. The candidates are numbered in alphabetical order as detailed in Table 8.5 and the ellipses are approximate 95% posterior sets which indicate the uncertainty in each of the candidate positions. Candidates with different political affiliations are coloured differently.

Fáil candidates are located on the far right with the more socialist candidates on the left. Interestingly in the one dimensional configuration, the Progressive Democrat candidate Morrissey is situated closest to the Fianna Fáil candidates — Fianna Fáil and the Progressive Democrats formed a coalition government as a result of this general election and Fianna Fáil voters had been encouraged to give high preference to the Progressive Democrat candidate. Some indication of candidate driven voter preferences is apparent from the one dimensional configuration — the ordering of the candidates on the left side of the dimension are different to the order observed in the Dublin North constituency. In Dublin North, the Labour candidate Ryan was located farthest from the Fianna Fáil candidates whereas the Dublin West Labour candidate Burton is more centrally located.

Figure 8.4 illustrates the two dimensional configuration of the Dublin West candidates. Again similarities to the Dublin North configuration are apparent. The first dimension which accounts for the larger proportion of the variance of the candidate locations divides the candidates by political ideals — Fianna Fáil lie on the right with the more socialist parties situated on the left. The second dimension appears to be candidate driven. Candidates such as Burton, Terry and Higgins received large numbers of first preferences (see Table 2.5) and are located towards the bottom of the dimension. Less popular candidates such as McDonald and Smyth lie near the top. In terms of number of first preference votes however, Lenihan was the clear leader but is located centrally in the second dimension.

The three dimensional configuration of the Dublin West candidates is illustrated in Figures 8.5 and 8.6. The striking feature of the three dimensional plot in Figure 8.5 is the isolation of the Fianna Fáil candidates (numbers 3 and 5) and the socialist candidates (numbers 4 and 6) representing the Socialist Party and Sinn Féin. Examining the configuration dimension by dimension as demonstrated in Figure 8.6 suggests that the first dimension separates candidates by political affiliation i.e. by a 'Fianna Fáil versus the rest' criterion. The second dimension appears to be candidate driven with the third dimension separating candidates on the level of their socialist views.

Evidence of the 'candidate centered but party wrapped' theory of Irish electoral campaigns detailed by Marsh (2000) is clear from the configurations of the Dublin

West configurations.

8.3.3 The 2002 General Election: Meath Constituency

Five seats in Dáil Éireann were available for election in the Meath constituency in the 2002 general election. Brady (Fianna Fáil), Bruton (Fine Gael), Dempsey (Fianna Fáil), English (Fine Gael) and Wallace (Fianna Fáil) were elected. A latent space model, incorporating the Plackett-Luce model for rank data, was fitted to the 64081 electronic votes over the range of dimensions $D = 1, \ldots, 4$. Proposal densities were fixed to be

$$\begin{aligned} N(0, \sigma_v^2) &= N(0, 3^2) \\ N(0, \sigma_c^2) &= N(0, 0.02^2). \end{aligned}$$

The DIC and Pritchard et al's criterion were computed (Table 8.6) to determine the appropriate dimension for the latent space. DIC suggested $D = 3$ whereas Pritchard et al's criterion suggested $D = 2$. Due to these contrasting results principal components analysis was applied to the different dimensional configurations of Meath candidates.

Table 8.6: Model selection criteria values to indicate the optimal latent space model for the Meath constituency. Latent space models were fitted over the range of dimensions $D = 1, \ldots, 4$. Entries in bold indicate the best fitting model according to each criterion.

Dimension	**DIC**	**Pritchard et al.**
1	1263878	1229179
2	1209849	**1224901**
3	**1193842**	1559845
4	1224017	5538406

Table 8.7 shows the variation captured by each principal component when different dimensions of latent space model were fitted to the data. Both dimensions $D = 1$ and $D = 2$ appear to summarize the data well in that each dimension accounts for more than one fifth of the total variance. When a three dimensional model was

fitted the additional principal component only accounted for 9% of the variance of the data.

Table 8.7: The proportion of the variance explained by each principal component computed for configurations of the candidates in the Meath constituency, for a range of dimensions. Principal components analysis was applied to the average candidate configuration only, as the main interest of this study lies in the relative locations of the candidates. Entries in bold indicate the models deemed as good models using a 20% minimum variance criterion.

	Variances			
Dimension	$\boldsymbol{\sigma_1^2}$	$\boldsymbol{\sigma_2^2}$	$\boldsymbol{\sigma_3^2}$	$\boldsymbol{\sigma_4^2}$
1	**1**	-	-	-
2	**0.67**	**0.33**	-	-
3	0.65	0.26	0.09	-
4	0.57	0.28	0.10	0.05

Both the one dimensional and two dimensional configurations are analyzed to examine relationships between the electoral candidates.

One Dimensional Results

Figure 8.7 illustrates the one dimensional configuration of the fourteen Meath candidates. Each candidate is represented by an abbreviation of their surname and political party as detailed in Table 2.4. It is immediately clear that party politics plays a large role in the electorates' view of the candidates. The Fianna Fáil candidates (Brady, Dempsey and Wallace) are located on the far right of the single dimension with the Fine Gael candidates (Bruton, English and Farrelly) located on the far left. Fianna Fáil and Fine Gael are the two largest (and rival) Irish political parties. The other candidates lie between the two poles created by the Fianna Fáil and Fine Gael candidates but closer to Fine Gael. Interestingly Ward, who is a Labour Party candidate, is located closest to the Fine Gael candidates — Fine Gael and Labour have a history of forming coalition governments (most recently from 1994–1997). Also of note are the narrow interval estimates for the estimated candidate positions (mean ±2 standard deviations are shown). This suggests low uncertainty in the candidate locations in one dimension.

The one dimensional configuration of candidates within the Meath constituency is similar to that of Dublin North and Dublin West in that there appears to be a Fianna Fáil versus non-Fianna Fáil division. However again evidence of candidate driven preferences is apparent from the different order of the candidates located on the left of the dimension. In Dublin North and Dublin West the Fine Gael candidates were quite centrally located whereas in Meath the Fine Gael candidates are situated farthest from the Fianna Fáil candidates. Thus within each constituency the candidates themselves are driving voter preferences to some degree; if this was not the case the same party order would occur across the single dimension in each constituency.

Two Dimensional Results

Good acceptance rates of 32% and 15% were achieved for the voter and candidate positions respectively when a two dimensional model was fitted.

Figure 8.8 illustrates the final average position of each of the fourteen candidates in the Meath constituency. Each candidate is denoted by the abbreviations detailed in Table 2.4. Party politics are again demonstrated as the mechanism which drives this election. The first principal component separates candidates by their political ideals — the estimated positions shows a clear divide between Fianna Fáil and the other parties in the x-axis direction. The second principal component illustrates the presence of an ideological cleavage (left to right wing) of the candidates. For example, the Christian Solidarity Party espouse right wing conservative values and their candidate Redmond (Rd) is located highest in the second principal component.

The plot also includes ellipses which show approximate 95% posterior set estimates of each candidate location to represent the uncertainty in the estimated locations. The uncertainty associated with all candidate locations is low. Furthermore, there is considerable overlap between candidates from the same party.

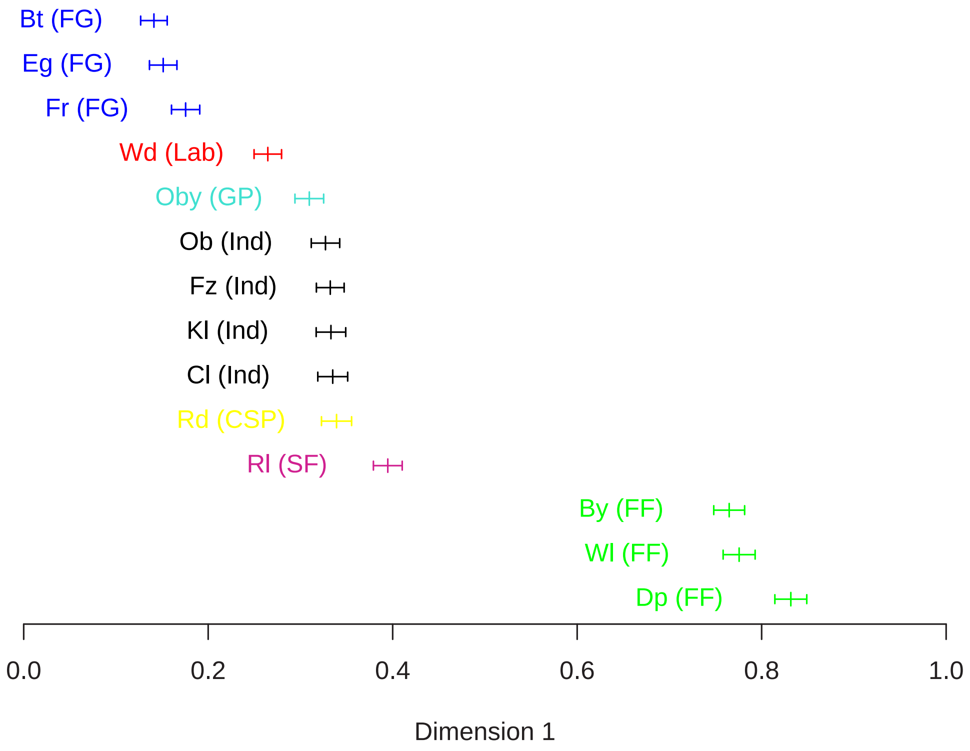


Fig. 8.7: The one dimensional configuration of the Meath candidate means, averaged over a Metropolis-Hastings algorithm, and their associated uncertainty (indicated by ±2 standard deviation intervals). Each of the fourteen candidates as detailed in Table 2.4 are denoted by an abbreviation of their surname and political party. Candidates from different parties are plotted in different colours. The mean acceptance rate for candidate locations was 35%.

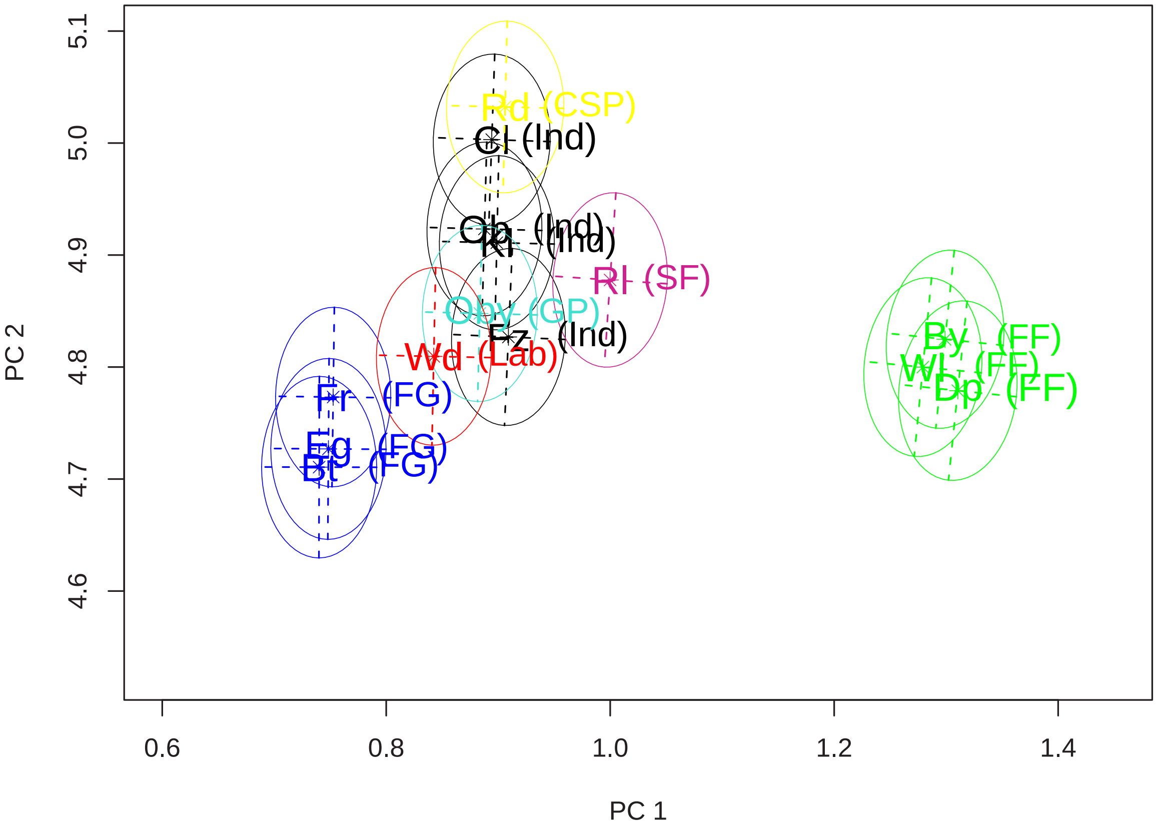


Fig. 8.8: The two dimensional configuration of the Meath candidate means with their associated uncertainty. The candidate initials indicate their posterior mean positions and the ellipses are approximate 95% posterior sets which indicate the uncertainty in the candidate positions. Candidates from different parties are plotted in different colours. The position of each candidate and the ellipses are estimated by 65000 Metropolis-Hastings iterations (post burn-in), thinned after every 100th iteration.

8.3.4 The 1997 Irish Presidential Opinion Polls

The latent space model incorporating the Plackett-Luce model was fitted to the Lansdowne exit poll over the range of dimensions $D = 1, \ldots, 4$. A zero centred Normal proposal density for the voters $N(0, \sigma_v^2)$ and candidates $N(0, \mu_c^2)$ was used throughout. Variance parameters were fixed to be $\sigma_v = 3$ and $\sigma_c = 0.005$ for $D = 1$ and 2, $\sigma_v = 10$ and $\sigma_c = 0.02$ for $D = 3$ and $\sigma_v = 12$ and $\sigma_c = 0.05$ for $D = 4$.

Table 8.8 details the Deviance Information Criterion and Pritchard et al.'s criterion (see Chapter 3) for models fitted to the exit poll data over the range of dimensions $D = 1, \ldots, 4$. These criteria are proposed as dimensionality selection techniques. While these criteria agree on a single dimensional optimal model in this case as has been previously illustrated they often appear to contradict each other. Thus Table 8.9 shows the variance of the data accounted for by each principal component when PCA was applied to configurations of different dimensions. Dimensions $D = 1$ and $D = 2$ appear to summarize the data well. The addition of further dimensions adds little value in terms of the variance of the data explained.

Table 8.8: DIC values and Pritchard et al.'s criterion values for latent space models of dimension $D = 1, \ldots, 4$ fitted to the exit poll data. Entries in bold font indicate the best fitting model according to each criterion. The criteria indicate a single dimensional model fits best.

Dimension	DIC	Pritchard et al.
1	**18447**	**17575**
2	18770	19726
3	18999	33938
4	18946	98719

One Dimensional Results

Figure 8.9 illustrates the relative one dimensional spatial locations of each of the five presidential candidates and their associated uncertainty. Acceptance rates of

Table 8.9: The variance, σ_d^2, captured by each principal component fitted to configurations resulting from the exit opinion poll data, for different dimensions. Principal components analysis was applied to the average candidate configuration only as the main interest of this study lies in the relative locations of the candidates.

	VARIANCES			
DIMENSION	$\boldsymbol{\sigma_1^2}$	$\boldsymbol{\sigma_2^2}$	$\boldsymbol{\sigma_3^2}$	$\boldsymbol{\sigma_4^2}$
1	**1**	-	-	-
2	**0.78**	**0.22**	-	-
3	0.75	0.24	0.01	-
4	0.75	0.24	0.01	0.00

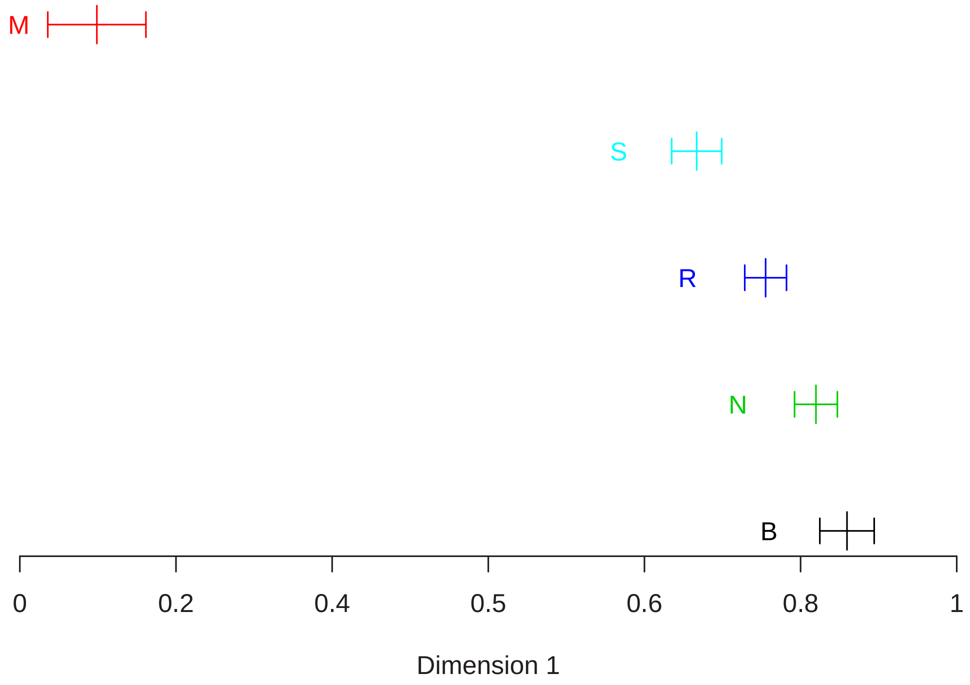


Fig. 8.9: The one dimensional configuration of the 1997 Irish presidential candidate means, averaged over a Metropolis-Hastings algorithm, and their associated uncertainty (indicated by ±2 standard deviation intervals). Each of the presidential candidates as detailed in Table 2.1 are denoted by an abbreviation of their surname.

63% and 84% for voters and candidates respectively were achieved during 27000 iterations (post burn-in) of the Metropolis-Hastings algorithm.

Mary McAleese was elected as President of Ireland and is clearly separated from the other four candidates. McAleese was supported by the current coalition government of Fianna Fáil and the Progressive Democrats. McAleese and Scallon were deemed to be the more conservative candidates and Scallon is situated closest to McAleese among the non-successful candidates. The one dimensional configuration reveals that the electorate partitioned the candidates according to their views on McAleese, with conservative and liberal candidate characteristics also playing a small role.

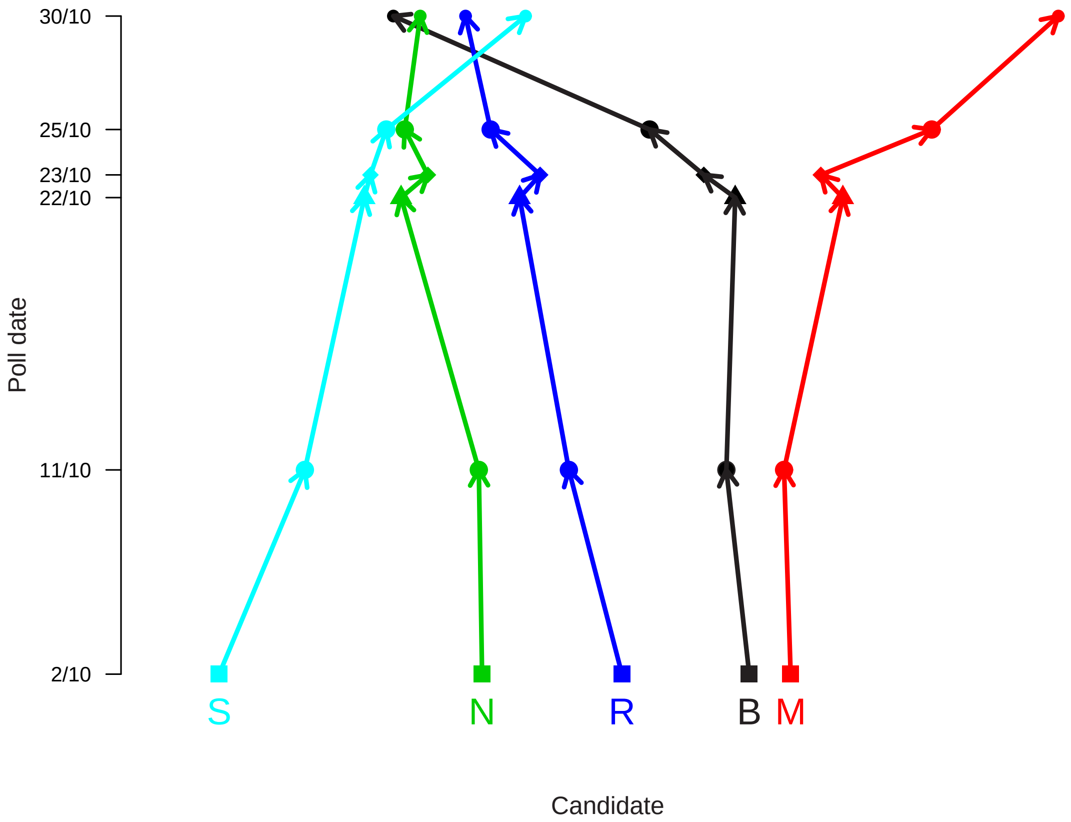


Fig. 8.10: A chronological trace of the spatial movement of each of the presidential candidates in one dimension over the course of six opinion polls conducted during the 1997 presidential campaign. Each shape represents a new poll and each candidate is denoted by their initial.

The one dimensional latent space model was also fitted to six opinion polls conducted during the campaign prior to the 1997 presidential election. Each new shape in Figure 8.10 represents one of the six polls; the arrows between polls indicate the spatial movement of the candidates during the electoral campaign. It is immediately clear that McAleese is viewed very differently to the rest of the candidates and that Scallon moved in the same direction as McAleese as the campaign progressed. It is also apparent that McAleese and Banotti became rival candidates in the later stages of the campaign.

Two Dimensional Results

Figure 8.11 demonstrates the two dimensional relative locations of each of the five presidential candidates estimated by the model. Mean acceptance rates for the two dimensional model were 44% and 87% for voter and candidate locations respectively. The center of each ellipse indicates the posterior mean location of each candidate — the ellipses are approximate 95% posterior sets which indicate the uncertainty in the candidate positions.

The candidate locations are plotted according to their two principal components — the first principal component suggests the electorate are mainly divided on their views on Mary McAleese. McAleese was supported by the large political party, Fianna Fáil who (in coalition with the Progressive Democrats) were in government at that time.

The second principal component suggests party politics are also at play. In the second component McAleese and Banotti (who are supported by the larger political parties) are located at higher position than the other three candidates who were either independent or supported by a smaller political party. Location uncertainty for all candidates is small.

The two dimensional latent space model, incorporating the Plackett-Luce model, was finally fitted to all six opinion polls conducted during the campaign prior to the 1997 presidential election. Each new shape in Figure 8.12 represents one of the six polls; the arrows between polls indicate the spatial movement of the candidates during the electoral campaign. As the campaign progresses McAleese and Banotti move in opposite directions, while Scallon becomes more centrally located. Scallon's popularity increased significantly during the electoral campaign. Roche and Nally follow a similar outward direction but Roche appears to change direction towards the end of the campaign. On the day of the election the final locations of the candidates suggest the largest divide is between the McAleese and Banotti camps, with the profile of the candidate (and in some way their associated political party) appearing as the second influential factor.

8.4 Conclusions

A latent space model, incorporating a Plackett-Luce model, provides good methodology for statistically modelling PR-STV rank data. The latent space aspect of the model gives an interpretable framework for the results of model fitting and the Plackett-Luce model works well in modelling PR-STV data.

The latent configurations suggest that party politics drive general elections in Ireland. Other factors such as the level of a candidate's public profile may also be influential but some of these factors would be confounded with party membership, when it comes to an interpretation of the model's estimates.

When defining the latent space, squared Euclidean distance was implemented as a measure of 'distance' between two members of the space. This distance worked well in the sense that the latent positions found using this distance measure are easily interpreted. Hoff (2005) made use of the inner product as a latent space distance measure and such a method could be implemented in this context.

Principal components analysis selected the optimal dimension of the latent space — the method worked well but is somewhat ad-hoc. Reversible jump Metropolis-Hastings with delayed rejection is an alternative but complicated method of selecting D.

In terms of the Bayesian tools used to fit the model, the random walk proposal worked well in practice but a more sophisticated proposal could be implemented. Also, a basic prior structure was used for the candidate and voter locations, yet a more structured prior on the voters could be employed — for example, a mixture of normals as was used in a social networks context by Handcock et al. (2005) may provide a more suitable prior.

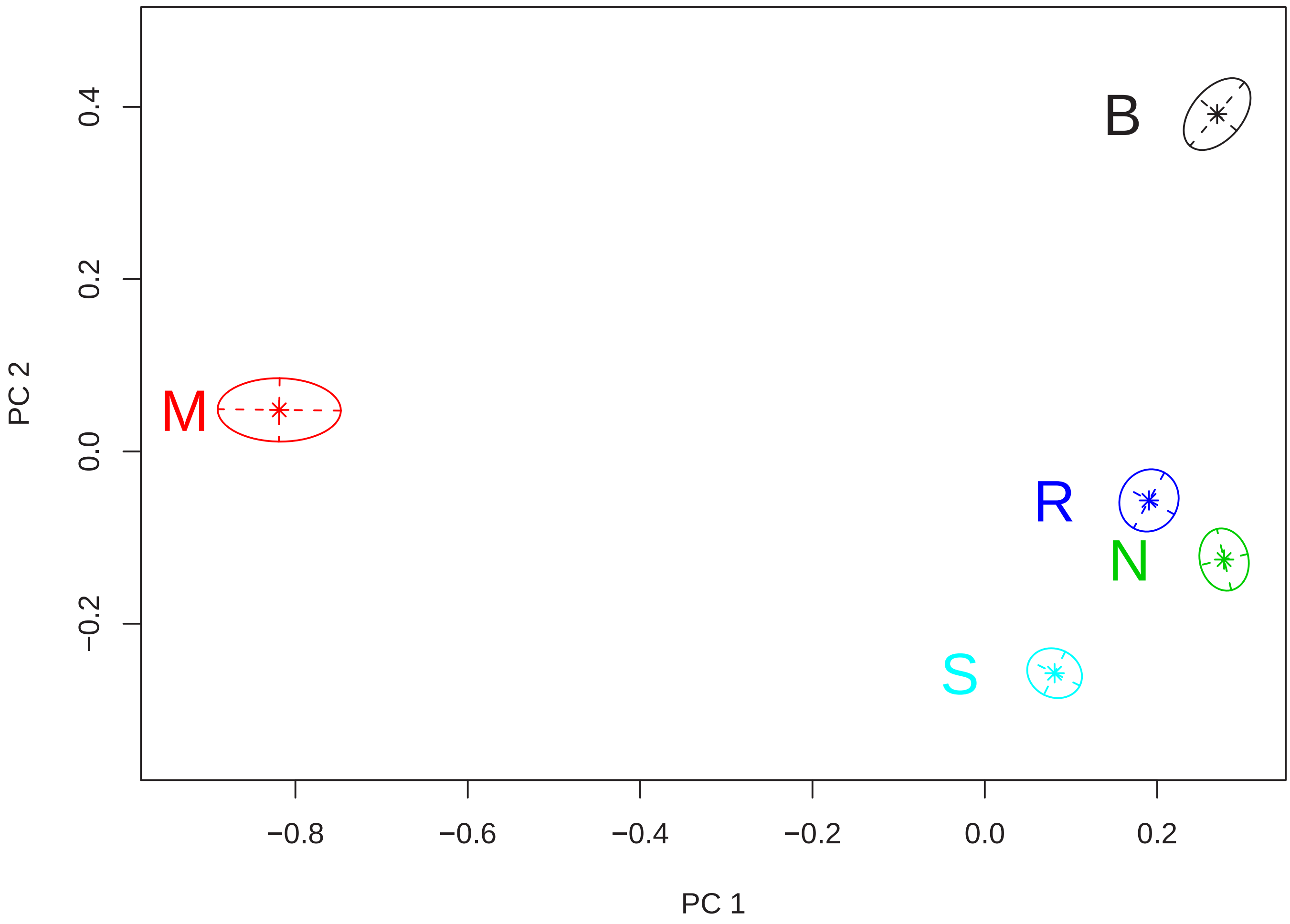


Fig. 8.11: The two dimensional configuration of the 1997 presidential candidate means with their associated uncertainty. The center of each ellipse indicates the posterior mean location of each candidate — the ellipses are approximate 95% posterior sets which indicate the uncertainty in the candidate positions. The position of each candidate and the ellipses are estimated by 27000 Metropolis-Hastings iterations (post burn-in), thinned after every 100th iteration. 87% of proposed candidate locations were accepted.

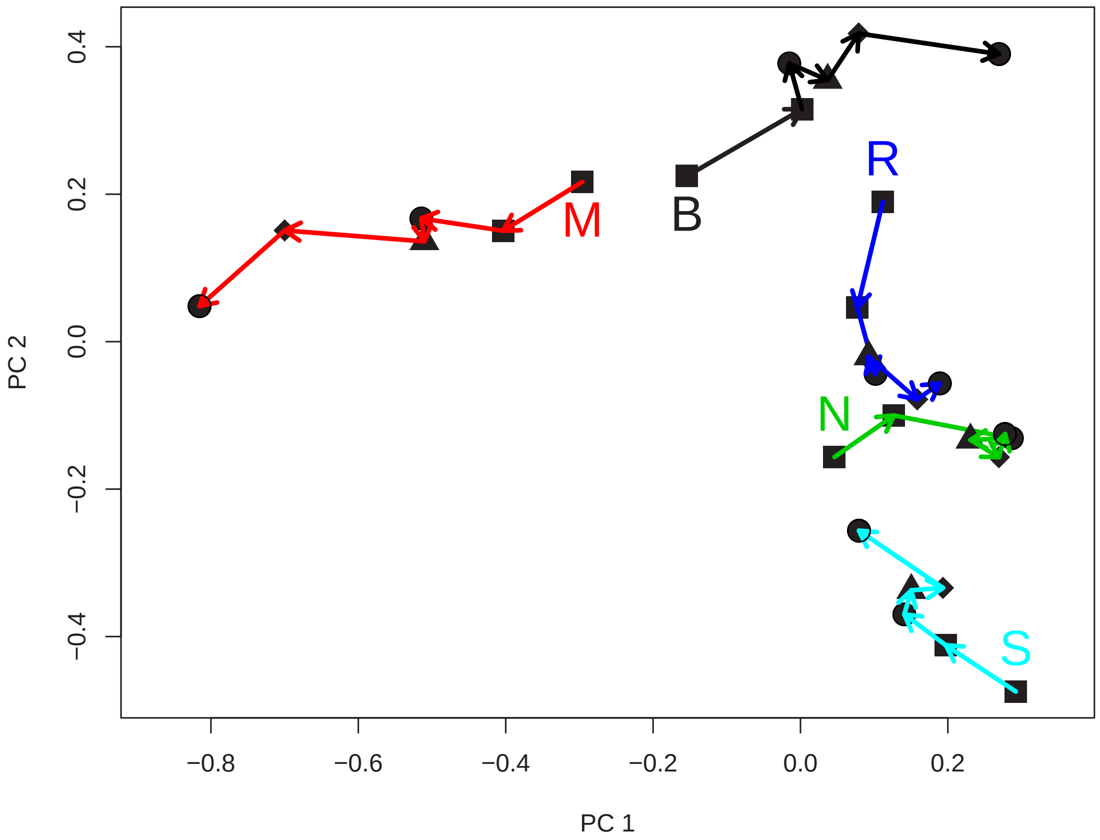


Fig. 8.12: A chronological trace of the spatial movement of each of the presidential candidates in two dimensions over the course of six opinion polls conducted during the 1997 presidential campaign. Each shape represents a new poll and each candidate is denoted by their initial.

Chapter 9

Conclusions and Further Work

9.1 Conclusions

From the analysis conducted it is apparent that many relationships exist between the set of the judges and between the objects they rank within the contexts of Irish third level college applications and Irish elections.

9.1.1 Central Applications Office Data

Analyzing the set of applicants to Irish third level institutions provides evidence of the presence of distinct homogeneous groups within the population. The resulting groupings reveal that applicants generally appear to be driven by their vocational interests as discipline emerges as the defining characteristic of applicant groups. For example, a group of applicants who give high preference to courses from the field of business and marketing emerges as does a group who applied for engineering courses.

The geographical position of the institution to which an applicant applies also transpires to have a significant influence on course choice. Groups of applicants who give high preference to courses within a specific geographical region of Ireland are frequent. Some applicants apply to Cork based institutions, some to Galway-Limerick based institutions and some are divided by their opinions on the same type of course but offered by institutions within and outside of Dublin based colleges.

Crucially however, some weight is added to the CAO system detractor's arguments who claim applicants are influenced by the prestige of some degree programs. A deeper analysis of the revealed groups highlights a subtle influence of the points on the applicants choices. It seems the points status of courses has an effect on health science applicants more so than on applicants in other components within the population.

Finally, a separate analysis of the male and female data suggests applicants of different gender have different course choice behaviours. Stereotypical differences between the two genders are apparent – female applicants who apply for courses within the fields of social science, art and design, music and education are distinctly separated whereas female applicants with an interest in engineering and computer science are grouped together. The largest group of male applicants involves those who applied for construction studies courses; such a group does not appear as a distinct component in the female results.

9.1.2 Irish Voting Data

Analysis of Irish voting data also provides evidence of relationships between both the members of the Irish electorate and between the electoral candidates.

It is observed that there is strong political party support in Irish general elections, because voters tend to give their high preferences to candidates from the same political party or to parties of a particular persuasion. Traditional coalition political parties are often grouped together or voters appear to have large probability of giving such coalition candidates high preferences.

There is also evidence however of candidate orientated voters within the Irish electorate in both the presidential and general elections. Within the 1997 Irish presidential election it is apparent that the candidates with good public profiles have high levels of support. Similarly in general elections, candidates with relatively unpopular party political affiliations can have large levels of support due to the public's perception of their personality. Evidence of the 'candidate centered but party wrapped' theory of Irish elections presented by Marsh (2000) is supported by the analysis presented here.

9.2 Further Work

As the research presented progressed many more questions and potential applications for alternative techniques arose.

1. In Chapters 4 and 5, when estimating standard errors within the context of the EM algorithm an approximation of the covariance matrix was computed with little extra overhead. The technique used however is only applicable in the case where the data is independently and identically distributed. In the general case, Meng and Rubin (1998, 1991) provide a method of estimating the asymptotic covariance matrix using only code from the EM algorithm itself, standard matrix calculations code and code to compute the complete data covariance matrix. Their method is based on the idea that the rate of convergence of the EM algorithm is governed by the amount of missing data. Using this, they find the increased variability due to the missing data and add it to the complete data variance-covariance matrix. They term this the supplemented EM algorithm. The implementation of this EM algorithm for rank data would broaden the range of cases the current framework could deal with. Also with reference to the approximation of standard errors, the use of a block technique in problems involving large data sets would improve computation time of the EM algorithm.

2. When fitting mixed membership models (Chapter 6) only the Deviance Information Criterion (DIC) was examined as a method of model selection. Other methods need to be examined. Airoldi et al. (2006) recently discussed methods of model selection for mixed membership models. Also Raftery et al. (2006) introduced the AICM (Akaike Information Criterion Monte (Carlo)) and BICM (Bayesian Information Criterion Monte (Carlo)) which are derived through the estimation of the harmonic mean estimator. Difficulties may arise when specifying the number of parameters within the BICM and AICM criteria but such techniques should be examined. The BICM and AICM should also be examined as a method of selecting the dimensionality of the latent space in latent space models. Gormley and Murphy (2006*a*) details some preliminary work in this area.

3. In both the application of mixed membership models and the application of the latent space model rather basic priors were specified. By fitting mixtures of rank data models to the data set and then basing the mixed membership prior distribution and hyperparameters on these a more informed prior distribution would be provided. Moreover, further model accuracy could be attained by imposing a hierarchical framework — a hyperprior could be introduced for the prior parameters of the mixed membership and support parameter priors respectively. Erosheva (2003) employed such hierarchical priors.

4. In Chapter 7 when fitting mixtures-of-experts models improved variable selection should be performed. While the use of the Bayesian Information Criterion appears to have highlighted consistent and informative covariates techniques with a more theoretically sound comparison of the models should be conducted. Raftery and Dean (2006) introduced a model based variable selection technique which could be adapted to suit the mixtures-of-experts model. Also, similar to Peng et al. (1996) who fitted the mixtures-of-experts and hierarchical mixtures-of-experts models within a Bayesian framework, it would be interesting to fit the rank data model version within the Bayesian paradigm.

5. When defining the latent space, squared Euclidean distance was implemented as a measure of 'distance' between two members of the space. Hoff (2005) made use of the inner product as a latent space distance measure and such a method could be implemented in this context. Also within the context of latent space models choosing the dimensionality of the unobserved space provided problems. Reversible jump MCMC methods could provide a complicated but alternative dimensionality selection technique.

6. An alternative application of such rank data models was suggested by Murphy and Gormley (2006) in the context of modelling pollen counts. Due to relatively sparse data the modelling of pollen abundance in terms of a partial ranking was suggested. This approach could offer an alternative to the presence/absence approach taken by Haslett et al. (2006).

7. Nested choice models (McFadden, 1978; Train, 2003) could be used to model the choice process which results in a ranking of college degree programs or

in a list of candidates on an election ballot form. Such models assume that choices are made in a hierarchical manner; the judges begin with coarse categories which are refined during the ranking process. For example, third level applicants may choose a field of study and then select courses within that field; voters may select political parties and then candidates within parties. Nested choice models could be extended into nested ranking models using a multi-stage ranking model approach. Also, the nesting structure in the set of objects is not often known. Product partition models (Hartigan, 1990; Barry and Hartigan, 1992; Crowley, 1997) could be employed to provide a probability distribution on possible nesting structures.

Appendix A

Data Sources

- The various 1997 Irish presidential election opinion poll data sets were collected by the three companies: Lansdowne Market Research, Irish Marketing Surveys (IMS), and Market Research Bureau of Ireland (MRBI). These data sets are available through the Irish Elections Data Archive

  `http://www.tcd.ie/Political_Science/elections/elections.html`

  and the Irish Opinion Poll Archive

  `http://www.tcd.ie/Political_Science/cgi/`

  which are maintained by Professor Michael Marsh in the Department of Political Science, Trinity College Dublin, Ireland.

- The voting data from the Dublin North and Dublin West constituencies are available from the constituency returning officer's web page

  `http://www.dublincountyreturningofficer.com`.

- The voting data from the Meath constituency are available from the Meath county council web page

  `http://www.meath.ie/election.html`.

Appendix B

Propositions

Proposition 1 (Dampening Entropy) *Let* $\underline{p} = (p_1, p_2, \ldots, p_N)$ *be such that* $p_j > 0$ *for all* j *and* $\sum_{j=1}^{N} p_j = 1$ *and let* $0 \leq \alpha \leq 1$. *Let*

$$q(\alpha) = (q_1, q_2, \ldots, q_N) = \left(\frac{p_1^\alpha}{\sum_{j=1}^N p_j^\alpha}, \frac{p_2^\alpha}{\sum_{j=1}^N p_j^\alpha}, \ldots, \frac{p_N^\alpha}{\sum_{j=1}^N p_j^\alpha} \right).$$

Then, the $\mathrm{Entropy}[q(\alpha)] = \mathbf{E}[q(\alpha)] = -\sum_{j=1}^{N} q_j(\alpha) \log q_j(\alpha)$ *is a decreasing function of* α.

Proof: We have:

$$\begin{aligned}
\frac{\partial \mathbf{E}}{\partial \alpha} &= -\sum_{j=1}^{N} q_j' [1 + \log q_j] \\
&= -\sum_{j=1}^{N} \left[\frac{p_j^\alpha}{\sum_{l=1}^N p_l^\alpha} \left\{ \log p_j - \frac{\sum_{l=1}^N p_l^\alpha \log p_l}{\sum_{l=1}^N p_l^\alpha} \right\} \right] \left[1 + \alpha \log p_j - \log \sum_{l=1}^{N} p_l^\alpha \right] \\
&= -\sum_{j=1}^{N} \Bigg[q_j \log p_j + \alpha q_j \{\log p_j\}^2 - q_j \log p_j \log \sum_{l=1}^{N} p_l^\alpha \\
&\quad -q_j \sum_{l=1}^{N} q_l \log p_l - \alpha q_j \log p_j \sum_{l=1}^{N} q_l \log p_l + q_j \log \sum_{l=1}^{N} p_l^\alpha \sum_{l=1}^{N} q_l \log p_l \Bigg] \\
&= -\alpha \left[\sum_{j=1}^{N} q_j \{\log p_j\}^2 - \left\{ \sum_{j=1}^{N} q_j \log p_j \right\}^2 \right] \\
&\leq 0
\end{aligned}$$

by the Cauchy-Schwarz inequality. ▮

Bibliography

Airoldi, E. A., Fienberg, S. E., Joutard, C. and Love, T. M. (2006), Discovering Latent Patterns with Hierarchical Bayesian Mixed-Membership Models, Technical Report CMU-ML-06-101, School of Computer Science, Carnegie Mellon University.

Aitkin, M., Anderson, D. and Hinde, J. (1981), 'Statistical Modelling of Data on Teaching Styles (with Discussion)', *Journal of the Royal Statistical Society, Series A: General* **144**, 419–461.

Akaike, H. (1973), 'Information theory and an extension to the maximum likelihood principle', *Second International Symposium on on Information Theory* pp. 267–281.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**.

Banfield, J. D. and Raftery, A. E. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics* **49**(3), 803–821.

Barry, D. and Hartigan, J. A. (1992), 'Product partition models for change point problems', *Annals of Statistics* **20**, 260–279.

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997), 'Inference in model-based cluster analysis', *Statistics and Computing* **7**, 1–10.

Benter, W. (1994), Computer-based horse race handicapping and wagering systems: A report, *in* W. T. Ziemba, V. S. Lo and D. B. Haush, eds, 'Efficiency of Racetrack Betting Markets', Academic Press, San Diego and London, pp. 183–198.

Biernacki, C., Celeux, G. and Govaert, G. (2000), 'Assessing a mixture model for clustering with the integrated completed likelihood', *IEEE Transactions on Pattern Analysis & Machine Intelligence* **22**(7), 719–725.

Blais, A. (1991), 'The debate over electoral systems', *International Political Science Review* **12**(3), 239–260.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B. (1994), 'The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family', *Annals of the Institute of Statistical Mathematics* **46**, 373–388.

Bowler, S. and Farrell, D. M. (1991), 'Voter Behavior Under STV-PR: Solving the Puzzle of the Irish Party System', *Political Behavior* **13**(4), 303–320.

Bradley, R. A. and Terry, M. E. (1952), 'Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons', *Biometrika* **39**, 324–345.

Bradlow, E. T. and Fader, P. S. (2001), 'A Bayesian Lifetime Model for the "Hot 100" *Billboard* Songs', *Journal of the American Statistical Association* **96**, 368–381.

Brams, S. J. and Fishburn, P. C. (1984), Some Logical Defects of the Single Tranferable Vote, *in* A. Lijphart and B. Grofman, eds, 'Choosing an Electoral System: Issues and Alternatives', Praeger, New York, pp. 147–151.

Carlin, B. P. and Louis, T. A. (2000), *Bayes and empirical bayes methods for data analysis.*, 2nd edn, Chapman & Hall, New York.

Casella, G. and Berger, R. L. (1990), *Statistical Inference*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

Casella, G. and George, E. I. (1992), 'Explaining the Gibbs Sampler', *The American Statistician* **46**(3), 167–174.

Celeux, G. and Diebolt, J. (1985), 'The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem', *Computational Statistics Quarterly* pp. 73–82.

Celeux, G., Hurn, M. and Robert, C. P. (2000), 'Computational and inferential difficulties with mixture posterior distributions', *Journal of the American Statistical Association* **95**(451), 957–970.

Chib, S. and Greenberg, E. (1995), 'Understanding the Metropolis-Hastings Algorithm', *The American Statistician* **49**, 327–335.

Clancy, P. (1995), *College Entry In Focus: A Fourth National Survey of Access to Higher Education*, Higher Education Authority, Dublin, Ireland.

Coakley, J. and Gallagher, M. (1999), *Politics in the Republic of Ireland*, 3rd edn, Routledge in association with PSAI Press, London.

Critchlow, D. E. (1985), *Metric methods for analyzing partially ranked data*, Lecture Notes in Statistics, 34, Springer-Verlag, Berlin.

Crowley, E. M. (1997), 'Product partition models for normal means', *Journal of the American Statistical Association* **92**(437), 192–198.

Dasgupta, A. and Raftery, A. (1998), 'Detecting features in spatial point processes with clutter via model-based clustering.', *Journal of the American Statistical Association* **93**, 294—302.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B* **39**(1), 1–38. With discussion.

Diaconis, P. (1988), *Group representations in probability and statistics*, Institute of Mathematical Statistics, Hayward, CA.

Dobson, A. J., ed. (2002), *An introduction to generalized linear models*, second edn, Chapman and Hall, London.

Erosheva, E. (2002), Grade of membership and latent structure models with application to disability survey data., PhD thesis, Department of Statistics, Carnegie Mellon University.

Erosheva, E. A. (2003), Bayesian Estimation of the Grade of Membership Model, *in* 'Bayesian Statistics, 7', Oxford Univ. Press, UK.

Fienberg, S. E. and Larntz, K. (1976), 'Log linear representation for paired and multiple comparisons models', *Biometrika* **63**(2), 245–254.

Fligner, M. A. and Verducci, J. S., eds (1993), *Probability models and statistical analyses for ranking data*, Springer-Verlag, New York. Papers from the conference held at the University of Massachusetts, Amherst, Massachusetts, June 8–13, 1990.

Fraley, C. and Raftery, A. E. (1998), 'How many clusters? Which clustering method? - Answers via Model-Based Cluster Analysis', *Computer Journal* **41**, 578–588.

Fraley, C. and Raftery, A. E. (1999), 'Mclust: Software for model-based clustering', *Journal of Classification* **16**, 297–306.

Fraley, C. and Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American Statistical Association* **97**(458), 611–612.

Gelfand, A. E. and Smith, A. F. M. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**(410), 398–409.

Geman, S. and Geman, D. (1984), 'Stochastic relaxation, Gibbs Distributions and the Bayesian Restoration of Images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., eds (1996), *Markov chain Monte Carlo in practice*, Chapman & Hall, London.

Gormley, I. C. and Murphy, T. (2006*a*), 'Discussion of Raftery et al.'Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity.", *Bayesian Statistics 8* .

Gormley, I. C. and Murphy, T. B. (2005), Exploring Heterogeneity in Irish Voting Data: A Mixture Modelling Approach., Technical Report 05/09, Department of Statistics, Trinity College Dublin.

Gormley, I. C. and Murphy, T. B. (2006*b*), 'A latent space model for rank data', Lecture Notes in Computer Science, Statistical Network Analysis: Models, Issues, and New Directions, Springer Verlag.

Gormley, I. C. and Murphy, T. B. (2006*c*), 'Analysis of Irish third-level college applications data', *Journal of the Royal Statistical Society, Series A* **169**(2), 361—379.

Graves, T., Reese, C. S. and Fitzgerald, M. (2003), 'Hierarchical Models for Permutations: Analysis of Auto Racing Results', *Journal of the American Statistical Association* **98**, 282–3291.

Handcock, M., Raftery, A. E. and Tantrum, J. (2005), Model-based clustering for social networks, Technical Report 482, Department of Statistics, University of Washington, Seattle.

Hartigan, J. A. (1990), 'Partition models', *Communications in Statistics, Part A – Theory and Methods* **19**, 2745–2756.

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S., Allen, J. R. M., Huntley, B. and Mitchell, F. J. G. (2006), 'Bayesian palaeoclimate reconstruction (with discussion)', *Journal of the Royal Statistical Society, Series A.* **169**(3), 395–438.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Hoff, P. D. (2005), 'Bilinear Mixed-Effects Models for Dyadic Data', *Journal of the American Statistical Association* **100**, 286–295.

Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002), 'Latent Space Approaches to Social Network Analysis', *Journal of the American Statistical Association* **97**, 1090–1098.

Hunter, D. R. (2004), 'MM algorithms for generalized Bradley-Terry models', *The Annals of Statistics* **32**(1), 384–406.

Hunter, D. R. and Lange, K. (2004), 'A tutorial on MM algorithms', *The American Statistician* **58**(1), 30–37.

Hyland, A. (1999), *Commission on the Points System: Final Report and Recommendations*, Commission on the Points System Reports, The Stationery Office, Dublin, Ireland.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991), 'Adaptive mixture of local experts', *Neural Computation* **3**(1), 79–87.

Johnson, V. E., Deaner, R. O. and van Schaik, C. P. (2002), 'Bayesian Analysis of Rank Data With Application to Primate Intelligence Experiments', *Journal of the American Statistical Association* **97**, 8–17.

Jordan, M. I. and Jacobs, R. A. (1994), 'Hierarchical mixtures of experts and the EM algorithm', *Neural Computation* **6**, 181–214.

Kass, R. E. and Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**, 773–795.

Katz, R. S. (1984), The single transferable vote and proportional representation, *in* A. Lijphart and B. Grofman, eds, 'Choosing an Electoral System: Issues and Alternatives', Praeger, New York, pp. 135–145.

Keribin, C. (1998), 'Estimation consistante de l'ordre de modèles de mélange', *C. R. Acad. Sci. Paris Sér. I Math.* **326**(2), 243–248.

Keribin, C. (2000), 'Consistent estimation of the order of mixture models', *Sankhyā Ser. A* **62**(1), 49–66.

Krzanowski, W. J. (1988), *Principles of Multivariate Analysis: A User's Perspective*, Clarendon Press.

Kullback, S. and Leibler, R. (1951), 'On information and sufficiency', *Annals of Mathematical Statistics* **22**, 79–86.

Lange, K., Hunter, D. R. and Yang, I. (2000), 'Optimization transfer using surrogate objective functions', *Journal of Computational and Graphical Statistics* **9**(1), 1–59. With discussion, and a rejoinder by Hunter and Lange.

Laver, M. (2004), 'Analysing structure of party preference in electronic voting data', *Party Politics* **10**, 521–541.

Lee, P. M. (2004), *Bayesian statistics*, third edn, Arnold, London. An introduction.

Leroux, B. G. (1992), 'Consistent estimation of a mixing distribution', *The Annals of Statistics* **20**(3), 1350–1360.

Lindsay, B. (1995), *Mixture Models: Theory, Geometry and Applications*, Institute of Mathematical Statistics, Hayward, CA.

Lynch, K., Brannick, T., Clancy, P. and Drudy, S. (1999), *Points and Performance in Higher Education: A Study of the Predictive Validity of the Points System*, number 4 *in* 'Commission on the Points System Research Papers', The Stationery Office, Dublin, Ireland.

Manton, K., Woodbury, M. and Tolley, H. (1994), *Statistical Applications Using Fuzzy Sets.*, Wiley-Interscience.

Marden, J. I. (1995), *Analyzing and modeling rank data*, Chapman & Hall, London.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979), *Multivariate analysis*, Academic Press [Harcourt Brace Jovanovich Publishers], London. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

Marsh, M. (1999), The Making of the Eighth President, *in* M. Marsh and P. Mitchell, eds, 'How Ireland Voted 1997', Westview and PSAI Press, Boulder, CO, pp. 215–242.

Marsh, M. (2000), Candidate centered but party wrapped: Campaigning in Ireland under STV, *in* S. Bowler and B. Grofman, eds, 'Elections in Australia, Ireland, and Malta under the Single Transferable Vote', The University of Michigan Press, Ann Arbor, MI, pp. 114–130.

McCullagh, P. and Nelder, J. (1983), *Generalized Linear Models*, Chapman and Hall, London.

McFadden, D. G. (1978), 'Modelling the choice of residential location', *Spatial Interaction Theory and Planning Models* pp. 75–96.

McLachlan, G. J. and Krishnan, T. (1997), *The EM algorithm and extensions*, John Wiley & Sons Inc., New York.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture models*, John Wiley & Sons, New York.

Meilijson, I. (1989), 'A fast improvement to the EM algorithm on its own terms', *Journal of the Royal Statistical Society, Series B* **51**(2), 127–138.

Meng, X.-L. and Rubin, D. B. (1991), 'Using the EM to obtain asymptotic variance-covariance matrices: the SEM algorithm', *Journal of the American Statistical Association* **86**, 899–909.

Meng, X.-L. and Rubin, D. B. (1993), 'Maximum likelihood estimation via the ECM algorithm: a general framework', *Biometrika* **80**(2), 267–278.

Meng, X.-L. and Rubin, D. B. (1998), 'Obtaining asymptotic variance-covariance matrices for missing-data problems using the EM algorithm', *Proceedings of the American Statistical Association (Statistical Computing Section)* pp. 140–144.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equations of state calculations by fast computing machine', *Journal of Chemical Physics* **21**, 1087–1091.

Murphy, T. B. and Martin, D. (2003), 'Mixtures of distance-based models for ranking data', *Computational Statistics and Data Analysis* **41**(3–4), 645–655.

Murphy, T. and Gormley, I. C. (2006), 'Discussion of Haslett et al.'Bayesian paleoclimate reconstruction.", *Journal of the Royal Statistical Society, Series A* **169**(3), 434–435.

O'Hagan, A. and Forster, J. (2004), *Kendall's Advanced Theory of Statistics: Volume 2B Bayesian Inference*, second edn, Arnold, London, UK.

Peng, F., Jacobs, R. A. and Tanner, M. A. (1996), 'Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With and Application to Speech Recognition', *Journal of the American Statistical Association* **91**(435), 953—960.

Plackett, R. L. (1975), 'The analysis of permutations', *Applied Statistics* **24**(2), 193–202.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1996), *Numerical Recipes in C: The Art of Scientific Computing Second Edition.*, Cambridge University Press, Cambridge.

Pritchard, J. K., Stephens, M. and Peter, D. (2000), 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics* **155**, 945–959.

Raftery, A. E. and Dean, N. (2006), 'Variable selection for model-based clustering', *Journal of the American Statistical Association* **101**, 168–178.

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N. (2006), Estimating the Intergrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity., Technical Report 499, Department of Statistics, University of Washington, Seattle, Washington, USA.

Richardson, S. and Green, P. J. (1997), 'On Bayesian analysis of mixtures with an unknown number of components', *Journal of the Royal Statistical Society, Series B* **59**, 731–758.

Roberts, G., Gelman, A. and Gilks, W. (1994), Weak convergence and optimal scaling of random walk Metropolis algorithms, Technical Report 94.16, Statistical Laboratory, University of Cambridge.

Rosén, B. (1972), 'Asymptotic Theory for Successive Sampling with Varying Probabilities Without Replacement, I', *The Annals of Mathematical Statistics* **43**(2), 373–397.

Schwartz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**, 461–464.

Sinnott, R. (1995), *Irish voters decide: Voting behaviour in elections and referendums since 1918*, Manchester University Press, Manchester.

Sinnott, R. (1999), The electoral system, *in* J. Coakley and M. Gallagher, eds, 'Politics in the Republic of Ireland', 3rd edn, Routledge & PSAI Press, London, pp. 99–126.

Smyth, P. (2000), 'Model selection for probabilisitc clustering using cross-validated likelihood', *Statistics and Computing* **9**, 63—72.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society, Series B* **64**(4), 583–639.

Stephens, M. (2000), 'Dealing with label-switching in mixture models', *Journal of the Royal Statistical Society, Series B* **62**(4), 795–810.

Tierney, L. (1994), 'Markov Chains for Exploring Posterior Distributions', *The Annals of Statistics* **22**, 1701–1762.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985), *Statistical analysis of finite mixture distributions*, Wiley, Chichester.

Train, K. E. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge.

Tuohy, D. (1998), *Demand for Third-Level Places*, number 1 *in* 'Commission on the Points System Research Papers', The Stationery Office, Dublin, Ireland.

Wasserman, L. (2004), *All of Statistics. A Concise Course in Statistical Inference.*, Springer-Verlag, New York.