Explanatory statistical modelling of influences of demographic experience on political identity

Lachlann McArthur

July 23, 2019

Thesis submitted for the degree of
Master of Philosophy
in
Statistics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

Contents

List of Tables

List of Figures

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: . . . .   Date: . . . .

Acknowledgements

I would like to thank my supervisors—Dr Jonathan Tuke, Dr Melissa Humphries, and Professor Nigel Bean—who have supported me relentlessly for the last two years. It's a wonder they not only didn't get tired of meeting with me every week, but right to the end, fought over who got to read my work first. I couldn't have asked for a more affable, encouraging and talented team.

I would like to thank my fellow students and staff at the School of Mathematical Sciences at the University of Adelaide, as well as the ARC Centre for Excellence in Mathematical and Statistical Frontiers, for showing me how vibrant and diverse the Australian mathematics community can be.

And I am forever indebted to Caro Mader, my mum, for always being around when I've needed to talk about anything, even that time when what I needed to talk about was maths; Rod McArthur, my dad, for telling me that my value is not determined by my achievements; Pip Wayte, for reminding me what's important (it's Twin Peaks and making every recipe in Simple), for agreeing to proofread 200 pages in a new language in two days, and for teaching me that there's nothing in the world better than being a dork surrounded by dorks; Justin McArthur, for letting me try to explain every single chapter of this thesis to him, and for helping me out with the Saturday Paper cryptic every weekend; Huy Nguyen, for sometimes letting me beat him at Ticket to Ride; ~~Sebastian Nguyen, who effortlessly brings me joy and so deserves mention despite protesting being acknowledged here~~; and Vinay Athreya, Rupert Piccoli and Edward Ramsey, for giving me confidence when I didn't think I deserved it. All these people have shaped me into the person that wrote this thesis. I'm incredibly lucky that the Venn diagram of my friends and my family is a circle.

Abstract

This thesis seeks to create, and relate, holistic depictions of demographic and political identity, using data from the Australian National University's 2016 Australian Election Study. We thus create new spectra of identity, and produce a multivariate model to explore demographic influences on political ideology. From a statistical perspective, we build a foundation for, and expand upon, techniques for model selection. From a political science perspective, we create a modern, data-driven Australian political spectrum, produce key findings on how this spectrum is influenced by demography, and build a stronger understanding of the ways political thought can diverge from expectations.

We first note that the 2016 Australian Election Study is too complicated to model directly. Instead, we seek simple representations of demographic and political identities. After exploring the literature on mathematical ways of reducing the dimensions of variables, we produce a new spectrum of political ideology for Australia, as well as a new spectrum of demographic identity. The spectra are mathematically designed to be comprised of axes representing the issues that most unite, and most divide, Australians. Our new political spectrum is interpreted in the context of current conceptions of political thought.

We explore models to connect our demographic and political spectra, with the goal of explaining relationships between them in a clear and concise manner. We do not attempt to make predictions about individuals, but rather, explain relationships that exist in the population at large. We seek to build a multivariate model, to describe all dimensions of our new political spectrum simultaneously. We introduce, construct, and prove results relating to four candidate models, each of which elucidates key relationships in subtly different ways. We explain that a traditionally-used model to predict multiple outcomes simultaneously, the multivariate regression model, makes assumptions about the data that cannot be justified.

We develop new tools for comparing models. In our model selection process, we put emphasis on the models' errors; it is dangerous to ignore the ways in which our expectations might be wrong. Our tools for comparing the four candidate models thus place emphasis on selecting a model that is most accurate when it comes to its error distribution's assumptions. Our tools are placed in the context of the historical literature on the issue, as well as in conjunction with other criteria by which a model could be selected.

Our model's results are interpreted in a political science context. We explore, with reference to other research in the area, emergent associations between demographic and political identity. These include the positive relationships between

socio-economic status and education, and social inclusivity, and between stage of life and trust in authority, as well as the negative relationships between stage of life and social inclusivity, and socio-economic status and high social spending. We discuss the relationships between different political views that persist after accounting for demographic influences. These associations are of particular relevance in light of trends towards populist movements around the world, with evidence of low trust in 'elites' being common especially among a small cluster of socially non-inclusive people.

With a process established for relating underlying constructs in large and complex surveys, our methodologies have the capacity to be implemented to surveys of varying geographical and temporal origin. This leads to two paths for future analysis: questioning how the associations between demographic and political identity have changed over time in Australia; and questioning how these associations differ internationally.

Chapter 1

Introduction

For a representative democracy to succeed, decision-makers should understand the values and beliefs of their citizens. Knowledge of the lines along which opinions diverge is thus vital for adequate representation. This thesis seeks to describe how political opinions diverge among different groups of Australian voters, through the novel construction of new spectra of identity. We seek to relate Australians' lived experiences and their views on political issues. At a time of perceived increased political polarisation [81, 116, 133], it is more important than ever to study what unites Australians, and what divides us.

To draw out relationships, we seek to create holistic depictions of demographic and political identity, using data from the 2016 Australian Election Study [121]. We then seek to build a statistical model to see what demographic influences exist on political ideology. We first note that the 2016 Australian Election Study, which comprises 32 demographic questions, and 103 political questions—that is, questions relating to political issues, rather than party politics—is too complicated to model directly. If we try to predict each political response individually, we would produce a model with 103 response variables. This is too many for any holistic analysis. Instead, we seek a parsimonious understanding of political ideology, which requires us to replace the observed political variables with a simpler political spectrum.

Construction of political spectra has been done before. But this thesis produces a political spectrum using data-driven, unsupervised machine learning techniques in an Australian context, and with recent data. The machine learning techniques used are specifically chosen to highlight the dimensions along which Australians' opinions differ. This means we are able to capture as much information as possible, in as concise a way as possible. Vitally, we produce a democratic political spectrum—one in which the axes are drawn to represent how Australians really think, not just how researchers or the media suppose Australians think.

This thesis further adds to the political science literature in that we elucidate relationships between demographic experience and political ideology. We highlight specific directions in which the lived experiences of communities within Australia lead to general trends in political thought. These relationships are modelled, in a statistically-justified manner, between underlying variables affecting demographic identity, and the underlying variables affecting political ideology. As such, we draw

out the key demographic trends in political ideology, not just realisations of them.

It has been widely acknowledged that single-issue political opinion polling is discourse-dependent [39, 157]. In other words, when polling political issues, there are so many ways to ask a question that results become strongly influenced by phrasing of the question. For example, with respect to immigration, the 2016 Australian Election Study asks a number of questions. These include, among others (see Appendix A):

- "What do you think is the best way to handle the processing and resettlement of asylum seekers who come by boat and manage to reach Australian waters?";

- "Do you think the number of immigrants allowed into Australia nowadays should be reduced or increased?";

- "How much do you agree or disagree with each of the following statements? Immigrants are generally good for Australia's economy"; and

- "How much do you agree or disagree with each of the following statements? Immigrants take jobs away from people who are born in Australia".

Each question places emphasis on, and makes distinctions about, different aspects of immigration policy. Not all participants who answer one question in a particular way, will also align on all the other questions. But all these questions are representative of a respondent's underlying belief when it comes to the value of immigration. By modelling underlying beliefs, rather than directly explaining their noisy realisations, we are able to average out discursive influences on our results, as well as produce a more holistic model. We are able to cut past the subjectivity of particular wordings in favour of the true relationships of interest.

Along the way, we notice that political modelling is ingrained with substantial error. We cannot make, and we do not seek to make, specific predictions about individuals based on their demographic identity. A specific goal of this thesis is to grapple with the size and scope of the region in which political opinions exist, rather than just relying on our model's conditional expectation of an individual's political ideology. We desire to produce a model that is best able to describe the error in making political predictions based upon demographic identity. This represents a novel approach in statistical modelling; while traditionally, models are sought to reduce the size of their errors, we first produce a range of models that reduce the size of the error, and then seek the model among them that best explain what distributions of errors might occur. We thus develop new tools for selecting the candidate model that best adheres to its error assumptions. This relies upon a new combination of model selection techniques, and requires us to further develop the literature on non-nested model selection.

Using these new techniques, alongside the wide-ranging, large-sample, high-dimensional 2016 Australian Election Study data set, we are able to understand not just the ways in which political opinion is influenced by demographic experience, but the ways in which political opinion diverges, once demographic experience has been accounted for. We explain these distributions in light of trends towards populist

movements around the world, movements centred around low trust in authority; we provide new insight into clusters of populist ideology.

This work thus provides five outcomes of note in a political and statistical context. These are:

1. We produce a current, data driven, statistically-justified, reproducible political spectrum for Australia.

2. We introduce four candidate explanations of relationships between underlying demographic experience and underlying political ideology in Australia.

3. We introduce, develop and explain techniques for non-nested model selection between multiple candidate models on the basis of unconventional criteria.

4. We explain key relationships between underlying demographic experience and underlying political ideology in Australia, and the consequences of this in light of the existing social science literature.

5. We explain key relationships within political ideology in Australia, once demographic experience has been accounted for, and the consequences of this in light of the existing social science literature.

Section 1.1 now outlines how this thesis will go about achieving these outcomes.

1.1 Outline of chapters

Chapter 2: The Australian Election Study

Chapter 2 describes the 2016 Australian Election Study ('AES'), the data that forms the basis for our analysis. The Australian Election Study is undertaken shortly after each Australian federal election, to survey potential voters' background, values and engagement in the election campaign. The 2016 survey, conducted by Ian McAllister, Clive Bean, Rachel Kay Gibson, and Toni Makkai, was designed to coincide with the 2 July 2016 federal election. The data set features 322 variables, 302 of which are survey questions asked of the 2818 respondents. The respondents are people who elected to participate, having been selected by a random draw from the Australian electoral roll, stratified by state. This random sample was supplemented by invitations sent to a random selection of Australian addresses.

The Australian Election Study is chosen as it is the most comprehensive public survey of Australians' political and cultural values. Besides questions about party politics, the study contains 103 questions which we consider to be about political issues, allowing for a broad-based understanding of Australian political beliefs and discourse. The study also contains 32 questions we consider to be about individuals' demography, meaning a model can be built that considers more than just, say, age and gender, a common limitation of opinion polling data. This means the Australian Election Study can be used to relate demographic and political identity much more comprehensively than can other studies.

In Chapter 2, we also demonstrate that the sample provides a relatively good representation of the demographics of Australian electors, making the sample more likely to be generalisable to the population.

Chapter 3: Dimension Reduction Techniques

Chapter 3 introduces mathematical techniques for reducing the dimension of observed variables. This invariably involves replacing the high-dimensional space with a lower-dimensional representation. These techniques are necessary in the context of analysing the Australian Election Study, since there are too many political and demographic variables present for a manageable explanatory model to be produced; we want to be able to explain the relationships present, which becomes difficult with 32 demographic predictors and 103 political response variables.

In Chapter 3, we explore the literature behind six techniques for dimension reduction:

1. Principal Component Analysis;

2. Multiple Correspondence Analysis;

3. Factorial Analysis of Mixed Data;

4. Multiple Factor Analysis;

5. Multidimensional Scaling; and

6. Factor Analysis.

Each methodology is introduced alongside notes for that technique's application. With each technique being more popular in some disciplines than others, and being more useful with some types of data than others, it is important to understand the practical context, as well as the theory, behind these dimension reduction techniques.

Chapter 4: Dimension Reduction Application

Chapter 4 chooses the most appropriate dimension reduction technique for the set of observed political variables, and for the set of observed demographic variables. Different techniques are chosen for each set of variables due to the differing data types. When applied, we produce a dimension-reduced demographic spectrum, and a dimension-reduced political spectrum.

When interpreting the dimension-reduced demographic variables, we find that the demographic spectrum is dominated by three variables, which we label 'socio-economic status and education', 'stage of life' and 'cultural background'. Meanwhile, when interpreting the dimension-reduced political variables, we find that the political spectrum is also dominated by three variables, which we label 'social inclusivity', 'attitudes to authority' and 'spending priorities'.

Finally, we explore how political spectra have been described in other literature throughout the last 250 years. We note that our new political spectrum falls neatly

in this lineage; it both reflects previous understandings of political ideology, and expands the political spectrum, in a statistically-justified way, to reflect modern developments in political discourse.

Chapter 5: Multivariate Regression

With dimension-reduced spectra of political ideology developed in Chapter 4, Chapter 5 introduces the first model for relating these spectra—the multivariate regression model. We do so by first introducing a univariate regression model more common in statistical literature, and expanding this to allow for a multivariate response, since our spectrum of political ideology is multi-dimensional. We explore the assumptions behind the multivariate model, as well as how to test them, noting that the most important assumption is that the residuals of the model are independent observations of a multivariate normal distribution. We also prove a key limitation of the model, that it can only produce hyperelliptical prediction regions; in other words the model does not allow for non-linear relationships in its error terms.

We then apply the multivariate regression model to the 2016 Australian Election Study data set. Building first a linear model in the demographics, we note that the assumption that the residuals of the model are independent observations of a multivariate normal distribution cannot be satisfied, since there are non-linear relationships between the model's residuals in different dimensions. We then try a number of more complicated terms in the multivariate regression model, including polynomial terms, interaction terms, logarithms and exponentials, with the goal of using more complicated functions of the demographics to explain away the non-linear relationships between the political variables. We find that the demographic variables are unable to do this; there is a non-linear relationship between different aspects of political thought that cannot be explained by demography alone. We conclude that different models must be considered, which are better able to explain these relationships.

Chapter 6: Alternatives to the Multivariate Regression Model

With Chapter 5 establishing that the multivariate regression model has room for improvement in terms of its applicability to the 2016 AES data set, Chapter 6 introduces three alternative candidates to model the relationships between demographic and political identity. All of the three models are extensions to regression models, either univariate or multivariate, built so as to be able to encapsulate non-linear relationships between errors in different dimensions.

The first, the recursive multivariate model, is built around the construction of a series of univariate regression models, with each successive model containing non-linear functions of the previous model's error term, to encapsulate non-linear relationships between errors.

The second, the Gaussian mixture model, is expressed identically to the multivariate regression model, save that the error term, which is a single multivariate normal random variable in the multivariate regression model, is replaced by a

mixture of a number of multivariate normal random variables, allowing for clusters of errors to be represented.

Similarly, the third model, the kernel density model, replaces the multivariate regression error term with a kernel density estimate on the residuals. This allows for even more complex relationships in the errors.

In Chapter 6, each of these models are introduced, and the method for fitting the models is explained or derived. Each model is fit to the 2016 AES data set.

Chapter 7: Assumption-Driven Model Selection

Chapters 5 and 6 introduced four candidate models to describe the relationships between demographic and political identity as expressed by the 2016 AES data set— the multivariate regression model, the recursive multivariate model, the Gaussian mixture model, and the kernel density model. With no predominating consideration for model selection, Chapter 7 then explores a number of bases for selecting a model. Methods introduced and applied include the accuracy and size of prediction regions, the models' parsimony, and their relevance to the political science domain.

Given the importance of accurately representing the direction, as well as the magnitude, of a model's errors, we also seek a model selection criterion able to compare how well data adheres to the distributional assumptions relating to a candidate model's error. We explore the literature on goodness-of-fit comparisons for non-nested models, and look for techniques able to compare distributional goodness-of-fit statistics across multiple models simultaneously. Discovering such an approach, called 'model mimicry', we place it in the context of its statistical lineage, and rigorously explain the technique's application and interpretation. We explain how numerical discriminants can be used to select a model. We hope that this thesis can help other statistical practitioners choose between multiple non-nested models on the basis of generalised goodness-of-fit criteria.

Finally, a model is selected on the balance of the model selection methods discussed in the chapter.

Chapter 8: Results in Context

With a model selected, model parameters and results can be discussed in light of the political science domain in which they reside. In Chapter 8, we explain the results in context. We firstly explain each linear association between a dimension-reduced demographic axis and a dimension-reduced political axis. Since we describe the three main demographic and political axes, there are nine such associations. Each is compared with relevant social science literature.

We then explain the relationships that emerge between the dimension-reduced political axes, once demographics have been accounted for. The relationships that emerge here have been described to some extent in previous social science literature, but this thesis, by splitting the political spectrum into the three axes described in Chapter 4, is able to clarify this work further. We note that our political spectrum findings echo recent populist political developments around the world. For this reason, we propose that future work reproducing this research in an overseas

context, and across different time periods could help determine when and where this development has emerged.

1.2 Publication status and plans

Three papers are planned arising out of this thesis: one theoretical and two presenting applications.

The first paper, currently in preparation, outlines the model comparison technique described in Chapter 7, which builds on the 'model mimicry' technique introduced by Wagenmakers *et al.* [172]. The paper places the technique we use in this thesis in the context of surrounding literature, and uses both synthetic and observed datasets to verify and demonstrate the technique's validity. This paper will provide an outline for other statisticians to select models on the basis of novel goodness-of-fit measures, including fidelity to model assumptions.

The second paper outlines the political spectrum developed in Chapter 4, alongside similarly-produced political spectra overseas. We will use the same techniques as in this thesis to produce political spectra using international data sets, so as to compare differences in political discourse around the globe. These spectra will reflect differences both in the kinds of political questions asked in different countries, and the questions in particular that most unite and divide people in those countries. If demographic data is available, this paper will also seek to model relationships between political and demographic identities.

The third paper will study how the political spectrum has changed in Australia over time, especially in relation to the axes emergent in the 2016 Australian Election Study. This thesis has noted that the three main axes of political thought in Australian, as at 2016, were 'social inclusivity', 'attitudes to authority' and 'spending priorities'. This is an extension of previous research, which has noted the social and economic axes, but not 'attitudes to authority' (see Chapter 4). Since this 'attitudes to authority' axis is incredibly relevant in light of recent political developments (see Chapter 8), we seek to determine at what point this axis emerged. Using previous editions of the Australian Election Study, we can perform the same analysis as done in this thesis to determine the relative importance of the 'attitudes to authority' axis over time. In effect, we can measure how populist thought has changed in Australia over time.

Chapter 2

The Australian Election Study

Shortly after each Australian federal election, the Australian National University ('ANU') undertakes the Australian Election Study ('AES'), a survey of potential voters' backgrounds, values and engagement in the election campaign [121]. Eleven surveys were conducted between 1987 and 2016, corresponding to the eleven federal elections in this period. Due to the scheduling of elections, each survey is between two and three years after the previous survey.

The survey seeks to sample from the population of people eligible to vote in Australia. Potential subjects are selected by the Australian Electoral Commission, who take a random sample of individuals from the federal electoral roll, stratified by state in proportion to number of enrolled voters, in a bid to ensure representativeness [121]. These subjects are sent a letter containing the survey, and, later, a follow-up postcard. More follow-ups are undertaken when necessary, though naturally, completion of the survey is not compulsory. The fact that the survey is not compulsory may lead to self-selection bias, especially in instances where the survey has a low response rate.

Sample sizes range from 1,769 (2004) to 3,955 (2013) [23, 24]. A downward trend in response rates has been observed, varying from a high of 62.8% in both 1987 and 1993, to a low of 22.5% in 2016 [123, 101, 121]. This has been accompanied by an increase in the number of subjects approached, with the potential sample reaching a peak of 12,497 in 2016 [121].

Over time, the number of questions in the AES has varied, from 193 in the inaugural 1987 study, to a maximum of 461 in 2004, decreasing to 302 in 2016 [123, 23, 121]. While attempting to keep the survey short enough to encourage high participation, each survey builds on those prior by adding new questions relevant to Australia's then-current political discourse. For example, despite an increase in the number of immigration-related questions in the 1996 and 1998 surveys [101, 102, 21], a question was not asked specifically relating to asylum seekers prior to the 2001 edition of the study [22], during an election campaign dominated by the Tampa Affair [70, 120, 125]. By contrast, a number of questions relating to the AIDS epidemic were removed between the 1987 and the 1990 surveys [123, 122]. Further, many questions are repeated word-for-word from one study to the next, making results comparable; the AES seeks to maintain a balance between keeping questions

consistent and keeping questions relevant.

All AES data used here has been made available through the Australian Data Archive (see, e.g., [121]). The original investigators for each edition of the AES can be found in the Bibliography. Those carrying out the original analysis and collection of the data sets bear no responsibility for the analysis presented in this thesis and the interpretation of these analyses.

The 2016 Study

The 2016 Australian Election Study was conducted between the 30th of June and the 6th of November, 2016, to coincide with the 2016 Australian federal election, held on July 2 [121]. As in previous incarnations of the AES, subjects were selected from a stratified random sample of the Australian electoral roll. A supplementary sample was taken using a random selection of Australian addresses from the Geocoded National Address File (G-NAF); this was done with the goal of increasing participation among younger Australians who may not yet be on the electoral roll, despite intending to vote. Of the 12,497 people contacted, 2,818 valid responses were obtained. While this 22.5% response rate is lower than in previous incarnations of the survey, it is much higher than for comparable polls internationally; for example, in 2016, response rates for the Gallup Poll Social Series, a leading social attitudes survey in the United States, were just 6% [118]. This may be a result of the AES's sampling methodology, which included four follow-up letters being sent to potential subjects after the initial contact [121].

2.1   Variables

The AES 2016 data set comprises 322 variables, 302 of which correspond to questions asked of the respondents, and 20 of which contain metadata about the interviews. ANU divides the variables into thirteen groups [121], the sizes and names of which are displayed in Table 2.1. The group 'Restricted variables' contains the subjects' postcodes, and is not used here for privacy reasons; only subjects' state of residence is included in this research. The weighting variables are also unnecessary here, since in building a model drawing associations based upon demography, it matters not how many people within a particular demographic group are used to train the model, so long as each demographic group is sufficiently well represented.

Grouping	Number of variables
The election campaign	53
Election issues	51
Personal background	47
Party preference and voting	37
Politicians and government	37
Social policy	36
General political views	26
Education and work	15
Administrative	13
Weight variables	3
Restricted variables	3
Mapping variables	1

Table 2.1: The 322 variables in the 2016 Australian Election Study, as grouped by the Australian National University, who collected the data. This grouping is not directly used for the purpose of this research, but is representative of the objectives of ANU in collecting the data.

Of the remaining variables, these have been separately divided for this research into three groups:

1. demographic variables;

2. political variables; and

3. other variables.

It should be noted that questions regarding political parties or politicians are not included in the 'political variables' grouping; the purpose of this research is to uncover the relationship between demography and political issues, not the relationship between demography and political personalities. As such, much of the 'the election campaign', 'party preference and voting' and 'politicans and government' groups as defined by ANU are treated as 'other variables' for the purpose of this research.

Political variables

Overall, there are 103 political variables used in this research. A full list of these variables can be found in the data dictionary attached in Appendix A, and examples of the political variables can be found in Table 2.2. All of the variables selected as 'political' are either ordinal or nominal categorical; 98 are ordinal and 5 are nominal categorical. The ordinal variables are mostly on the 5-point Likert scale, with responses ranging from Strongly Disagree to Strongly Agree, but other

responses also have a natural ordering; for example, the question "Which one of these statements comes closest to how you feel about abortion in Australia" has three answers–"Women should be able to obtain an abortion readily when they want one", "Abortion should be allowed only in special circumstances" and "Abortion should not be allowed under any circumstances"'–these answers are in increasing order of opposition to abortion.

While not all subjects responded to all questions, the issue of missingness as it pertains to the political variables is a relatively minor one; all political questions were answered by a minimum of 80.4% of respondents. The manner with which missingness has been dealt is addressed in Section 2.2.

Demographic variables

Similarly, a grouping of 'demographic variables' was produced. This grouping contains all variables which refer to the backgrounds and experiences of respondents. There are 32 such variables in the 2016 AES data set. The demographic variables are wide-ranging in scope. Examples of the demographic variables, their data types and the number of valid responses obtained can be found in Table 2.3. A full list of the demographic variables can be found in the data dictionary attached in Appendix A.

Of the demographic variables, 25 are categorical nominal variables, four are ordinal, and three are numeric. Of the nominal variables, there is much variation in the number of levels, ranging from 2 to 98. Where a large number of levels exist, these may be recoded to make analysis more simple; this is discussed in greater detail in Chapter 4.

Similarly to the political variables, there is some missing data in the demographic variables. With the exception of questions regarding respondents' partners, all questions were answered by at least 76.61% of respondents. Missingness in the questions about respondents' partners, which all had between 50% and 60% valid responses, is as a result of the number of single participants in the survey; for each of these questions, less than 10% of respondents did not provide a valid answer.

Political question	% valid responses	Type of variable	Type of responses
How serious a threat do you think global warming will pose to you or your way of life in your lifetime?	95.24%	Ordinal	4 levels (Very serious; fairly serious; not very serious; not at all serious)
Do you strongly agree, agree, disagree or strongly disagree with the following statement? The smoking of marijuana should NOT be a criminal offence	94.39%	Ordinal	5 levels (Strongly agree; agree; neither agree nor disagree; disagree; strongly disagree)
Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?	94.14%	Categorical nominal	2 levels (most people can be trusted; you can't be too careful)

Table 2.2: Examples of political variables in the 2016 Australian Election Study. Example variables are chosen to reflect the mixture of data types in the AES.

Demographic question	% valid responses	Type of variable	Type of responses
In what country was your father born?	92.83%	Categorical nominal	98 levels (97 countries, 1 'born at sea')
As an official part of your job, do (or did) you directly supervise the work of other employees or tell other employees what work to do?	84.53%	Categorical nominal	2 levels (yes; no)
How old were you when you left secondary school?	86.66%	Numerical	Age (years)
Would you say you now live in…?	93.86%	Ordinal	5 levels (a rural area; a small country town; a larger country town; a large town; or a major city)

Table 2.3: Examples of demographic variables in the 2016 Australian Election Study. Example variables are chosen to reflect the mixture of data types in the AES.

2.2   Missing data

As has been discussed in Section 2.1, there is a small amount of missing data in the results of the AES 2016 survey. Missing values will be filled using the imputation methods for mixed data in the `missMDA` [104] package in R [149], a companion package to the dimension reduction package `FactoMineR` [109]. The imputation procedure involves using principal components-related techniques (see Chapter 3) to find relationships between the variables, allowing missing values to be estimated using these relationships. This assumes that the missingness in the data is uninformative, an assumption which may not be justified; for example, people who do not wish to share information about their gross annual income may be of a particular economic position, that is not captured by the other socio-economic variables. Without any information about the nature of question-by-question selection biases that may be present, and with all of the studied variables containing less than 24% missing data, and the relevant variables containing an average of 6.8% missing data, dealing with missingness will not form a significant focus of this project. Additional information on the techniques used can be found in the work of Audiger, Husson and Josse (2016-7) [13, 14, 15].

The `missMDA` package imputes numerical values for quantitative variables, and for categorical variables, provides the option of either probabilistically imputing categories, or taking the most probable factor [104]. For example, if a categorical variable took two values, *yes* or *no*, and the imputation algorithm gives a 60% chance of a missing value being *yes* and a 40% chance of *no*, the `missMDA` imputation gives the option of taking a value of either $0.6 \times yes + 0.4 \times no$, or simply imputing a value of *yes*. Since the probabilistic imputation is compatible with the dimension reduction techniques outlined in Chapter 3, and more accurately reflects the uncertainty about the imputed value, it will be preferred here.

2.3   Ethics

As with any use of survey data, ethical issues might arise where subjects might be identifiable. Given the specific and personal nature of the large number of survey questions, avoiding the identification of subjects is particularly important. This being said, there is negligible risk of re-identification of subjects for this data. Subjects have been de-identified, and no geographical identifiers are used on a finer scale than State [121]; while the postcodes of subjects were recorded, access to this information is restricted by ANU and was not sought for this research.

The use of data relating to human subjects complies with the National Health and Medical Research Council (NHMRC) National Statement on Ethical Conduct in Human Research 2007 [134], as cited in the Australian Code for the Responsible Conduct of Research 2018 ('the Code') [135]. This thesis' use of the 2016 AES data has been deemed 'Negligible Risk' under the Code.

2.4   Representativeness

For the 2016 AES data set to be used in political modelling, there must be sufficient representation of all relevant demographic groups. This does not necessarily mean that the sample need have an identical demographic make-up to the population, since inference is made for particular demographic groups as opposed to the whole population. However, it is necessary that there are enough people of all relevant demographics in the sample to be able to infer relationships in those demographics. The following paragraphs compare the 2016 AES sample with the population of those eligible to vote in Australia.

Gender   According to the Australian Electoral Commission, in the 2016 federal election, 51.4% of the electorate was female, 48.6% of the electorate was male, and 0.003% was of indeterminate or unknown gender [18], while in the 2016 Australian Election Study, 48.3% of subjects responded that they were female, 46.4% identified as male, while 5.3% of subjects did not wish to say. Of those who identified their gender as either male or female, the 2016 AES skews 51.0% female to 49.0% male. The sample is thus a close representation of the population male/female gender divide.

It should be noted that options other than male and female were not given to the gender question in the 2016 AES, and electoral roll entries were only required by law to recognise individuals identifying as neither male nor female as recently 2013 [12, 138]. This means that the 0.003% of electoral roll entries corresponding to indeterminate or unknown gender is very likely an underestimate, and this demographic is not explicitly captured in the AES at all. As such, whether the AES sample has accurate non-binary representation of the population of eligible electors is difficult to assess. Since non-binary options were not provided in the 2016 AES, gender-based inference here will only be drawn for male and female subjects.

State   The percentage of eligible electors residing in each state in the 2016 Australian federal election can be found in Table 2.4, alongside the number of respondents in the 2016 Australian Election Study [18, 121]. It can be observed that there is slight under-representation of New South Wales and Western Australia in the AES, and substantial under-representation of the Northern Territory, who while being 0.85% of the electorate, are only 0.46% of the AES sample. This makes inference on the basis of residence in the Northern Territory less well-founded; 0.46% of the AES sample corresponds to just 13 subjects.

Income   The median gross household income for subjects in the 2016 AES was calculated using responses to the question "What is the gross annual income, before tax or other deductions, for you and your family or others living with you from all sources? Please include any pensions and allowances, and income from interest or dividends." Since responses were given only within particular income

State/Territory	Electorate (%)	AES Sample (%)
New South Wales	32.5	30.5
Victoria	25.3	25.9
Queensland	19.6	19.8
Western Australia	10.1	9.3
South Australia	7.6	9.3
Tasmania	2.4	2.6
Australian Capital Territory	1.8	2.2
Northern Territory	0.9	0.5

Table 2.4: Representation of Australia's six states and two self-governing territories in the electoral roll for the 2016 Federal Election [18], and in the 2016 Australian Election Study [121]. The Northern Territory, New South Wales and Western Australia are each under-represented in the AES, though only under-representation of the Northern Territory is so substantial as to be of concern.

increments (for example, $40,001 to $45,000 per year), it was assumed that responses were uniformly distributed within these increments. Thus, since the median was assessed to be the 101st smallest income in the $60,001 to $70,000/year increment, of 153 subjects in that increment, the median was calculated to be $\frac{101}{153} \times 70000 + \left(1 - \frac{101}{153}\right) \times 60001 = \$66,602$.

Among reported gross household incomes in the 2016 Australian census, the median income was calculated to be $74,493 [16]. This calculation was done in the same manner as for the AES sample, since household income responses in the census are given only within particular income brackets. This value appears to suggest that the AES 2016 sample has a lower median income ($66,602) than the population ($74,493). However, there is a key difference in methodology between the collection of household incomes in the AES and the census; while the AES requires an annual income figure to be reported for an entire household, income for the census is collected for each subject separately, and aggregated for each household. As a result, a household in which not all members report their income is treated as missing data for the census, while this issue does not appear in the AES sample. There were 881,074 such households in the 2016 census, in which 649,875 involve some but not all members of a family reporting their income. Households in which one or more members does not report their income, but other members do report income, are likely to be single-income families. Since single-income families have, on average, smaller income, the missingness in the census data may result in the calculated median of $74,493 being an overestimate. In light of this, the difference in the medians between the 2016 AES and the 2016 census are not so dramatic as to indicate that the 2016 AES sample is unrepresentative of Australia's electorate at large, in terms of income.

Place of origin   Since this research seeks to represent the diversity of background and political opinion that exists in the Australian electorate, it is important that there is adequate representation of migrant communities in the Australian Election Study. Since the AES seeks to represent people on the electoral roll, it would be useful to compare the countries of birth of those in the AES, with those on the electoral roll. Unfortunately, country of birth for those on the electoral roll is not made public by the Australian Electoral Commission (see, e.g., [45] at Section 90B), so census figures for country of birth are used [16]. It should be noted that the census figures have a higher proportion of people not born in Australia than the electoral roll, since overseas-born Australians are more likely to be captured by the census and yet be ineligible to vote, than those born in Australia, who in most circumstances [18] are on the electoral roll.

Table 2.5 contains the ten most common countries of birth for those captured by the 2016 census [16], and their representation in the 2016 AES [121]. There is close correspondence between the population proportion of Australians from particular countries, and their representation in the AES. Discrepancies can be largely accounted for by the fact that the AES seeks to represent those on the electoral roll, as opposed to all people residing in Australia. The largest difference between the 2016 census and the AES sample is the AES's over-representation of people from the United Kingdom. This group is more likely than others to be on the electoral roll due to Section 93(1)(b)(ii) of the Commonwealth Electoral Act 1918 [45], which allows British subjects who were on the electoral roll prior to 1984 to remain on the roll. The remaining over-representation is unlikely to be an issue since this is not at the expense of the representation of any particular other migrant group.

Aboriginal and Torres Strait Islander Status   The Australian Electoral Commission does not ask people whether they identify as Aboriginal and Torres Strait Islander when they enrol to vote, so the population proportion of Aboriginal and Torres Strait Islander people on the electoral roll must be estimated. According to the Australian Electoral Commission estimates, 58% of eligible Aboriginal and Torres Strait Islander people were enrolled to vote at the 2016 election. This is much lower than the enrolment rate for non-indigenous Australians. As such, the proportion of Aboriginal and Torres Strait Islander people on the electoral roll is estimated to be 58% of voting-age Aboriginal and Torres Strait Islander people. Using the number of Aboriginal and Torres Strait Islander-identifying people from the 2016 census, this number is calculated to be 387,270, or 2.47% of the electoral roll. This compares with 1.05% of the Australian Election Study sample, indicating that Aboriginal and Torres Strait Islander people are under-represented in the AES. This may be somewhat mitigated by the fact that 5.50% of the sample did not respond to the question regarding Aboriginal and Torres Strait Islander status, but this still indicates that inference regarding the views of Aboriginal and Torres Strait Islander people from the AES may be less reliable than inference for non-indigenous people.

Country	Australians (%)	AES Sample (%)
Australia	66.7	70.6
United Kingdom	4.6	8.4
New Zealand	2.2	1.6
China (excludes SARs and Taiwan)	2.2	0.8
India	2.0	1.1
Philippines	1.0	0.5
Vietnam	0.9	0.5
Italy	0.7	1.1
South Africa	0.7	0.5
Malaysia	0.6	0.6

Table 2.5: Representation of the most common countries of birth for Australians in the Australian Election Study. While Australian-born people are over-represented, relative to the Australian population, in the AES, this is not especially problematic, since the AES seeks to reflect the population of potential electors, not the entire Australian adult population. There is substantial over-representation of people born in the United Kingdom, who by action of the Electoral Act [45] are much more likely than other overseas-born people to be on the Electoral roll. Under-representation of Chinese- and Indian-born Australians may be as a result of the higher number of recent migrants from these countries [98], who may not yet be on the electoral roll.

Age group	Electoral Roll (%)	AES (%)
18-19	2.7	1.5
20-24	7.9	2.8
25-29	8.1	3.7
30-34	8.4	4.6
35-39	8.2	5.6
40-44	8.8	7.0
45-49	8.8	7.5
50-54	8.8	8.8
55-59	8.6	11.8
60-64	7.8	13.0
65-69	7.2	12.7
70+	14.8	21.0

Table 2.6: Representation of age groups in the AES sample, as compared with the proportions of the electoral roll in each group. Older Australians are over-represented in the AES, despite efforts to reduce bias by sampling outside the electoral roll [121]. Since this research aims to compare these demographics, this over-representation will not pose too great a problem, since there is still a sufficiently large sample from each key demographic group.

Age   The AES should reflect the age distribution of the electoral roll. The Australian Electoral Commission provides figures for the ages of enrolled electors at the 2016 federal election, in five-year intervals, as can be seen in Table 2.6. This table also has the corresponding proportions in the AES sample. This table demonstrates that younger demographics are under-represented in the Australian Election Study, for example, those under 20 represent 2.68% of the electoral roll, but only 1.51% of the AES. In fact, all groups under the age of 50 were under-represented. This is despite methodological changes in the 2016 AES to increase the representation of young people; while in previous years, the electoral roll was the sole source of subjects, in 2016 potential subjects were chosen by a combination of random sampling from the electoral roll and from the Geocoded National Address File (G-NAF), to increase the proportion of the sample who were of age to vote but had not yet enrolled.

If inference was to be made without consideration of the under-representation of younger subjects in the sample, and age is indeed a factor in political opinion, results would be skewed in favour of the views of older members of the population. Since age will be included as a predictor in models built herein, and no demographic is so under-represented as to lead to too-small sample sizes in a particular age group, inference can still be made for people of particular age groups.

Summary As a general rule, a data set should be representative of the subjects it seeks to describe. Where a demographic group is over-represented, inference regarding the population will be biased in favour of that group. In this research, however, inference is not sought regarding the views of the population as a whole; no effort is made to establish what the 'average' Australian's view is on an particular issue. Instead, here inference is done with an aim to establish what contrasts exist in the relationships between individuals' experiences, and their political perspectives. This inference is not directly influenced by the under-representation or over-representation of certain groups; while sample sizes may be different between two groups we seek to compare, this comparison will be unbiased so long as each of the samples are unbiased.

This being said, representativeness is relevant for two key reasons: to ensure that there is a sufficient sample of each relevant demographic group, since a group may be so under-represented that it is unwise to draw conclusions from the sample available; and to prevent contrasts between groups being biased by representation of subgroups. An example of the latter consideration is as follows. Suppose contrast is drawn between the political views of people with higher incomes, and those with lower incomes. If inference based upon a sample which under-represents migrants, then conclusions will be biased in the direction of contrasts between Australian-born people with higher incomes and Australian-born people with lower incomes. As such, especially where intersectional relationships are sought, representativeness of the sample is a relevant consideration.

In the 2016 AES data set, the following conclusions can be made regarding representation of Australian electors' demographic groups:

- **Gender** was accurately reflected, in a binary sense; non-binary gender options were not provided in the AES and are not accurately recorded in the electoral roll;

- **State** of residence was accurately reflected, with the exception of a slight under-representation of subjects from the Northern Territory, which makes conclusions for such subjects tenuous;

- **Income** looks to be accurately reflected, accounting for methodological differences in the manner of calculating income between the census and the AES;

- **Place of origin** is largely accurately represented, though with some under-representation of recent migrant communities;

- **Aboriginal and Torres Strait Islander Status** is poorly represented, with Aboriginal and Torres Strait Islander people comprising 1.05% of the AES but 2.47% of the electoral roll; and

- **Age** is not well-represented, with an over-sample of older subjects.

Despite some groups being under-represented, the conclusions drawn from analysis of the 2016 AES sample should not be tainted by strong bias in favour of any particular demographic groups. In light of the fact that the 2016 AES data set arose

from a non-compulsory survey, that the data set is somewhat representative is to be commended. While it may be that future such surveys should be undertaken with specific demographic targets in mind, especially to combat the under-representation of Aboriginal and Torres Strait Islander peoples and migrants, the 2016 AES data set should nonetheless provide accurate insights into the relationships between experiences and beliefs.

Chapter 3

Dimension Reduction Techniques

In Chapter 2, the data used in this thesis, the 2016 Australian Election Study ('AES'), was introduced. This data set is used with the goal of building a model connecting demographic and political experiences of Australians; we seek to use the demographic variables in the 2016 AES as predictors, and the political variables as responses. As seen in Chapter 2, there are 103 political variables in the data set, and 32 demographic variables. Naturally, building a model with 103 response variables is unwieldy. We thus seek to replace the AES's high-dimensional representation of political identity with a lower-dimensional representation. The act of replacing a high-dimensional variable with a lower-dimensional representation is called dimension reduction.

This chapter presents six methods of dimension reduction, as well as notes for their application. These methods are

- Principal Component Analysis (Section 3.1);

- Multiple Correspondence Analysis (Section 3.2);

- Factorial Analysis of Mixed Data (Section 3.3);

- Multiple Factor Analysis (Section 3.4);

- Multidimensional Scaling (Section 3.5); and

- Factor Analysis (Section 3.6).

The first four of these techniques—namely, Principal Component Analysis, Multiple Correspondence Analysis, Factorial Analysis of Mixed Data, and Multiple Factor Analysis—are all variants of Principal Component Analysis, which focuses on producing dimension-reduced variables which capture as much of the variation in the observed data as possible. Meanwhile, Multidimensional Scaling is focused on best ensuring subjects in the dimension-reduced variable space are as similar, or as different, as they were in the observed data set. Finally, Factor Analysis provides a model-centric approach to dimension reduction, where it is assumed that the observed variables are linear functions of some given smaller number of underlying variables, with some noise added.

Once these dimension-reduction techniques are understood, we can go on to select the most appropriate method for the types of data in the 2016 AES. This technique selection, and application, is done in the following chapter, Chapter 4.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is an exploratory data analysis technique which seeks to represent random variables $\boldsymbol{X}^{\mathrm{T}} = [X_1, X_2, \ldots, X_p]$ through linear combinations thereof. Since these linear combinations are chosen to be linearly independent, PCA is a process involving and determining a change of basis. The transformed set of random variables $\boldsymbol{Y}^{\mathrm{T}} = [Y_1, Y_2, \ldots, Y_p]$ can be written as $Y_i = \boldsymbol{a_i}^{\mathrm{T}} \boldsymbol{X}$, where the components $\boldsymbol{a_i}, i = 1, 2, \ldots, p$ are vectors of the form $\boldsymbol{a_i}^{\mathrm{T}} = [a_{i1}, a_{i2}, \ldots, a_{ip}]$. Alternatively, we can write

$$\boldsymbol{Y} = A^{\mathrm{T}} \boldsymbol{X}, \tag{3.1.1}$$

where $A = [\boldsymbol{a_1}^{\mathrm{T}}, \boldsymbol{a_2}^{\mathrm{T}}, \ldots \boldsymbol{a_p}^{\mathrm{T}}]$.

The new basis is chosen such that the first variable Y_1 captures as much of the variance of the random variables $\boldsymbol{X}$ as possible, that the second variable Y_2 captures as much of the remaining variance as possible while being uncorrelated with Y_1, and so on until the final variable Y_p captures all remaining variance. Thus at each stage of PCA, we seek to maximise $\mathrm{Var}(Y_i)$, such that all Y_i are uncorrelated with each other: $\mathrm{Cov}(Y_i, Y_j) = 0, \ \forall \ j < i$. The fact that the variables Y_i are uncorrelated is useful in a setting where they are to be used as predictors in regression, as this prevents any problems arising from collinearity. The argument in Section 3.1.1 is adapted from that outlined by Chatfield and Collins (1980) [41].

3.1.1 Obtaining the components

Firstly, the components $\boldsymbol{a_i}$ should be subject to some normalisation constraint, since otherwise the problem would be ill-constrained; unless all of the observed random variables $\boldsymbol{X}$ have zero variance, the variance of Y_i could be arbitrarily increased by increasing components $\boldsymbol{a_i}$, leading to $\mathrm{Var}(Y_i) = \infty$. As such, we normalise $\boldsymbol{a_i}$ by setting $\boldsymbol{a_i}^{\mathrm{T}} \boldsymbol{a_i} = 1$ for $i = 1, 2, \ldots p$.

The first principal component can then be obtained by solving the optimisation problem

$$\underset{\boldsymbol{a_1} \in \mathbb{R}^p}{\mathrm{argmax}} \, \mathrm{Var}(Y_1) : \boldsymbol{a_1}^{\mathrm{T}} \boldsymbol{a_1} = 1. \tag{3.1.2}$$

We define

- $\mathrm{Var}(\boldsymbol{X})$ to be Σ; this means

- $\mathrm{Cov}(X_i, X_j) = [\Sigma]_{ij} = [\Sigma]_{ji}$

- and $\mathrm{Var}(X_i) = [\Sigma]_{ii}$,

meaning

$$\mathrm{Var}(\boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}) = \boldsymbol{A}^{\mathrm{T}}\Sigma\boldsymbol{A}, \text{ and}$$
$$\mathrm{Var}(Y_i) = \mathrm{Var}(\boldsymbol{a_i}^{\mathrm{T}}\boldsymbol{X}) = \boldsymbol{a_i}^{\mathrm{T}}\Sigma\boldsymbol{a_i}. \tag{3.1.3}$$

Thus Equation (3.1.2) can be solved by employing a Lagrange multiplier to find the critical points with respect to $\boldsymbol{a_1}$ of

$$L(\boldsymbol{a_1}) = \boldsymbol{a_1}^{\mathrm{T}}\Sigma\boldsymbol{a_1} - \lambda(\boldsymbol{a_1}^{\mathrm{T}}\boldsymbol{a_1} - 1):$$
$$\frac{dL}{d\boldsymbol{a_1}} = 2\Sigma\boldsymbol{a_1} - 2\lambda\boldsymbol{a_1}.$$
$$\Rightarrow \frac{dL}{d\boldsymbol{a_1}} = 0 \text{ iff } (\Sigma - \lambda\boldsymbol{I})\boldsymbol{a_1} = 0. \tag{3.1.4}$$

Equation (3.1.4) means that $\boldsymbol{a_1}$ is a normalised eigenvector of Σ. The eigenvector $\boldsymbol{a_1}$ must be selected to maximise $\mathrm{Var}(Y_1)$. Using Equation (3.1.3),

$$\mathrm{Var}(Y_1) = \boldsymbol{a_1}^{\mathrm{T}}\Sigma\boldsymbol{a_1}$$
$$= \boldsymbol{a_1}^{\mathrm{T}}\lambda\boldsymbol{I}\boldsymbol{a_1} \text{ (by (3.1.4)}, \Sigma\boldsymbol{a_1} = \lambda\boldsymbol{I}\boldsymbol{a_1})$$
$$\mathrm{Var}(Y_1) = \lambda \text{ (since } \boldsymbol{a_1}^{\mathrm{T}}\boldsymbol{a_1} = 1). \tag{3.1.5}$$

As such, the variance of Y_1 will be maximised when the component $\boldsymbol{a_1}$ is the eigenvector associated with the greatest eigenvalue λ_1.

The second component is obtained by a similar procedure, with the additional requirement that Y_2 and Y_1 are uncorrelated. In other words, $\mathrm{Cov}(Y_2, Y_1) = 0$. Since

$$\mathrm{Cov}(Y_2, Y_1) = \mathrm{Cov}(\boldsymbol{a_2}^{\mathrm{T}}\boldsymbol{X}, \boldsymbol{a_1}^{\mathrm{T}}\boldsymbol{X})$$
$$= \boldsymbol{a_2}^{\mathrm{T}}\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{a_1}$$
$$= \boldsymbol{a_2}^{\mathrm{T}}\Sigma\boldsymbol{a_1}$$
$$= \boldsymbol{a_2}^{\mathrm{T}}\lambda_1\boldsymbol{a_1},$$

$\boldsymbol{a_2}^{\mathrm{T}}\boldsymbol{a_1} = 0$. The requirement that the Y_is are uncorrelated is therefore equivalent to a requirement that the $\boldsymbol{a_i}$s are orthogonal.

Because our covariance matrix Σ is a real symmetric matrix, all its eigenvectors are orthogonal where they do not correspond to the same eigenvalue. As a result of this, if Σ has p distinct eigenvalues, there are p orthonormal solutions to Equation (3.1.4) – the eigenvectors of Σ. Thus if these eigenvectors are normalised, they are the desired vectors $\boldsymbol{a_i}, i = 1, 2, \ldots, p$, since they satisfy

$$\underset{\boldsymbol{a_i} \in \mathbb{R}^p}{\mathrm{argmax}}\ \mathrm{Var}(Y_i), \text{ s.t. } \begin{cases} \boldsymbol{a_i}^{\mathrm{T}}\boldsymbol{a_i} = 1, \text{ for } i = 1, 2, \ldots, p, \\ \boldsymbol{a_i}^{\mathrm{T}}\boldsymbol{a_j} = 0, \text{ for } i \neq j. \end{cases}$$

The case of repeated eigenvalues will be discussed in Section 3.1.2.

The components thus form an orthonormal basis of $\mathbb{R}^p$, meaning that the transformation $\boldsymbol{Y} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}$ is invertible. This means that given some observation of the variables $\boldsymbol{X}$, we can transform to obtain the principal-components-transformed set $\boldsymbol{Y}$, and given an observation of the transformed set $\boldsymbol{Y}$, we can calculate their real-world realisations $\boldsymbol{X}$.

3.1.2   Ranking the components

The components are chosen in such an order that $\text{Var}(Y_1)$ is maximised in preference to $\text{Var}(Y_2)$, and so on. By Equation (3.1.5), the variance of Y_i is the eigenvalue corresponding to the eigenvector $\boldsymbol{a_i}$. Thus the components $[\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_p}]$ are the eigenvectors of Σ corresponding to the descending eigenvalues, $\lambda_1 > \lambda_2 > \ldots > \lambda_p$, respectively.

The eigenvalues allow for a comparison of the importance of components, since they represent the amount of variance of $\boldsymbol{X}$ accounted for by each component of $\boldsymbol{Y}$. Pagès labels the quantity $\lambda_i \Big/ \sum_{j=1}^{p} \lambda_j$ the 'inertia' of component i, a marker of the usefulness of Y_i [142]. Often, components below some arbitrary eigenvalue threshold are discarded, as they do not provide a substantial amount of information about $\boldsymbol{X}$ [41]. In this thesis, discarding of lower-ranked components will be done in a manner informed by the political science domain in which we work; components will be used in descending order of eigenvalue until they are no longer informative, after which point they will be discarded.

Equal eigenvalues

The case in which some of the eigenvalues are equal presents a difficulty wherein the associated eigenvectors are no longer unique. In this case, any k equal eigenvalues will be associated with a set of k eigenvectors lying in a k-dimensional subspace of $\mathbb{R}^p$ orthogonal to the other $p - k$ eigenvectors. To fulfil the orthogonality requirement, the components can be chosen by selecting any orthonormal basis for this subspace, along with the other $p - k$ eigenvectors. This issue rarely appears in practice, since sample covariances are usually used, reducing the likelihood of exact orthogonality within $\boldsymbol{X}$.

3.1.3   PCA in practice: using samples from X

In obtaining the components above, the covariance matrix Σ was used. Unfortunately, this is not available when the variables $\boldsymbol{X}$ are only known through observations thereof. Hence to estimate the principal components the sample covariance matrix S can be used instead of the covariance matrix Σ.

As PCA is usually an exploratory analysis technique, the sampling error arising from this is not discussed here. Because we are not interested in directly predicting the observed values of the variables $\boldsymbol{X}$, but rather of exploring associations between the transformed variables $\boldsymbol{Y}$, we do not need to be concerned with the error associated with the transformation; we are not making predictions about particular subjects.

Errors associated with sampling are discussed by Osborne and Costello (2004), and Guadagnoli and Velicer (1988) [141, 79].
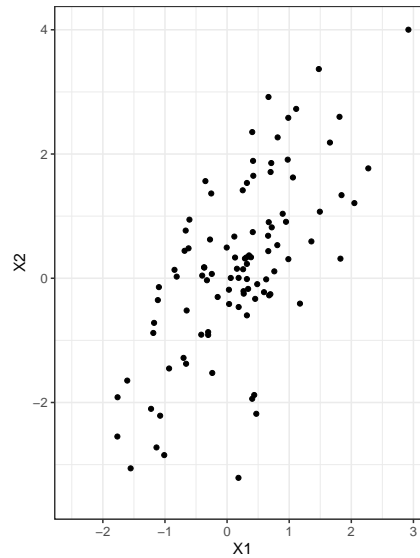
Figure 3.1: Plot of a simulated two-dimensional random variable, to demonstrate Principal Component Analysis. Note that X_1 and X_2 are strongly correlated.

3.1.4 A two-dimensional example

To visualise the operation of PCA, a two-dimensional example is presented in this section. This is done using a simulated two-dimensional observed random variable $\boldsymbol{X} = [X_1, X_2]$, where

$$X_1 \sim N(0, 1), \text{ and}$$
$$X_2 | X_1 \sim N(X_1, 1).$$

100 variates from $\boldsymbol{X}$ are plotted in Figure 3.1.

It can be observed in Figure 3.1 that X_1 and X_2 are strongly correlated. In fact, the sample covariance matrix from this particular data set is

$$S = \begin{pmatrix} 0.92 & 0.97 \\ 0.97 & 2.12 \end{pmatrix}.$$

The normalised eigenvectors of S are $[0.487, 0.873]$ and $[-0.873, 0.487]$, and correspond to eigenvalues 2.66 and 0.38. The first two principal components-transformed axes are thus given by

$$Y_1 = 0.487 X_1 + 0.873 X_2 \text{ and}$$
$$Y_2 = -0.873 X_1 + 0.487 X_2.$$

The two transformed axes can be seen in Figure 3.2. This figure demonstrates that the Y_1 value for each data point in $\boldsymbol{X}$ is given by that point's projection onto the red line—the representation of Y_1. Note that the axes are orthogonal to each other.
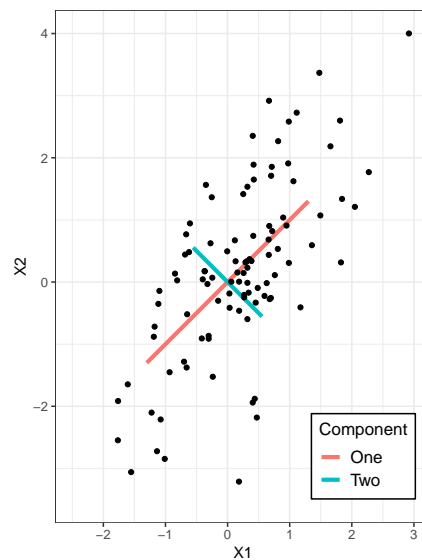
Figure 3.2: Plot of a simulated two-dimensional random variable, to demonstrate Principal Component Analysis, with its principal components overlaid. The red axis is the first principal component, and the blue axis is the second principal component. The lengths of the components are equal to their eigenvalues. Note that little information would be lost by representing all data just by their values along the red axis, while much information would be lost by representing the data just by their values along the blue axis.

If we sought to reduce the dimension of $\boldsymbol{X}$, we could discard Y_2, and retain just Y_1. This is because Y_1 is the the transformation that captures the most variation possible in the observed data set (see Equation (3.1.2)). In fact, with an eigenvalue of 2.66, compared to 0.38 for the component associated with Y_2, Y_1 captures

$$\frac{2.66}{2.66 + 0.38} = 87.5\%$$

of the variation in the observations from $\boldsymbol{X}$.

Thus, since this synthetic data set is quite strongly correlated, we can reduce its dimension by taking the first principal component, and discarding the other component. More generally, in PCA we reduce the dimension of some correlated set of variables by using the same number of principal components, and discarding axes below some eigenvalue threshold.

3.1.5 Scaling variables

If variables take different scales, different principal components will be derived. This can be observed heuristically from the fact that PCA intends to capture the most variance possible with each successive component; thus if one variable operates on a much larger scale than another, it will have greater variance the other, and so the PCA will be skewed in the direction of the variable on a much larger scale.

As a result of this, one common approach is to normalise all variables in the form

$$X_i^* = \frac{X_i}{\sqrt{\mathrm{Var}(X_i)}}$$

beforehand, called standardised PCA [41]. Standardised PCA is equivalent to applying PCA techniques to the sample correlation matrix of $\boldsymbol{X}$, rather than the sample covariance matrix. This reduces the potentially arbitrarily large effect variables which take wider ranges have on components, by forcing each variable to be of equal importance in component calculation. Arguably, this decision is just as arbitrary, however; it may be that in some given analysis, it is more important to explain some variables than others, so a bespoke approach may make output more relevant when targeted exploration is the goal of the analysis.

3.1.6 Advantages and disadvantages of PCA

PCA is a popular technique for two main reasons, the first of which is that the components are easy to calculate; anybody with a personal computer has access to eigenvector computation algorithms.

More importantly, PCA does not require any assumptions to be made about the underlying structure of the data [41]. For example, no assumption is made that the data follows any particular distribution. This means PCA can be applied to any multidimensional data set, subject to the proviso that PCA is only useful when the variables $\boldsymbol{X}$ covary in some meaningful way.

A disadvantage of PCA is that it does not take advantage of any distributional form that exists within the data. By contrast, a technique like Factor Analysis acknowledges and makes some attempt to quantify the error structure present in the data. Further, the scaling issue discussed above means practitioners of PCA need to make active choices about weighting the variables, meaning the researcher's subjective beliefs can have almost as much impact on the reduced dimensions as the data itself.

Finally, PCA cannot directly deal with non-continuous data. PCA assumes that data can take any value on the real line, making it unclear how to deal with categorical, or qualitative, data. Other techniques for categorical data, most notably Multiple Correspondence Analysis, exist (Section 3.2), and hybrid quantitative/categorical techniques such as Factorial Analysis of Mixed Data will be discussed in Section 3.3.

3.2   Multiple Correspondence Analysis

Closely related to Principal Component Analysis, Multiple Correspondence Analysis (MCA) is an unsupervised dimension reduction technique with no underlying assumptions about the distribution of $\boldsymbol{X}$. Unlike PCA, MCA deals solely with categorical variables. Historically, MCA has been treated as an expansion of Correspondence Analysis [78], which is used to deal with contingency tables relating to two categorical variables. However, it can also be expressed as an adjustment to PCA. Placing MCA in the same framework as PCA allows for methods dealing with mixed data to be pursued [142].

3.2.1   MCA as a variation of PCA

The following treatment of MCA, as PCA applied to an adjusted and weighted matrix W, is as described by Pagès (2016) [142]:

The categorical data for MCA need first be placed in a disjunctive table X of all the categories across the variables; that is, each entry x_{lj} is equal to 1 if subject l falls within j, and 0 otherwise. There are thus a number of columns dedicated to each variable–one for each of its levels. Each column is then weighted by dividing by its mean p_j. For example, for a subject l possessing a level j shared by one-fifth of their cohort, we multiply x_{lj} by 5.

Weighting the variables is valuable in that it increases the influence of rarer attributes, which are usually more characteristic of those possessing them than more common attributes. To draw an example from demography, suppose subjects are asked about their current employment status, as well as their annual income. Knowing that a person is unemployed (a rarer level) provides more information to a researcher than knowing that they earn less than $100,000 per year (a more common level). Since it is more characteristic, the rarer category of unemployment is given greater leverage in MCA.

Finally, the data is mean-centred. Since the columns have been scaled by their mean, each has mean 1, so mean-centring entails subtracting 1 from all values. As

such, the adjusted disjunctive table W is defined such that $w_{lj} = x_{lj}/p_j - 1$. MCA is simply PCA applied to W. The output $Y = \boldsymbol{a}^{\mathrm{T}} W$ is used as in Section 3.1.

The method of MCA can thus be summarised as:

1. Define disjunctive table X as follows:

$$x_{lj} = \begin{cases} 1 & \text{if subject } l \text{ possesses category } j \\ 0 & \text{otherwise,} \end{cases}$$

for $l = 1, \ldots, n$, $j \in J$; $J = J_1 \cup J_2 \cup \cdots \cup J_p$; and J_m the set of all levels of the mth categorical variable.

2. Define adjusted disjunctive table W as follows:

$$w_{lj} = \frac{x_{lj}}{p_j} - 1, \tag{3.2.1}$$

where $p_j = \dfrac{1}{n} \displaystyle\sum_{l=1}^{n} x_{lj}$.

3. Apply Principal Component Analysis to W, using the sample covariance matrix obtained by treating the rows of W as independent observations from some set of $|J|$ variables $\boldsymbol{W}$.

3.2.2 Weighted MCA

Sometimes it may be natural to group categorical variables according to themes in the data. A survey, for example, might have themes of equal value to a researcher, but have different numbers of questions about each theme. A solution to this is weighted MCA, which groups variables according to theme.

It can be observed above that each column in a standard MCA is weighted according to its mean. This means that for each variable, the sum of the weights p_j of its levels is 1. In a weighted MCA, variables can be grouped, and scaled such that the sum of the weights of all categories in each group is also 1. Such weighting is also necessary when undertaking Multiple Factor Analysis on qualitative groups (discussed in Section 3.4).

3.2.3 Describing the variables

The MCA method above does not make use of, or attempt to describe, the categorical variables; rather, the components are derived from correlations between individual categories. Determining what influence each variable has over each component is thus a task that must be undertaken separately. Since we seek to measure the relationship between a categorical variable (an observed categorical variable M) and a quantitative variable (a component Y_i of $\boldsymbol{Y}$), the estimated

squared correlation ratio is used [142]. In this circumstance, the ratio can be simplified to be

$$\eta^2(M, Y_i) = \frac{\sum_{k=1}^{|J_M|} n_k \overline{Y_{i,k}}^2}{n} \Bigg/ \sum_{l=1}^{n} Y_{il}^2,$$

where M is the categorical variable, $|J_M|$ is the number of categories in M, n_k is the number of observations possessing category k, Y_{il} is the lth observation of component Y_i, and $\overline{Y_{i,k}}$ is the mean value of Y_i for observations falling within category k.

The squared correlation ratio takes a value between 0 and 1, where value of 1 indicates that the variable M is completely coincident with the axis Y_i.

3.3 Factorial Analysis of Mixed Data

Since MCA and PCA can be framed as identical procedures applied to different forms of data, we may desire to combine the approaches in order to transform a mixture of qualitative and quantitative variables to a set of quantitative components. This process is called Factorial Analysis of Mixed Data ('FAMD') [142].

In order to undertake FAMD, a matrix of observations of the set of both quantitative and qualitative variables must be produced. Each quantitative variable takes up one column; it is normalised such that it has variance one, and is centred about zero. The qualitative variables are presented as they are in MCA—each level of each variable is placed in a column, each column is divided by its mean, and 1 is subtracted from every entry. The matrix produced is called the 'complete disjunctive table'. In FAMD, unstandardised PCA is applied to the complete disjunctive table of the quantitative and qualitative variables.

3.4 Multiple Factor Analysis

Multiple Factor Analysis (MFA) presents an extension to FAMD in that it allows for like variables to be grouped, to balance these groups of variables' influence on the components [142]. For example, an opinion survey might ask respondents a number of questions about different issues, but focus on some issues more than others. MFA would allow for each issue to form a group of equal influence, reducing the potential for dimension reduction to be dominated by particular issues. Within each group, the process performs locally as PCA, MCA or FAMD, depending on the variable types contained within. It should be noted that Multiple Factor Analysis has no direct relation to Factor Analysis; the nominal resemblance is due to a mistranslation from the original French [142].

The process for Multiple Factor Analysis relies on weighting the columns of $\boldsymbol{X}$ so that the groups are of equal value, accounting for the correlation within groups. Firstly, each qualitative variable is scaled as it is in MCA, and each quantative

variable is normalised and mean-centred. Each column in a purely qualitative group I is then divided by the number of variables in that group N_I; this is to ensure that the sum of the weights of each group is equal to one, a requirement for weighted MCA but not for PCA or FAMD (see Section 3.2.2).

All groups must then be balanced. Instinctively, we might then attempt to weight the groups by dividing each column by the number of variables within that group. However, this would still lead to a closely correlated group dominating other groups. To account for this, separate PCA, MCA or FAMD analyses are taken of each group I, depending on which is applicable to the data within I. All columns are then divided by the largest eigenvalue of this analysis, λ_1^I. As such, a group is weighted by the following:

- $\frac{1}{\lambda_1^I}$ if the group contains quantitative variables, or is mixed, or

- $\frac{1}{\lambda_1^I N_I}$ if the group contains only qualitative variables.

Finally, unstandardised PCA can be applied to the complete adjusted matrix, yielding balanced principal components.

3.4.1 FAMD and MFA compared

Both FAMD and MFA are useful developments in dimension reduction as they allow for the analysis of mixed data. Both simply entail data transformation and then application of PCA.

MFA is preferred by many, including its creator Jérôme Pagès, for its additional flexibility; it allows the user to alter the focus of the data when undue emphasis has been placed on a particular area [142]. For example, if an opinion poll asks single questions on some issues, and then a large number on one other issue, MFA can correct for this. Where a study is initially balanced, all variables can be placed in separate groups, reducing MFA to FAMD.

However, this flexibility forces the data analyst to make subjective decisions. MFA applied to the same data can produce substantially varying results according to the choice of groupings [142]. For example, suppose an opinion poll, among other things, asks questions about finance, taxation, education, health, water and the environment. A data analyst could consider these all separate themes, or could group finance with taxation, and water with the environment. The former choice gives greater weight to the economics and sustainability areas, and the latter might under-emphasise them; both choices are somewhat arbitrary. In many circumstances, these weighting decisions should be made by the primary researcher, not a data analyst.

The decision to use MFA should thus not be taken lightly, since it assumes an imbalance exists in the data that a data analyst can identify and then rectify. By contradistinction then, FAMD assumes a natural balance exists in the data. One should be aware of the design assumptions implicit in choosing either method.

3.4.2   Ordinal variables in mixed analysis

Given that techniques exist to deal with both quantitative and qualitative variable dimension reduction, in practice we must decide whether to treat ordinal variables as quantitative or categorical. Ordinal variables have the natural ordering of quantitative variables, but take only a few categorical values. If treated as quantitative variables, the categories will have the same linear association with all principal components, meaning components will be much easier to interpret. This is especially true if the ordinal variable has a large number of categories; for example, a variable containing respondents' incomes in $10k/year brackets should be treated as a quantitative variable, rather than a qualitative variable with many categories.

On the other hand, where the categories are associated in a non-linear way, treating the categories as distinct may be advantageous. This is especially important to consider when we are unsure whether the categories are linearly related. For example, suppose survey respondents are asked a question with the responses "Strongly agree", "Agree", "Neither agree nor disagree", "Disagree" or "Strongly disagree" (the Likert scale [112]). It may be that there is a closer association between the "Disagree" and "Strongly disagree" respondents, who are on the same side of the issue, than between the "Disagree" and the "Neither agree nor disagree" respondents. It may contrastingly be that the "Strongly" held views align with one another in some way, because people with extreme beliefs may have things in common. In either case, treating the variable as qualitative allows for the non-linearity presented by this scenario.

As such, we might undertake preliminary analysis treating ordinal variables as qualitative. This can determine whether any non-linearity exists in the effect of the ordinal categories. If the assumption of linearity does appear justified, the variables can be treated as continuous in the main analysis.

3.5   Multidimensional Scaling

Like PCA-related methods, multidimensional scaling (MDS) is a dimension reduction method with no underlying distributional assumptions about the data. MDS seeks a low-dimensional approximation of high-dimensional data using a method which attempts to preserve distances between subjects. This highlights the difference in focus between PCA and MDS: PCA focuses on 'explaining' covariance, while MDS focuses on the distances between subjects.

MDS first requires a distance metric to be defined between all variables. For quantitative variables, simple distance metrics include Euclidean distance, or L_2 distance, defined as

$$d_{i,j}^2 = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2},$$

where $d_{i,j}^2$ is the Euclidean distance between the vectors $\boldsymbol{x_i} = [x_{i1}, \ldots, x_{in}]$ and

$\boldsymbol{x_j} = [x_{j1}, \ldots, x_{jn}]$, and the Manhattan distance, or L_1 distance, defined as

$$d_{i,j}^1 = \sum_{k=1}^{n} |x_{ik} - x_{jk}|,$$

where $d_{i,j}^1$ is the Manhattan distance between the vectors $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$.

Given that distance metrics can be determined, the distances between all pairs of subjects can be contained in a matrix D. For some given lower dimension d of the data, the task then becomes to find an arrangement of the subjects in $\mathbb{R}^d$ such that the distances between all pairs of subjects in $\mathbb{R}^d$, D', is as close to the true distances D as possible. This 'closeness' may be defined by any of a number of loss functions. The loss function defined by Torgerson (1952) and Gower (1966) establishes a method often called 'classical multidimensional scaling' [169, 77]. It can be demonstrated that the results of Torgerson's MDS are equivalent to taking the first d components derived from PCA performed on the correlation matrix [84].

The drawing of k components from standardised PCA might thus be considered a special case of multidimensional scaling. MDS can also be applied to other loss functions, including commonly [159, 107] the least squares criterion

$$\sum_{i=1}^{n} \left\| \boldsymbol{D}_i' - \boldsymbol{D}_i \right\|^2,$$

where $\boldsymbol{D}_i'$ and $\boldsymbol{D}_i$ are the ith columns of D' and D respectively.

3.5.1 Categorical variables and MDS

Multidimensional scaling requires the distances, or dissimilarities, between all observations of variables to be calculable. For continuous variables, this is simply defined; one can take the absolute difference between observed values. Distance is harder to define for categorical variables, as it is difficult to establish which responses are more or less alike. For example, if subjects are asked what religious views they hold, a researcher might need to determine the distance between any pair of religious affiliations. This problem makes MDS, like standard PCA, infeasible for most categorical variables.

3.6 Factor Analysis

By contrast to PCA-related methods and to MDS, Factor Analysis is a model-based dimension reduction technique, with assumptions both on the number of components to be extracted and on the relationships between them. The p variables $\boldsymbol{X}$ are mean-centred, and reduced by expressing them in the form

$$\boldsymbol{X} = \Lambda \boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{3.6.1}$$

where $\boldsymbol{f} = [f_1, f_2, \ldots, f_k]^\mathrm{T}$ are the k dimension-reduced factors, Λ is a $p \times k$ matrix of coefficients, and $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p]^\mathrm{T}$ are error terms.

Equation (3.6.1) appears similar to a multivariate regression, as we express one set of variables as a linear combination of some others, with an error term. However, in this instance, a least squares solution is not immediately attainable; unlike in a multivariate regression, here $\boldsymbol{f}$ is unknown in addition to Λ and $\boldsymbol{\varepsilon}$.

As such, additional constraints need to be placed on $\boldsymbol{f}$ and $\boldsymbol{\varepsilon}$. Chatfield and Collins (1980) state common assumptions of Factor Analysis as follows [41]:

- the factors $\boldsymbol{f}$ are independent of one other;

- the residuals $\boldsymbol{\varepsilon}$ are independent of one another;

- $f_i \sim N(0, 1), i = 1, 2, \ldots, k$; and

- $\varepsilon_j \sim N(0, \psi_j), j = 1, 2, \ldots, p$, for some unknowns ψ_j.

Using this, we can write $\boldsymbol{f} \sim N(\boldsymbol{0}, I)$, and since independent univariate normal random variables also form a joint multivariate normal distribution, $\boldsymbol{\varepsilon} \sim N(0, \Psi)$, where $\Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)$. Thus $\boldsymbol{X} \sim N(\boldsymbol{0}, \Lambda\Lambda^{\mathrm{T}} + \Psi)$. Defining $\text{Var}(\boldsymbol{X}) = \Sigma$, finding the factors becomes a matter of solving

$$\Sigma = \Lambda\Lambda^{\mathrm{T}} + \Psi. \tag{3.6.2}$$

There are two issues in this being a solvable system of equations. Firstly, there need be at least as many equations as parameters. Here, we have $pk + p$ parameters, since there Λ is $p \times k$ and Ψ contains p values, and $\frac{1}{2}p(p+1)$ equations, since Σ is a symmetric $p \times p$ matrix. Thus there are more equations than parameters if $k + 1 < \frac{1}{2}(p+1)$.

A more crucial issue is that it can be easily shown that any solution for Λ is not unique, as $\Lambda\Lambda^{\mathrm{T}}$ is invariant under rotation of Λ. If we take any orthogonal (rotation) $k \times k$ matrix A, $(\Lambda A)(\Lambda A)^{\mathrm{T}} = \Lambda AA^{\mathrm{T}}\Lambda^{\mathrm{T}} = \Lambda\Lambda^{\mathrm{T}}$. As a result of this, additional constraints usually need to be placed to make the system solvable. One approach is outlined below.

3.6.1 The maximum likelihood method

Since it is assumed that $\boldsymbol{X}$ is normally distributed, we can write its likelihood and log-likelihood function.

Suppose we have n observed data points from $\boldsymbol{X}$, $\boldsymbol{x_i} \sim N(\boldsymbol{0}, \Sigma), i = 1, 2, \ldots, n$. We can then write the likelihood and log-likelihood function for $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}]$,

$$F(X; \Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{x_i}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x_i}\right), \text{ and}$$

$$l(\Sigma; X) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{x_i}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x_i}.$$

If we define $S = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x_i} \boldsymbol{x_i}^{\mathrm{T}}$, it can be shown that $\mathrm{tr}(S\Sigma^{-1}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x_i}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{x_i}$, and thus

$$l(\Sigma; X) = -\frac{n}{2}(p \log(2\pi) + \log|\Sigma| + \mathrm{tr}(S\Sigma^{-1})). \tag{3.6.3}$$

It should be noted that since Σ is a function of Λ and Ψ, the likelihood function can also be interpreted as a function of these parameters.

Jöreskog (1967) demonstrates [103] that maximising (3.6.3) is equivalent to minimising

$$H(\Lambda, \Psi) = \log|\Sigma| + \mathrm{tr}(S\Sigma^{-1}) - \log|S| - p; \text{ where } \Sigma = \Lambda\Lambda^{\mathrm{T}} + \Psi, \tag{3.6.4}$$

notable because at the MLEs $\hat{\Lambda}$ and $\hat{\Psi}$, $nH(\hat{\Lambda}, \hat{\Psi})$ is the likelihood ratio test statistic for the estimate of variance matrix Σ.

Approximate solutions to Equation (3.6.4) can be then found using an iterative method described in detail by Jöreskog (1967) [103]. Since the solutions found are invariant under orthogonal transformation, the factors can be rotated in such a way that they might then be interpreted more easily.

3.6.2 Applications of Factor Analysis

Factor Analysis, since it relies on a statistical model, is useful in that it accounts for the error in dimension reduction. This means goodness-of-fit tests like the likelihood ratio test can be applied to Factor Analysis estimates, unlike MDS or PCA-based approaches.

Another feature distinguishing Factor Analysis is that the factors are invariant under rotation, allowing us to rotate the factors for ease of interpretation. One such rotation, the varimax method [41], seeks that the observed variables $\boldsymbol{X}$ have loadings on only a small number of factors, making the real-world meaning of the factors more discernible. Chatfield and Collins, by contrast, see this ability to rotate factors as a disadvantage [41]. They argue that this rotational invariance can lead to researchers obtaining entirely different factors from the same set of variables.

Criticism has also been levelled at Factor Analysis as a result of the large number of assumptions necessary to produce a solution [41, 90]. The assumption of normality, for example, greatly restricts the variety of datasets to which Factor Analysis should be applied, as normality is often an unreasonable assumption to make.

Further, the number of underlying factors k needs to be predetermined, a difficult task without pre-existing research suggesting what latent factors are present in a particular area. This can lead to substantial error, since the choice of the number of factors can have a large effect on the factors' composition.

Due to these substantial criticisms, Chatfield and Collins cite Hills (1977) [90] in concluding that Factor Analysis is "not worth" carrying out [41]. For these reasons, which are also described in Chapter 4, Factor Analysis will not be used in this thesis; it is described in this chapter both to describe the wide variety of available techniques, and since Factor Analysis is particularly popular in social

science analysis [41]; previous literature in the political science field has regularly used Factor Analysis (see Section 4.5).

Despite these criticisms, in some certain circumstances, where its assumptions are valid and there is extensive literature discussing the number of latent factors that may apply, Factor Analysis' model-centric technique allows for a more thorough approach to dimension reduction.

3.7   Summary

This Chapter has presented six techniques for dimension reduction:

- Principal Component Analysis (Section 3.1);

- Multiple Correspondence Analysis (Section 3.2);

- Factorial Analysis of Mixed Data (Section 3.3);

- Multiple Factor Analysis (Section 3.4);

- Multidimensional Scaling (Section 3.5); and

- Factor Analysis (Section 3.6).

Each technique is valuable in certain applications. In Chapter 4, we select the most appropriate techniques for the demographic and political variables in the 2016 Australian Election Study data set, and apply the chosen methods. The dimension-reduced political and demographic variables can then be interpreted in context, and a model can be built linking political and demographic identity.

Chapter 4

Dimension Reduction Application

In Chapter 2, the 2016 Australian Election Study ('AES') data set was introduced. This data set is of high dimension, with 32 demographic variables and 103 political variables. In Chapter 3, methods for reducing the dimension of high-dimensional data were introduced. In this chapter, we select the most appropriate method for dimension reduction from Chapter 3, and apply it to the 2016 AES data set introduced in Chapter 2. Section 4.1 outlines in more detail why dimension reduction is needed for the 2016 AES data set. This is followed by the selection of a dimension reduction method in Section 4.2. We then apply this technique, interpreting the dimension-reduced demographic variables in Section 4.3 and the dimension-reduced political variables in Section 4.4. We then explore how dimension-reduced views of political ideology have evolved over the last 250 years in Section 4.5, and place our political spectrum in that context in Section 4.6.

Once the dimension-reduced spectra of demographic and political identity have been produced, we can build a model connecting them; the process for doing this begins in the following chapter, Chapter 5.

4.1   The need for dimension reduction

As introduced in Chapter 2, the 2016 Australian Election Study contains 32 demographic variables, and 103 political variables. In this work, a model is sought to relate these demographic and political variables. Clearly, there are too many political variables to create a coherent set of response variables. While techniques exist for handling multiple responses, and these techniques will be used in this work, 103 variables is unwieldy for such analysis.

Further, the purpose of modelling political opinion here is not to predict, or explain, relationships between *individual* demographics and views on *individual* political issues. Rather, relationships between demographic *identity*, and some notion of an overarching political *identity*, are sought. If individuals' political thoughts can be represented by some underlying set of beliefs, which are realised through opinions on individual political issues, it is this underlying set of beliefs that is sought to be related to demography through modelling. As a result, we seek to replace the 103 observed political variables with some smaller set of representative

variables. Dimension reduction is the tool that enables this.

There are similar domain-based motivations for reducing demographic dimension. It is clear that some of the demographic variables are highly related to one another, and so rather than attempting to explain political belief using a convoluted combination of these, it is more useful for an explanatory model to explain political belief using the commonalities between them. An explanatory model is about connecting the constructs that data represents, rather than their realisations. In this instance, the observed variables are proposed to be just realisations of underlying demographic and political identities, so a model should connect the underlying variables. For example, the 2016 AES asks subjects about the number of years of tertiary education they have completed, as well as whether they have completed high school, and if so, at what age they completed high school. All of these variables are strongly correlated with each other. Rather than using each of these variables to individually predict political belief, it would be more reasonable in an explanatory modelling domain to relate the underlying variable 'level of education' with political beliefs.

From a modelling perspective, the need to reduce the number of predictors is also evident. While 32 predictor variables is not necessarily exceedingly high, it should be noted that each of these variables is categorical, with a number of levels.

For example, the question "In what country was your father born?" received 98 different responses. In a univariate regression context, the number of degrees of freedom added to a model with the addition of categorical variable with k levels is $k - 1$. In other words, without dimension reduction, the question "In what country was your father born?" would add 97 degrees of freedom to the model.

Naturally, including all 32 demographic variables as predictors would lead to overfitting, since every additional degree of freedom is an additional opportunity for a spurious correlation to be incorporated into the model. As a result of this, a method to reduce the number of predictor variables, while losing as little of the information in the data as possible, is desired. The dimension reduction methods explored in Chapter 3 are specifically designed to achieve this, and so will be applied in this Chapter.

4.2   Selection of dimension reduction methods

Chapter 3 outlines a number of methods for undergoing dimension reduction. Some methods can be used only for qualitative variables, while others can only be used for quantitative variables. A list of methods, and their compatibility with variable types, can be found in Table 4.1. This table shows the restrictions on feasible methodologies, depending on the types of data being analysed. In this research, multidimensional scaling will not be used, as establishing compatible distance metrics for combinations of quantitative and qualitative variables proves to be a limitation (see Section 3.5). Further, Factor Analysis will not be used for the reasons in Chatfield and Collins [41], and Hills [90] (outlined in Section 3.6.2), who argue that Factor Analysis is insufficiently robust and relies upon unrealistic assumptions. The remaining techniques to be chosen between are those based on or related to principal

Technique	Abbreviation	Qualitative	Quantitative
Principal Component Analysis	PCA	No	Yes
Multiple Correspondence Analysis	MCA	Yes	No
Factorial Analysis of Mixed Data	FAMD	Yes	Yes
Multiple Factor Analysis	MFA	Yes	Yes
Multidimensional Scaling	MDS	No	Yes
Factor Analysis	FA	No*	Yes

Table 4.1: Dimension reduction techniques, and their compatibility with variable types. *Note that Factor Analysis can theoretically be applied to a disjunctive table of categorical variables, but this is not explicitly addressed in relevant literature.

component analysis.

For the political variables, we first note that of the 103 variables, 98 are ordinal, and 5 are categorical nominal. Since in general, simpler techniques are preferred, it might be argued that the ordinal variables should be treated as categorical nominal, with the ordering ignored, allowing for Multiple Correspondence Analysis to be used. However, this would substantially impact the ability of a researcher to interpret the reduced variables at the end of the analysis. With each dimension-reduced variable being a linear combination of all the levels of all the un-reduced variables, it would become more difficult to draw out the most important factors relating to each component. This is because treating an ordinal variable as quantitative makes the variable one-dimensional, while treating an ordinal variable with k levels as categorical makes the variable $k - 1$-dimensional. Thus, the ordinal variables are much more easily interpreted when treated as quantitative (see Section 3.4.2). Since the categorical nominal variables cannot be recoded as numerical, the only available methods for analysis are Factorial Analysis of Mixed Data (see Section 3.3), and Multiple Factor Analysis (see Section 3.4).

As discussed in Section 3.4, Multiple Factor Analysis ('MFA') allows the person undertaking data analysis to weight groups of variables arbitrarily. Thus, MFA requires the analyst to make a subjective decision about the relative values of different groups of variables. As a result, MFA will not be preferred here, since we prefer to trusts social science researchers at ANU to decide what political and demographic questions are relevant and important, than our own personal judgement. Factorial Analysis of Mixed Data will thus be the preferred approach.

Of the demographic variables in the 2016 AES, 25 are categorical nominal, four are ordinal, and three are numerical (see Section 2.1). With so many categorical nominal variables, in light of a preference for parsimonious methods, MCA seems *prima facie* the method of choice. In order to undertake this analysis, the ordinal and numerical variables need to be re-coded as categorical nominal. This was not preferred for the 98 ordinal variables in the case of the political part of the survey, since splitting 98 ordinal variables into all their constituent levels would substantially increase their dimensionality, but this is not a substantial issue for the

demographic variables, of which only four are ordinal.

Further, as will be discovered in Section 4.3, the preference for MCA allows non-linear cohort effects to be captured in dimension reduction, which are present in demography but not in political ideology. For example, by splitting age into distinct categories, rather than treating it as quantitative, we can allow for features of middle-aged people to be captured which are not present either in younger or older generations.

Re-coding the ordinal variables to nominal categorical is easy to do. To re-code the three numerical variables to nominal categorical, values are restricted into ranges, which form categories. Likewise, to reduce the number of levels for the categorical variables with a large number of responses, these are also re-coded into broader categories. These re-codings are summarised in Table 4.2. After recoding, there are 115 components of the original (and dimension-reduced) demographic spaces, since

$$\sum_{i=1}^{p}(|J_i| - 1) = 115,$$

for J_i the number of levels of categorical variable J, and p the number of variables. This is because the dimensionality of a particular variable is the number of levels subtract one, since the final level is linearly dependent on the others.

Question	Previous encoding	New encoding	Previous dimensionality	New dimensionality
How many years' tertiary study have you completed?	Numerical	0, 1-2, 3-4, 5-8, 9+	1	5
How old were you when you finished secondary school?	Numerical	15, 16, 17, 18, 19, 20+	1	6
In what year were you born?	Numerical	[Ages] 18-34, 35-49, 50-64, 65+	1	4
In which country were you born?	[Country name]	Australia, Overseas	92	1
In which country was your mother born?	[Country name]	Australia, Overseas	97	1
In which country was your father born?	[Country name]	Australia, Overseas	98	1
What is your household's gross annual income?	[$10,000 ranges]	$0-$25k, $26k-$50k, $51k-$80k, $81k-$120k, $121k-$160k, $161k+	22	6

Table 4.2: Recoding of demographic variables from the 2016 Australian Election Study.

4.3   Interpretation of demographic components

MCA was applied to the demographic variables with recoding as outlined in Table 4.2. Since MCA is simply a change of basis, there are as many components produced as there are dimensions in the original demographic space. A natural consequence of the variance-maximisation MCA procedure is that many of 115 transformed variables have very low influence, and can be discarded.

In fact, in total, 65 of the 115 transformed variables have associated eigenvalue less than the average eigenvalue, indicating that they explain less of the original dataset's variance than did the average level within the original dataset. The proportion of variance explained by each dimension in sequence is shown in Figure 4.1, which demonstrates that while the first three reduced demographic variables explain substantially more variance than any single one of the original levels was able to, the values of the next variables decline thereafter. This is to be expected for MCA, which in general produces smaller eigenvalues than PCA; in MCA, each categorical variable is associated with a subspace of the original variable space, while a quantitative variable is associated with just a single axis [142]. MCA cannot reduce the dimension of the subspaces associated with each categorical variable, but rather reduce the overall dimension by taking advantage of associations between the subspaces [142].

Tables 4.3, 4.4 and 4.5 contains the key survey questions influencing each of the six reduced demographic variables with the highest associated eigenvalues. The first six reduced variables are chosen, since these are the only reduced demographic variables whose eigenvalue suggests they explain more than twice as much variation within the data as a single level within the original data set. The table also lists the estimated squared correlation ratios between each reduced demographic variable and the most relevant survey questions.

The following subsections discuss how the first three of these reduced demographic variables, those with very high eigenvalues, might be interpreted. Each dimension-reduced demographic variable is briefly explained using the demographic variables most correlated with it. Since all of the reduced demographic variables are a linear combination of all of the observed demographic variables, in reality each of the reduced variables is much more complicated than this depiction.

These three reduced variables do not form a complete picture of demographic identity; all 115 components would still be needed to fully represent the demographic space. However, the three components here are a more complete description of demographic identity than any other three linear combinations of the observed demographic variables.

Note: in the following sections, the phrase 'mean coordinate' is used to refer to the average value on the relevant axis of an individual who has a stated characteristic. For example, the statement 'on the first component, completion of postgraduate education has mean coordinate 0.800' indicates that among individuals who have a postgraduate qualification, the sample mean coordinate on the first dimension-reduced demographic axis is 0.800.

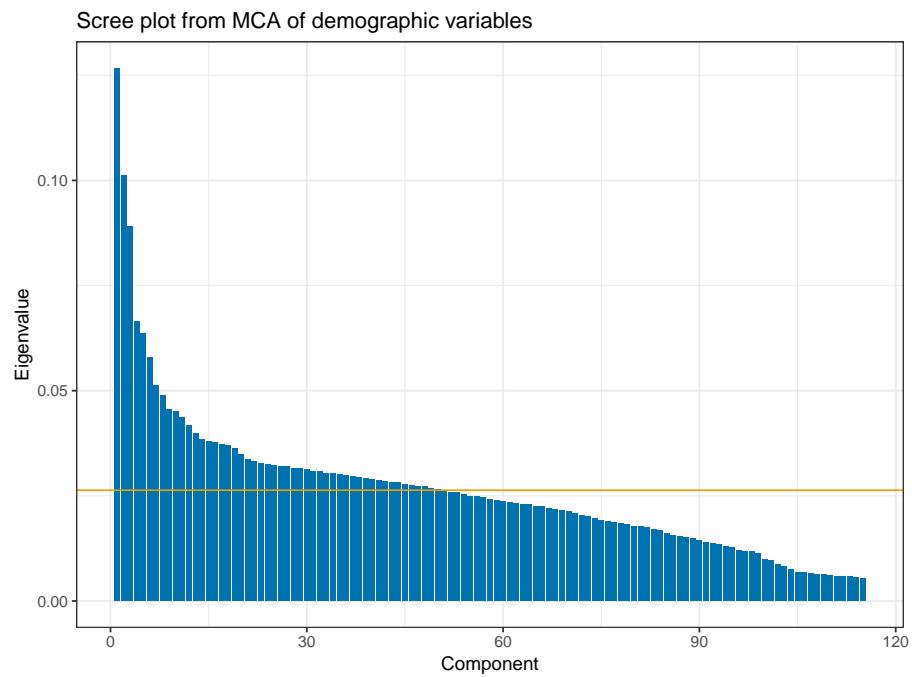Figure 4.1: Scree plot from MCA of the demographic variables, with a horizontal line representing the average eigenvalue. Of the 115 reduced components, 50 capture a larger proportion of the variance in the demographic variables than a the average level in the untransformed space. After the first three components, there is a sharp drop-off in variables' explanatory power, with a roughly linear decline after roughly 10 components.

Further, η^2 indicates the squared categorical correlation ratio, and cor^2 indicates the squared Pearson correlation coefficient, between the reduced variable and the observed variable in question.

4.3.1   The first component: 'Socio-economic status and Education'

The first reduced demographic variable is most highly correlated with gross annual income ($\eta^2 = 0.564$), employment status ($\eta^2 = 0.475$), year of birth ($\eta^2 = 0.430$) and level of tertiary education ($\eta^2 = 0.362$). Individuals with a high value on the first reduced demographic axis are more likely to have participated in high school (mean coordinate 0.05), have an annual income over \$160,000 (mean coordinate 1.01), and be between 35 and 49 years old (mean coordinate 0.746), while a low value is associated with a lack of formal schooling (mean coordinate -1.804), living in public housing (mean coordinate -1.397), and currently looking for full-time work (mean coordinate -0.168).

As a result of the first reduced component's strong correlation with variables associated with current income and amount of education, the component is labelled 'socio-economic status and education'. This moniker is indicative of the types of characteristics possessed by people at different extremes of this component, but does not give a complete description; as seen in the preceding paragraph, being between 35 and 49 years old, for example, is related to the variable. This is because there are a large number of factors which correlate with income, and these factors are not limited to level of education. It is this high level of correlation between education, income and the variety of other demographic variables in the AES 2016 data set that makes the 'socio-economic status and education' variable the component associated with the highest variance.

Another interesting feature of the first component is that despite being correlated with level of education, the state of having participated in high school has a mean coordinate of just 0.05–a relatively low magnitude. This can be explained by the fact that 97.1% of survey participants fall into this category, dulling its effect; by design, rarer levels are more likely to take extreme values in MCA (see Section 3.2, and Pagès 2016 [142]). By contrast, the somewhat rarer category of having undertaken between 5 and 8 years of tertiary education has a mean coordinate of 0.278.

4.3.2   The second component: 'Stage of Life'

The second reduced demographic variable is most highly correlated with the survey questions "Do you own outright, [mortgage] or rent the dwelling in which you now live?" ($\eta^2 = 0.449$), "In what year were you born?" ($\eta^2 = 0.378$), and employment status ($\eta^2 = 0.377$). The second reduced variable is thus most closely related to similar variables as the first variable, but closer inspection reveals the information captured by the second demographic variable is quite different to that of the first. While the first reduced demographic variable contrasts those who are

well-educated, currently employed, and middle aged, with those who are not, the second demographic variable more strongly captures a generational divide between older people, who are more likely to be married and own their own home, with younger people, who are more likely to be renting or be single.

Individuals who score highly on the second reduced demographic variable are more likely to be older than 65 (mean coordinate 0.562), own shares in more than 10 companies (mean coordinate 1.199), and be retired from paid work (mean coordinate 0.643). By contrast, those with a low score are more likely to be full time students (mean coordinate -2.569) or unemployed (mean coordinate -1.530), be between 18 and 34 years old (mean coordinate -1.456) or have never been married (mean coordinate -1.203). This focus on levels relating to age, and characteristics associated with age, lead to the labelling of this component 'stage of life'.

Other less obvious levels relating to one's 'stage of life' are also relevant in the second categorical variable. For example, Australia's Muslim population is younger than the population at large, and this is reflected in the second component, with the mean coordinate of Muslim subjects being -1.429. For similar reasons, the mean coordinate of Aboriginal and Torres Strait Islander people on the second component is -0.798.

As has been observed in this section, there is a distinction between the first and second reduced demographic variables, despite these variables both being strongly related to the same key demographic survey questions. This is because it is different levels within each survey question that relate to each of the reduced demographic components. This might be regarded as a key reason why, when given the choice between treating an ordinal variable as categorical nominal or continuous, ordinal variables are sometimes treated as categorical in an MCA, as opposed to as ordinal in a mixed analysis like FAMD (see Section 3.3); MCA allows for categories to be caught by multiple components in non-linear ways. For example, here age was treated as a categorical variable, rather than as a continuous variable, and was related to both the first and second reduced demographic variables. In MCA, levels common among older and younger people (lower incomes) can be grouped and juxtaposed against levels common among middle aged people (higher incomes), while this cannot be done if age was treated as continuous.

4.3.3 The third component: 'Cultural Background'

The third reduced demographic variable is most highly correlated with the survey questions "In which country was your mother born?" ($\eta^2 = 0.659$), "In which country were you born?" ($\eta^2 = 0.649$), and "In which country was your father born?" ($\eta^2 = 0.627$). The next most related variable is religion ($\eta^2 = 0.209$). Thus, this variable is labelled 'cultural background'. Looking to survey participants, those with positive scores on the 'cultural background' axis were more likely to be born outside Australia (mean coordinate 1.377), and identify their religion as one of Muslim (mean coordinate 2.336), Hindu (1.996), Orthodox (1.457), Buddhist (1.456), Jewish (1.125), or 'Catholic other than Roman' (1.052). Those with negative scores on the 'cultural background' axis were more likely to be born in Australia (mean coordinate

Component 1 (4.18% of variance)

Question	Squared correlation ratio
What is the gross annual income, before tax or other deductions, for you and your family or others living with you?	0.564
Last week, what were you mainly doing? (employment status)	0.475
In what year were you born?	0.430
Have you obtained a trade qualification, a degree or a diploma, or any other qualification since leaving school?	0.362
What was your partner's main activity last week? (employment status)	0.358

Component 2 (3.34% of variance)

Question	Squared correlation ratio
Own outright, buying or renting the dwelling in which you now live?	0.449
In what year were you born?	0.378
Last week, what were you mainly doing? (employment status)	0.377
What is your current marital status?	0.273
Which of the following best describes the position that you hold (or held most recently)? [Range "non-supervisory" to "upper managerial"]	0.247

Table 4.3: The five most correlated survey questions with the first and second components of the demographic Multiple Correspondence Analysis.

-0.471), have a mother born in Australia (-0.600), have a father born in Australia (-0.631), and identify as Aboriginal or Torres Strait Islander (mean coordinate -0.638).

It is notable that a more extreme mean coordinate is associated with having parents being born in Australia, than with being born in Australia oneself; this is because being a nation with a large migrant population, the category of having parents born in Australia is rarer than being born in Australia oneself.

4.4   Interpretation of political components

FAMD was applied to the political variables. Since the original political variables comprised a 123-dimensional space, the transformed variables is also 123-dimensional. 92 of the 123 transformed variables have associated eigenvalue less

Component 3 (2.94% of variance)	
Question	Squared correlation ratio
In which country was your mother born?	0.659
In which country were you born?	0.649
In which country was your father born?	0.627
What is your religion or faith?	0.209
In all, how many years of tertiary study have you completed since you left secondary school?	0.079

Component 4 (2.20% of variance)	
Question	Squared correlation ratio
Last week, what were you mainly doing? (employment status)	0.220
Own outright, buying or renting the dwelling in which you now live?	0.217
In what year were you born?	0.195
Whom do (or did) you work for? [self-employed/ family business/ public sector/ private sector]	0.148
Which of the following best describes the position that you hold (or held most recently)? [Range "non-supervisory" to "upper managerial"]	0.135

Table 4.4: The five most correlated survey questions with the third and fourth components of the demographic Multiple Correspondence Analysis.

Component 5 (2.10% of variance)

Question	Squared correlation ratio
Whom do (or did) you work for? [self-employed/ family business/ public sector/ private sector]	0.224
Last week, what were you mainly doing? (employment status)	0.208
Have you obtained a trade qualification, a degree or a diploma, or any other qualification since leaving school?	0.193
What was your partner's main activity last week? (employment status)	0.163
In what year were you born?	0.149

Component 6 (1.91% of variance)

Question	Squared correlation ratio
Last week, what were you mainly doing? (employment status)	0.271
Are you male or female?	0.237
Have you obtained a trade qualification, a degree or a diploma, or any other qualification since leaving school?	0.154
What is your religion or faith?	0.145
Which social class would you say you belong to?	0.125

Table 4.5: The five most correlated survey questions with the fifth and sixth components of the demographic Multiple Correspondence Analysis.
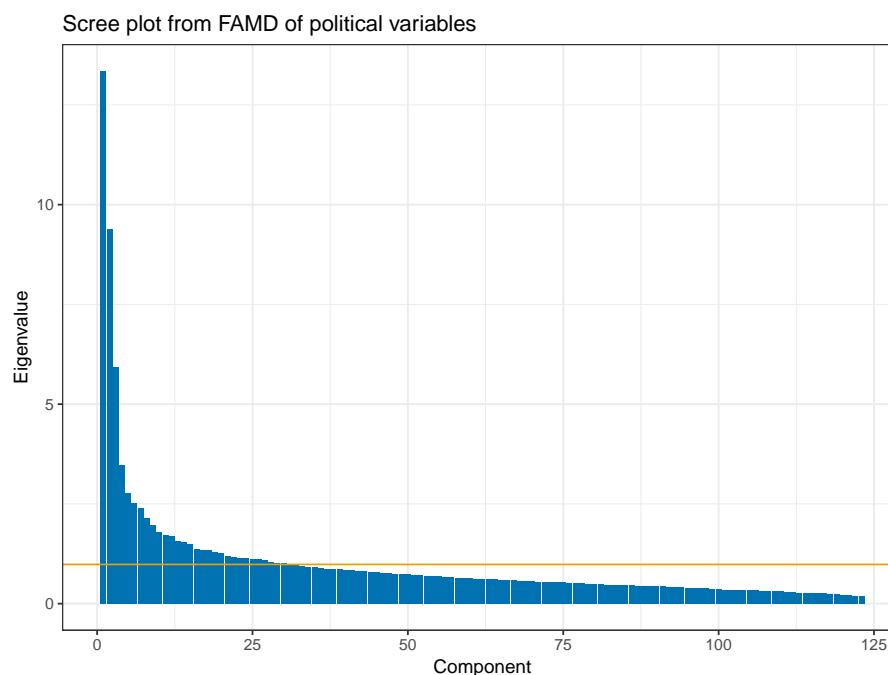
Scree plot from FAMD of political variables

Figure 4.2: Scree plot from FAMD of the political variables. Of the 123 reduced components, 31 capture a larger proportion of the variance in the political variables than a dimension in the un-transformed space. After the first three components, there is a sharp drop-off in variables' explanatory power, with a roughly linear decline after the first 20 variables.

than the average eigenvalue, indicating that much of the variance in the data is skewed towards the earliest principal components.

The proportion of variance explained by each dimension in sequence can be found in Figure 4.2, which demonstrates that the first three reduced political variables explain substantially more variance than the original variables were able to, and after roughly the 20th such variable, the variances of the reduced variables decline in a somewhat linear manner.

It is to be noted that the variance captured by these first few variables is substantially higher than in the MCA in Section 4.3. Potential reasons for this include the presence of a larger number of continuous variables, allowing for a larger number of dimensions to be explained simultaneously [142], and that demographic space may be of higher underlying dimension than the political space. This second explanation might hold if there are many more different sets of intersecting demographic identities than there are political philosophies.

Tables 4.6, 4.7 and 4.8 contains the key survey questions influencing each of the six reduced political variables with the highest associated eigenvalues. Tables 4.6, 4.7 and 4.8 also list the estimated correlations between each reduced political variable and the most relevant survey questions. In the following sections, the first three reduced political variables will be interpreted, with reference to the

survey questions and responses that characterise them. This will be followed by an explanation of the difficulty associated with interpreting the fourth political axis, which is of much less use in political science. Each of these descriptions are by necessity a simplification of the axes, since each is built as a linear combination of all 103 observed political questions, comprising 123 dimensions.

Following these descriptions, the axes will be discussed in light of existing research in the area of political opinion; spatial representations of political opinion have been discussed extensively in social science research (see, e.g., [34, 35, 137, 143, 151, 171, 176]), but there does not appear to be previous research in Australia using the rigorous data-driven approach used here.

4.4.1   The first component: Social inclusivity

The first reduced political variable is most highly correlated with questions regarding the policy of turning back boats carrying irregular maritime arrivals ($cor^2 = 0.589$), immigration numbers ($cor^2 = 0.531$) and perceptions of a relationship between immigration and crime ($cor^2 = 0.509$). Individuals with a high value on the first reduced political axis are more likely to:

- Strongly disagree with the statement "all boats carrying asylum seekers should be turned back" (mean coordinate 5.099);

- Be strongly in favour of allowing same-sex marriage (mean coordinate 2.126);

- Strongly disagree with the notion that immigrant increase rates of crime (mean coordinate 5.337)

- Think that changes advancing aboriginal land rights "have not gone nearly far enough" (mean coordinate 4.114);

- And strongly favour the decriminalisation of marijuana (mean coordinate 1.588).

All of these ordinal survey responses are treated as numerical variables for the purpose of this analysis, meaning that opposite responses for the survey questions above are also associated with mean coordinates at the opposite side of the first reduced political axis. For example, those who think that changes advancing aboriginal land rights "have not gone nearly far enough" have mean coordinate 4.114, while those who think that this change has "gone much too far" have mean coordinate -4.268.

While the most strongly correlated survey questions with the first reduced political variables are all related to immigration (see Table 4.6), the above demonstrates that immigration alone does not fully characterise this part of the political spectrum. Questions relating to LGBT+ people, Aboriginal and Torres Strait Islander peoples, and drug policy also form key divides on the first reduced political axis. For this reason, this reduced variable is referred to as the 'social inclusivity' axis, being related to a series of questions about how people treat those who are different

Component 1 (11.27% of variance)	
Question	Squared correlation
[To what extent do you agree] with the following statement? All boats carrying asylum seekers should be turned back	0.589
Do you think the following change [. . .] has gone too far, not gone far enough, or is it about right? The number of migrants allowed into Australia at the present time	0.531
[To what extent do you agree] with each of the following statements? Immigrants increase the crime rate	0.509
Do you think the number of immigrants allowed into Australia nowadays should be reduced or increased?	0.494
Do you think the following change [. . .] has gone too far, not gone far enough, or is it about right? Equal opportunities for migrants	0.473
Component 2 (7.92% of variance)	
Question	Squared correlation
How much confidence do you have in the [. . .] federal government in Canberra	0.542
How much confidence do you have in the [. . .] federal parliament	0.512
Would you say the government is run by a few big interests looking out for themselves, or that it is run for the benefit of all the people?	0.456
How much confidence do you have in [. . .] the Australian political system	0.429
How much confidence do you have in [. . .] Australian political parties	0.408

Table 4.6: The five most correlated survey questions with the first and second components of the political Factorial Analysis of Mixed Data.

Component 3 (4.98% of variance)

Question	Squared correlation
Should there be more or less public expenditure in the following area? Police and law enforcement	0.264
Should there be more or less public expenditure in the following area? The National Disability Insurance Scheme	0.260
Should there be more or less public expenditure in the following area? Old-age pensions	0.247
When you were deciding about how to vote, how important was each of these issues to you personally? Health and Medicare	0.240
Should there be more or less public expenditure in the following area? Child care	0.216

Component 4 (2.61% of variance)

Question	Squared correlation
How much confidence do you have in the following organisation? Television	0.191
How much confidence do you have in the following organisation? The press	0.159
How much confidence do you have in the following organisation? Trade unions	0.136
Should there be more or less public expenditure in the following area? Public transport infrastructure	0.127
In your opinion, are any of the following countries likely to pose a threat to Australia's security? Japan	0.119

Table 4.7: The five most correlated survey questions with the third and fourth components of the political Factorial Analysis of Mixed Data.

Component 5 (2.34% of variance)	
Question	Squared correlation
When you were deciding about how to vote, how important was each of these issues to you personally? Immigration	0.195
When you were deciding about how to vote, how important was each of these issues to you personally? Refugees and asylum seekers	0.188
Some people say it makes a big difference who is in power. Others say it doesn't make any difference who is in power. Using the scale below, where would you place yourself?	0.117
When you were deciding about how to vote, how important was each of these issues to you personally? The environment	0.106
Some people say that no matter who people vote for, it won't make any difference to what happens. Others say that who people vote for can make a big difference to what happens. Using the scale below, where would you place yourself?	0.100

Component 6 (2.16% of variance)	
Question	Squared correlation
Do you strongly agree, agree, disagree or strongly disagree with the following statement? Terminally ill patients should be able to end their own lives with medical assistance	0.218
When you were deciding about how to vote, how important was each of these issues to you personally? Taxation	0.198
When you were deciding about how to vote, how important was each of these issues to you personally? Superannuation	0.161
Which one of these statements comes closest to how you feel about abortion in Australia?	0.153
Do you think that Australia should become a republic with an Australian head of state, or should the Queen be retained as head of state?	0.128

Table 4.8: The five most correlated survey questions with the fifth and sixth components of the political Factorial Analysis of Mixed Data.

to them. Within this axis, it is views relating to immigration that are most characteristic of ones' social policy positions.

4.4.2 The second component: Attitudes to authority

The second reduced political variable is most highly correlated with questions regarding confidence in the federal government ($cor^2 = 0.541$), confidence in the federal parliament ($cor^2 = 0.512$) and whether "the government is run by a few big interests looking out for themselves, or that it is run for the benefit of all the people" ($cor^2 = 0.456$). Individuals with a high value on the first reduced political axis are more likely to:

- Have no confidence at all in the federal government in Canberra (mean coordinate 4.434);

- Believe the government is entirely run for a few big interests (mean coordinate 3.627);

- Have no confidence at all in large Australian companies (mean coordinate 4.102)

- Have no confidence at all in banks and financial institutions (mean coordinate 2.927);

- And be 'not at all satisfied' with the way democracy works in Australia (mean coordinate 3.701),

while those with low scores on this axis exhibit attitudes a greater level of trust in government, democracy, banks and large corporations. For this reason, the second political axis is labelled 'attitudes to authority'.

There may be some hesitation about referring to levels of trust as an explicitly political position, since the dominant survey questions relating to the second axis are not explicitly policy questions. However, politics is, at its essence, a study of the relationships between those in power, and the populace [10]. This means general views about authority are an inherently political position. A more extensive discussion of the importance of considering trust in authority in light of recent political developments can be found in Section 4.5.

4.4.3 The third component: Spending priorities

The third reduced political variable is most highly correlated with questions regarding expenditure in police and law enforcement ($cor^2 = 0.264$), expenditure in the National Disability Insurance Scheme ('NDIS') ($cor^2 = 0.258$), expenditure in old-age pensions ($cor^2 = 0.247$), and the importance of health and medicare to one's voting choice ($cor^2 = 0.240$). Individuals with a high value on the third reduced political axis are more likely to:

- Favour much more spending on police and law enforcement (mean coordinate 2.543);

- Favour much more spending on the NDIS (mean coordinate 2.709);

- Favour much more spending on old-age pensions (mean coordinate 2.114)

- Strongly agree that wealth should be "distributed towards ordinary working people" (mean coordinate 1.250);

Meanwhile, individuals with a low value on the third reduced political axis are more likely to:

- List 'management of the economy' as the most important factor influencing their vote (mean coordinate -0.792);

- List 'taxation' as the most important factor influencing their vote (mean coordinate -0.677);

- Strongly disagree that wealth should be "distributed towards ordinary working people" (mean coordinate -2.291);

- Rate "health and Medicare" as "not very important important" in influencing their vote (mean coordinate -3.553).

Since this axis contrasts those who favour spending on health, education and ageing, with those who are more concerned about the economy, and are less inclined towards increased government spending, the third component is referred to as the 'spending priorities' axis.

4.4.4 The fourth component

The fourth reduced political variable is most highly correlated with questions regarding confidence in television ($\text{cor}^2 = 0.191$), confidence in the press ($\text{cor}^2 = 0.159$), confidence in trade unions ($\text{cor}^2 = 0.136$), and whether there should be more investment in public transport ($\text{cor}^2 = 0.127$). Individuals with a high value on the fourth reduced political axis are more likely to:

- Think Japan is 'very likely' to pose a security threat to Australia (mean coordinate 1.656);

- Have a great deal of confidence in the press (mean coordinate 1.621);

- Rate "management of the economy" as "not very important" in influencing their vote (mean coordinate 1.310)

Meanwhile, individuals with a low value on the fourth reduced political axis are more likely to:

- Have no confidence in trade unions (mean coordinate -0.438);

- Have no confidence in television (mean coordinate -0.482);

- Think there should be much less expenditure on healthcare (mean coordinate -0.598).

There is no immediate connection to be drawn between the kinds of characteristics common among people with high or low values on the fourth reduced demographic axis, from a political science perspective. In order to interpret this axis, a political scientist may need to be consulted.

The fact that it is difficult to interpret the fourth reduced political axis is concordant with the sharp drop-off in variance captured by the fourth axis, relative to the first three. As visualised in Figure 4.2, a substantial decline in the amount of information captured by successive variables occurs after the third reduced variable. Numerically, the variance captured by the first 5 variables are 11.27%, 7.92%, 4.98%, 2.61%, and 2.34%. While a variable along which 11.27% of divergences in a 123-dimensional space may be of substantial interpretive value, a variable capturing just 2.61% of this divergence, as does the fourth reduced political variable, is much less useful.

It is concluded that analysis of the political spectrum should be concluded after the third reduced political variable is modelled. In other words, as represented in the 2016 AES data set, the Australian political spectrum dominantly consists of three clear and interpretable axes.

4.5   Literature review of the political spectrum

4.5.1   Early political spectra

Section 4.4's depiction of a political spectrum as some low dimensional space does not stand alone in the field of political science. In fact, the most simplistic representation of the political spectrum, as a dichotomy between two, categorical positions 'Left' and 'Right', originates from the French Revolution in 1789, when those in support of the King sat to the right of the president in the National Assembly, and those opposed sat to the left [71, 147]. The recognition of a spectrum of political thought came at the same time as the recognition of political power's derivation from the populace. Later in the summer of 1789, the same French National Assembly passed an affirmation that political authority 'proceed[s] directly from the nation', in Article III of the Declaration of the Rights of Man and the Citizen [51]. That democratic ideals emerged at the same time as greater consideration of the underlying political views of the people is no coincidence [71].

Throughout the 19th century, the initially rigid dichotomy in French discourse between 'Left' and 'Right' dispersed into more spectral terms than simply a designation of geographical position relative to the president. The terms 'centre-left', 'centre-right', 'far-left' and 'far-right' entered political discourse [71, 128], with the centrist views being distinguished by politicians from the extreme views ([168], quoted in [71]), indicating a perception that political opinions fell at some point on an axis, rather than just being a series of disparate positions. The 'Left' side of the

axis represented progressive and revolutionary views, such as democracy and rights for women, while the 'Right' side favoured conservatism and even counter-revolution [71].

With the rise of socialist thought in the 1890s and early 1900s, the use of a Left-Right spectrum gained credence across Europe and North America as it allowed for the contrast between a tendency toward universality and equality (the Left) with a tendency towards existing power structures (the Right) [71]. Gauchet writes that this is a similar underlying Left-Right divide to that which existed at the time of the French Revolution, since both contrast a desire to retain the status quo (the Right) with a desire to reduce the power of wealth and aristocracy (the Left) [71].

4.5.2 'Data-driven' approaches

Social scientists soon discovered, however, that a single Left-Right axis is insufficient to describe the range of underlying political identities that exist in an increasingly free and democratic society [57]. A new political spectrum was needed, and it was recognised that this spectrum should, in some part, be driven by data on citizens' views. The use of survey data to draw out the key underlying factors of political opinion was used in deeply flawed ways by Leonard Ferguson [62] and Hans J. Eysenck [57]. Their work, and its criticisms, will be discussed in the following paragraphs.

Ferguson's spectrum

The first isolation of a political spectrum using data-driven approaches was done by Leonard Ferguson [62]. In a series of papers from 1939 to 1942, Ferguson undertook Factor Analysis (see Section 3.6) to draw out a three-axis 'social attitudes' chart [62, 63, 64, 65, 66]. The Factor Analysis was based on a small sample of a small number of variables; the original analysis was based on 185 respondents' answers to questions on 10 scales [62, 67]. Adding to the difficulties of this analysis are that Ferguson chose the questions himself [66], leading to the output variables being influenced by what he thought were important political questions.

Further, Ferguson's Factor Analysis was done by trial and error, in a way deliberately designed to draw out the most substantial results. Ferguson initially struggled to find a unique solution in applying Factor Analysis to all of his ten variables, no matter how many reduced variables he sought [66]. He then split his ten variables into a group of six, and a group of four, choosing combinations of variables he thought important, and then subjecting the first six to a Factor Analysis with two output variables, and the next four to a Factor Analysis with one output [66]. He then concluded that the political spectrum comprised the three output variables he had drawn out, which he named 'religionism', 'humanitarianism' and 'nationalism' [64, 66].

In deliberately selecting groups of variables that would draw out desired results, Ferguson introduced bias to his dimension-reduction procedure (see Chapter 3). This combined with his small sample size, small number of variables, writing his own survey, and the use of the highly temperamental Factor Analysis procedure

(see Section 3.6.2) make his outputs untrustworthy. His work remains important, however, as the first attempt to build a political spectrum using survey data, rather than qualitative analyses.

The Eysenck Chart

Improving slightly upon the methods of Ferguson, Hans Eysenck published his version of a Factor Analysis-based political spectrum in his book *The Psychology of Politics* in 1954 [57], with the outputs discussed for a popular audience in *Sense and Nonsense in Psychology* (1957) [58]. The Eysenck Chart has become the standard representation of political thought in political science and the public imagination. For example, charts based on the Eysenck Chart are used in popular online tools PoliticalCompass.org [34], VoteCompass [171], and the (openly ideological) World's Smallest Political Quiz, which purports to have been taken over 20 million times [82].

However, reliance on Eysenck's chart should be grounded in a strong methodology, not just its ubiquity. As such, we now interrogate Eysenck's methodology.

Eysenck's work improved upon Ferguson's in the following ways [58]:

- The survey producing the political spectrum contained 60 variables, instead of just 10.

- Factor Analysis was undertaken on all variables simultaneously, instead of the variables being grouped prior to analysis. This reduces the level of intervention in the analysis on the part of the researcher.

However, Eysenck's work has been criticised [35, 152, 153] for maintaining the following errors, present also in Ferguson's work:

- Eysenck arbitrarily chose that there would be two output dimensions for his Factor Analysis, despite Factor Analysis being very sensitive to the number of dimensions chosen (see Section 3.6).

- Eysenck wrote the survey questions himself, based upon newspaper headlines he had read [58], meaning they are influenced by his view on what questions define political perspectives. For example, 13 of Eysenck's 60 questions specifically relate to religion [58].

The latter of these observations has been noted by Eysenck's critics [35, 152, 153] who note that Eysenck's own political thinking was heavily influenced by his own views, on race among other things; Eysenck believed in the refuted [163] notion that people of different ethnic backgrounds have differing levels of intelligence [59], and Eysenck was interviewed for the newspaper of the UK's far-right National Front [152, 153]. This might have led to his survey's inclusion of nine questions on race out of the 60 in total, and 13 on religion, meaning characteristics relating to views on race and religion were more likely to be drawn out by the Factor Analysis. On the other hand, only two questions relating to education were included in the survey, one regarding racial segregation in schools, and the other on religious education

in schools. Clearly, Eysenck's survey is too heavily influenced by his own views on what political variables are important, to be taken as an impartially-produced political spectrum.

Learning from the mistakes of Eysenck, a newly created political spectrum should use a survey whose questions are not written by the same person as that undertaking the analysis. Further, more robust techniques than Factor Analysis should be preferred for dimension reductions, since FA does not allow for an easy determination of the number of output variables to keep, and is sensitive to the number of reduced dimensions selected.

Despite its many flaws, Eysenck's political chart takes an important place in political science discourse [34, 35, 137, 171], so its axes should be discussed. Eysenck refers to the two axes of his political chart as 'Radicalism' and 'Tender-mindedness'. 'Radicalism' is essentially the traditional Left-Right divide as had evolved in the 19th century, with socialism on the left hand side of the 'Radicalism' axis and economic conservatism on the right hand side of this axis [58], with variables relating to class and economic issues being most influential over this axis. Recent adaptations of the Eysenck Chart have simply labelled this the Economic axis [34, 171]. 'Tender-mindedness' encompasses a view of how restrictive and 'tough' on its citizens a government should be with relation to social issues, with 'liberals' taking a low value on this axis, and 'authoritarians' taking a high value [58]. Recent adaptations of the Eysenck Chart have labelled this the Social axis [34], or the Libertarian/Authoritarian axis [171].

As can be observed in Figure 4.3, while different implementations of the Eysenck chart have been used over the last 60 years, all feature these same two axes of economic and social views. It will be seen in Section 4.6 that these two axes also arise from the Australian Election Study data set, in a much more rigorous analysis in an Australian context.

Recent data-driven literature

Recent papers have produced political spectra using Principal Components Analysis in, among other places, China (see Pan and Xu, 2015 [143]), Germany (see Riemann *et al.*, 1993 [151]), and Western Europe more generally (see Warwick, 2002 [176]). Many of these works use citizen survey data [143, 151], while others focus on surveying political party platforms [176]. It is encouraging to see that these works favour Principal Components Analysis over Factor Analysis, as PCA is more robust (see Chapter 3). All use sets of unreduced variables designed by the researchers themselves, for the express purpose of then being used to draw out common themes. This has the potential to lead to bias in survey design, in favour of what issues the researchers see as important. As such, it might be that secondary analyses, those undertaken after the primary survey collection has taken place, provide for more insightful results.

These recent papers also often suffer from being produced without regard to the mathematical context of the PCA procedure. For example, Riemann *et al.*'s 1993 [151] analysis of German political spectra undertakes numerous arbitrary factor
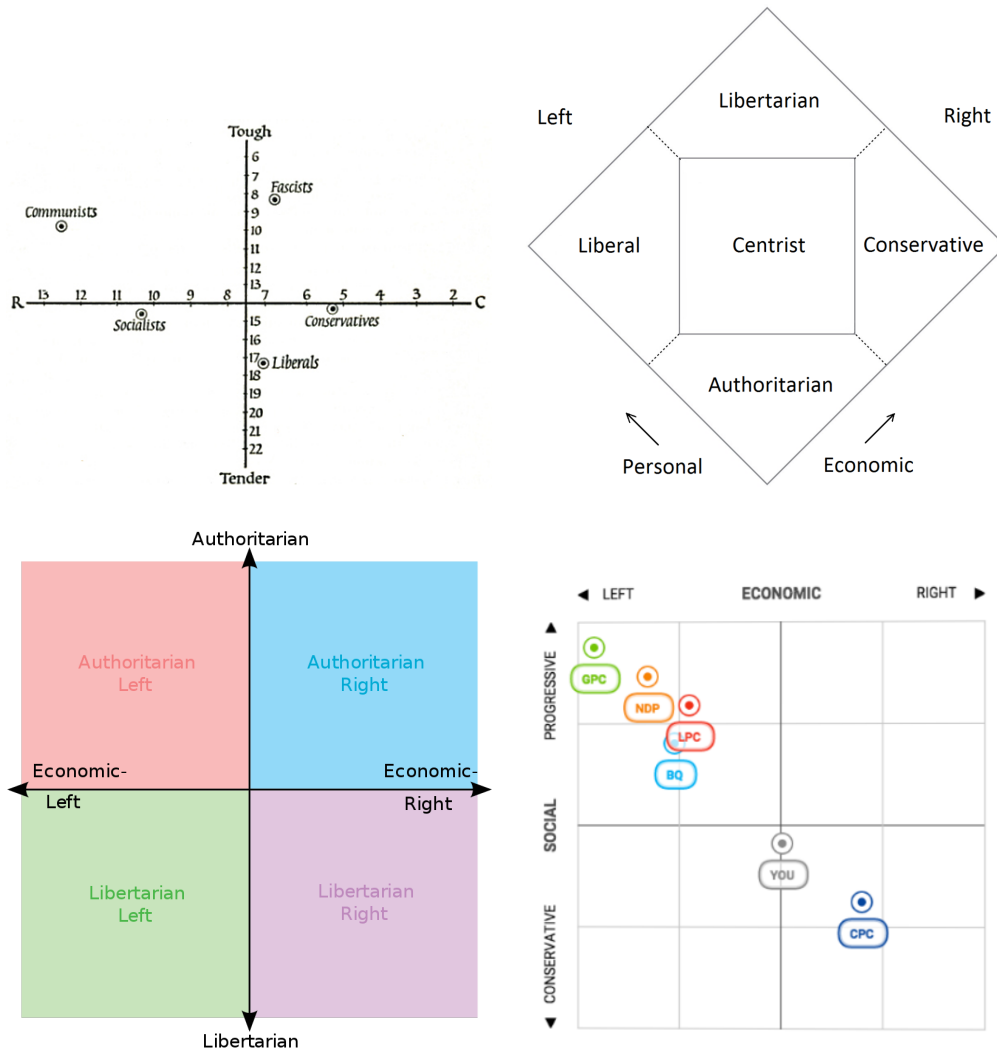
Figure 4.3: Four implementations of the Eysenck Chart. **Top-left** is from Eysenck's *Sense and Nonsense in Psychology* [58], and features 'radicalism-conservatism' as the economic axis, and 'tough-tender' as the social axis. Positions of contemporary (1957) UK political parties' platforms are placed on the axes for reference. **Top-right** is the Nolan Chart, introduced by US libertarian David Nolan in 1969 [137], and features 'liberalism-conservatism' as the economic axis, and 'authoritarian-libertarian' as the social axis. **Bottom-left** is the Political Compass produced by Pace News Limited [34], which features 'left-right' as the economic axis, and 'authoritarian-libertarian' as the social axis. **Bottom-right** is sample VoteCompass output produced by Vox Pop Labs [171], which features 'left-right' as the economic axis, and 'social conservative-progressive' as the social axis. Positions of contemporary (2015) Canadian political parties' platforms are placed on the axes for reference. The position 'you' is placed by the tool's creators as an example. Note that the social axis is flipped in both the Nolan (top-right) and VoteCompass (bottom-right) output, but this does not influence results.

rotations on the resulting principal components, selecting that which is easiest to interpret, and using this interpretation to select the number of components. By contrast, the analysis in this thesis selects reduced variables based primarily on the proportion of variation in the data they are able to explain, using ease of interpretation to illustrate, rather than justify, the decision to select a number of components.

More importantly from the perspective of mathematical rigour, all these papers use Principal Components Analysis without explaining the foundations of this procedure, nor how their survey data has been transformed to fit into the PCA framework. As explained in Section 3.1, PCA can only be applied to numerical variables. Survey questions in each work are a combination of categorical and numerical variables, with no explanation of how these have been transformed before PCA has been applied. For a political spectrum to be readily accepted as being accurate, it must be rigorous, and so this thesis builds on the useful but flawed work of Pan and Xu, Riemann *et al.* and Warwick [143, 151, 176], among others, by presenting a mathematically sound approach to the production of a political spectrum.

4.6 The AES political spectrum in context

As explained above in Section 4.4, the political spectrum derived from the Australian Election Study comprises three main axes:

1. 'Social inclusivity', which describes views as to how societies should treat marginalised groups;

2. 'Trust in authority', which describes the extent to which one is disenchanted with existing power structures; and

3. 'Spending priorities', which contrasts economic conservatism, with higher levels of government spending.

The first axis of our political spectrum

Looking at its constituent variables, the 'social inclusivity' axis constructed from the 2016 AES data is remarkably similar to Eysenck's 'tender-tough' axis [57, 58], now referred to as simply the 'social' axis [34, 171]. Both axes' primary concern is the treatment of those from varying backgrounds, or who lead varying lives. For example, the 'social inclusivity' axis is dominated by variables relating to the treatment of refugees and other immigrants, Aboriginal and Torres Strait Islander peoples, and LGBT+ people (see Section 4.4). Similarly, the Eysenck chart's 'tender-tough' axis is derived from variables including 'should European refugees should be left to fend for themselves?' and whether the dropping of the first atomic bomb on Hiroshima was morally wrong [58].

There is clearly a substantial overlap between the axes as expressed in this research and in Eysenck's work, despite the work having been undertaken over

60 years apart, and the more recent work utilising a considerably more sound methodology. Any differences between the axes seem to be a result of temporal and geographical differences; for example, questions relating to asylum seekers arriving by boat, present in the AES survey, would be out of place in 1950's England, as would questions relating to same-sex marriage or Aboriginal land rights. Eysenck's question of whether 'Jews are pretty much alike' [58] would be similarly out of place in modern political discourse.

The third axis of our political spectrum

Similarly, the 'spending priorities' axis constructed from the AES is very similar to Eysenck's 'radicalism' axis [57, 58], now referred to as the 'economic' axis [34, 171]. This axis relates to the question of whether governments should actively intervene to attempt to give its citizens equal access to health, education and a good standard of living, at the cost of higher taxation [182]. Progressive economic viewpoints at one end of this axis were perceived to be radical in the 1950s, as they required one to rethink the role of government (see, e.g., [69]); hence Eysenck's reference to this as a 'radicalism' axis [57, 58]. This label was also likely influenced by Eysenck's political tendencies, being himself conservative [35, 152, 153].

Looking at the substance of the axes, our 'spending priorities' axis contrasts views favouring high spending and distribution of wealth towards the working class, with views favouring low taxation and lower government assistance to less wealthy people (see Section 4.4). Similarly, Eysenck's 'radicalism' axis contrasts agreement with statements such as 'Capitalism is immoral because it exploits the worker' with agreement with statements such as 'Production and trade should be free from government interference'.

Our third axis and Eysenck's second are very similar. Some minor differences occur, however, as a result of Eysenck's neglect to include any survey questions relating to education policy (beyond relating to issues of race and religion in schools) [58], while favouring increased spending on education is strongly associated with a high score on the third axis of our political spectrum; favouring increased education spending has mean coordinate 1.371 on the 'spending priorities' axis. This does not reflect a dissimilarity between the underlying beliefs expressed by the compared axes, but rather, issues with Eysenck's methodology.

The second axis of our political spectrum

It can be concluded that because of the great similarity between the first axis of our dimension-reduced political space, and the first axis of Eysenck's political spectrum, and the great similarity between the third axis of our dimension-reduced political space, and the second axis of Eysenck's political spectrum, this thesis provides a rigorous, data-driven reinforcement of some of Eysenck's conclusions. Indeed, social and economic dimensions form a core of Australia's political spectrum, just as Eysenck theorised in a British context in 1954 [57].

However, Eysenck's work did not identify our second axis, regarding attitudes to authority. Those who score highly on the 'attitudes to authority' axis are likely

to distrust government, financial institutions, and democracy, while those who have a low score are likely to be less disenchanted, and trust institutions.

Though not present in Eysenck's work, the second axis of our political spectrum is grounded in a theoretical understanding of the spectrum of political identity. Miller (1974) and Citrin (1974) have previously noted that trust in authority is not captured in traditional political spectra [127, 44], and that low levels of trust are not particular to certain sets of social and economic identities. For example, Miller and Citrin both note that views cynical of government are popular both on the far-left and far-right [44], while Hibbing and Smith (2003) note that frustration in government can come from individuals in the political 'centre' [88]. Due to its importance in the current political moment, and the fact that it is not captured in current political models, Canovan (2004) suggests adding a 'trust in authority' axis to the political spectrum [38]. This proposal has been made from a social science perspective, and the research here provides statistical evidence for the existence of this axis.

Low levels of trust in authority are often manifested politically in a support for populist government [9, 37, 144]. Mudde (2004) defines populism as a political approach characterised by a distrust of "corrupt elites" ([132], cited in [25]).

Populism transcends traditional political spectra [96], so it is no surprise that it appears orthogonal to the 'social inclusivity' and 'spending priorities' axes. The emergence and recognition of this aspect of the political spectrum is particularly prescient at the current moment in political history, in which populist thought abounds. For example:

- A socially non-inclusive, predominantly economically conservative populist politician [95, 139], Donald Trump, is President of the United States;

- A socially non-inclusive, economically centrist populist coalition [61, 68, 187], consisting of Lega Nord and Five Star Movement, form Italy's government;

- A socially centrist, economically socialist populist party [162], SYRIZA, is the major party in Greece's governing coalition;

- A socially inclusive, economically socialist populist party [7, 148], MAS, is the governing party in Bolivia; and

- Actions by the previous (2005-2015) economically conservative Canadian government have been described as 'libertarian populist' [150].

Essentially, populist thought is not limited to particular ideologies with respect to social inclusivity, or to spending priorities, meaning it is not captured in current political spectra, despite it being a dimension which is becoming increasingly relevant. As such, this research's finding that attitudes to authority form a key axis in Australia's political spectrum is both novel and important.

Summary

The three axes produced by dimension reduction of the political variables in the Australian Election Study are not new from the perspective of political science. All three dimension-reduced variables reflect core aspects of political discourse. However, they have been discussed in the literature in a largely haphazard way; notions of 'left' and 'right' are used frequently without reference to whether the context is social or economic, while 'populism' is often referred to without regard to its status as an axis of identity orthogonal to economic or social views.

While data-driven political spectra have been previously produced, they have been done without due mathematical rigour, leading to descriptions of axes heavily influenced by the points of view of the researchers who produced them. These political spectra are also out of date; while often cited (see, e.g., [43, 85, 111]), Eysenck's poorly constructed model from 1954 should not form the basis of political science research.

This work thus provides four key advances on previous models of political opinion:

1. Our political spectrum is based on a recent study, from 2016, and usefully for Australian researchers, is based on Australian data;

2. Our political spectrum is mathematically rigorous, with its methodology open and unbiased;

3. Our political spectrum reinforces previous research in the area, and goes some way to validate the otherwise deeply flawed Eysenck model [57]; and

4. Our political spectrum adds a new 'attitudes to authority' axis, which is orthogonal to the other political axes, and reflects recent developments in the political science space.

This chapter presented a clearer understanding of what political ideology means in light of the 2016 AES (Section 4.4), and an understanding of demographic identity (Section 4.3). The model of political ideology should prove valuable in the realm of political science, where the addition of a third axis to the political spectrum is both strongly justified, and timely.

The next task is outlining how demographic identities intersect and are reflected in political ideology. The following chapter (Chapter 5) discusses the simplest model available for exploring the intersection of demographic and political identities, the multivariate regression model. Chapter 5 will introduce this model, before applying it to the dimension-reduced variables defined in this chapter, and determining if this model is adequate.

Chapter 5

Multivariate Regression

In Chapter 4, we produced dimension-reduced spectra of demographic and political identity. We now seek to build a model to determine how demographic identities are reflected in political ideology. In this chapter, we attempt to do so by building a multivariate regression model connecting the spaces. This is done by first recapping the theory behind the traditional multiple linear regression model (here denoted the 'univariate regression model') in Section 5.1.1. We then outline how this is extended to deal with the multivariate response present in the 2016 Australian Election Study data set, in Section 5.1.2. We explain the assumptions of, and prove key results surrounding, this model. Finally, in Section 5.2, we attempt to apply this model to the dimension-reduced spectra of demographic and political identity from Chapter 4. We apply a variety of forms of the multivariate regression model, with the goal of abiding by the model's assumptions.

We conclude in this Chapter that we cannot build a model that fully abides by the assumptions of the multivariate regression model. As such, we may need to consider alternative models to adequately describe the relationships between spectra of demographic and political identity, and to adequate describe the pitfalls of estimating political opinion using demography alone. Alternative models will be introduced in Chapter 6.

5.0   Notation in this chapter

In this chapter, and elsewhere in this work, algebraic notation follows these conventions:

- A scalar is denoted by a lower case letter; e.g. x.

- A vector is denoted by a bold lower case letter; e.g. $\boldsymbol{x}$. All vectors used here are column vectors.

- A matrix is denoted by a capital letter; e.g. X.

- A univariate random variable is denoted by a capital letter; e.g. X.

- A vector random variable is denoted by a bold capital letter; e.g. $\boldsymbol{X}$.

- A matrix random variable is denoted by a capital letter; e.g. X.

For example, suppose we have an l-dimensional vector random variable denoted $\boldsymbol{X} = [X_1, X_2, \ldots, X_l]^\mathrm{T}$. The elements of $\boldsymbol{X}$ are univariate random variables $X_1, \ldots, X_l$. A single realisation of $\boldsymbol{X}$ may be expressed as $\boldsymbol{x} = [x_1, x_2, \ldots, x_l]^\mathrm{T}$, while a set of n realisations of $\boldsymbol{X}$ may be represented by

$$
X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nl} \end{bmatrix}.
$$

Note that here, each *row* of the data matrix X corresponds to one realisation of the random variable $\boldsymbol{X}$. By contrast, a single realisation of $\boldsymbol{X}$, when considered on its own, is represented by a *column* vector. Both these conventions will be followed in this chapter, with the result that some expressions may contain a number of transpositions and thus appear inelegant. This is a natural consequence of extending notation conventions designed for use in univariate statistics to multivariate situations.

5.1   Theory of multivariate regression

5.1.1   A univariate regression model

A basic multiple linear regression model estimates a single response variable Y, based upon some set of p predictor variables, and an intercept, $\boldsymbol{X} = [1, X_1, X_2, \ldots, X_p]^\mathrm{T}$, using the model formulation

$$
Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon, \tag{5.1.1}
$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, for some scalar σ_ε^2, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p,]^\mathrm{T}$ is a vector of coefficients.

Parameter estimation from training data

Usually, $\boldsymbol{\beta}$ and σ_ε^2 are estimated from some set of observed data in the form

$$
\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}.
$$

Each element of $\boldsymbol{y}$ corresponds to a single realisation of Y, and each row of X corresponds to a single realisation of the predictor variables $\boldsymbol{X}$, with the first column of X representing the intercept term.

The model is produced with the goal of best estimating response variable Y, given predictor variables $\boldsymbol{X}$, so correspondingly $\boldsymbol{\beta}$ is chosen to minimise the error in the observed data. The best estimate of $\boldsymbol{\beta}$ is thus that which minimises the sum of squared errors

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) \right)^2. \tag{5.1.2}$$

Note the correspondence between the model in Equation (5.1.1) and the objective function in Equation (5.1.2).

In matrix form, this expression can be rewritten as

$$RSS(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2.$$

It can be easily shown that $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} RSS(\boldsymbol{\beta}) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y}$.

Further, σ_ε^2 can be estimated from the training data. The most common unbiased estimator is

$$s_e^2 = \frac{1}{n - (p+1)} RSS(\hat{\boldsymbol{\beta}}). \tag{5.1.3}$$

Assumptions of univariate regression

A coherent univariate regression requires four distributional assumptions of ε. These are

1. Linearity—that the expected error for some new observation is independent of that observation: $E[\varepsilon|\boldsymbol{X}] = E[\varepsilon] = 0$;

2. Homoscedasticity—that the variance of the error for some new observation is independent of that observation: $\operatorname{Var}(\varepsilon|\boldsymbol{X}) = \operatorname{Var}(\varepsilon) = \sigma_\varepsilon^2$;

3. Normality—that the error is normally distributed: $\varepsilon \sim N(0, \sigma_\varepsilon^2)$; and

4. Independence—that all errors in the training data are independent of one another: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2), i = 1, 2, \ldots, n$.

Linearity is necessary to ensure that the estimator $\hat{Y} = \boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is unbiased at all values of the predictor set $\boldsymbol{X}$. The model is always unbiased *on average* since $\hat{\boldsymbol{\beta}}$ is constructed to be unbiased for $\boldsymbol{\beta}$. However, without the assumption of linearity, it may be that the regression model is an inadequate fit for some or most values of $\boldsymbol{X}$.

Homoscedasticity is necessary to ensure error bounds can be accurately calculated. If spread is wider at one point than another, then error will be overestimated for some values of $\boldsymbol{X}$, and underestimated at others.

The errors must also be normally distributed. This ensures prediction and confidence intervals can be accurately calculated for all bounds.

Independence is necessary for s_e^2 (Equation (5.1.3)) to be unbiased for σ_ε^2.

Application of univariate regression to multivariate data

Where some multivariate response is to be measured, the regression techniques developed for univariate analysis allow one to build estimators for each response variable separately. From these estimates of marginal response distributions, some joint distribution of errors might be constructed. Since there are infinite constructions of joint distributions from any pair of marginals, some assumptions are necessary to construct a particular multivariate model from some set of univariate regressions. The simplest such construction assumes independence of residuals.

Where residuals are independent and normally distributed, they form a joint multivariate normal distribution with a diagonal covariance matrix. For example, suppose there are two response variables, Y_1 and Y_2, which are modelled by separate univariate regressions. If it is assumed that Y_1 and Y_2 are independent, and noting that $\boldsymbol{X}$ includes the intercept, the joint model can be expressed as

$$\boldsymbol{Y}^{\mathrm{T}} = \begin{bmatrix} Y_1 & Y_2 \end{bmatrix} = \boldsymbol{X}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{\beta_1} & \boldsymbol{\beta_2} \end{bmatrix} + \boldsymbol{\varepsilon}^{\mathrm{T}},$$

where $\boldsymbol{\varepsilon} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\varepsilon_1}^2 & 0 \\ 0 & \sigma_{\varepsilon_2}^2 \end{bmatrix} \right)$.

Using this model, multivariate estimates can be made. Conditional expectations of the response variables, those at which error takes its expected value of zero, are given by

$$\hat{\boldsymbol{y}_0}^{\mathrm{T}} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 \end{bmatrix} = \boldsymbol{x_0}^{\mathrm{T}} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 & \hat{\boldsymbol{\beta}}_2 \end{bmatrix}, \tag{5.1.4}$$

where $\boldsymbol{x_0}$ is a vector of new observations of $\boldsymbol{X}$. In other words, the univariate estimates are concatenated into a vector of multivariate estimates.

Prediction regions are obtained by taking hyperellipsoidal contours of the density of the distribution around $\hat{\boldsymbol{y}_0}$, corresponding to the desired level of confidence. Since no off-diagonal term exists in the assumed covariance matrix, the ellipsoids are restricted to those obtainable by directional scaling of a unit sphere. A derivation of prediction regions for the more involved multivariate regression model case is provided in Theorem 5.1.3.

5.1.2   A multivariate regression model

The univariate regression model is severely limited in multivariate analysis, as the assumption of independence of residuals is rarely justified. For example, consider a model of the health outcomes of a patient suffering a disease, with response variables representing the duration of the condition, and its severity. If a patient suffers the disease longer than expected, it may be that she is more likely to experience more severe symptoms. In this example, there is a positive relationship between the residuals in the 'severity' and 'duration' directions, so the assumption of independence of residuals does not hold. In this circumstance, some model encompassing the potential for correlation between residuals may prove valuable.

The multivariate regression model is motivated by this desire to describe correlated residual structures, while being relatively parsimonious. For some set of

k response variables $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_k]^{\mathrm{T}}$, modelled on some set of p predictor variables $\boldsymbol{X} = [1, X_1, X_2, \ldots, X_p]^{\mathrm{T}}$, with the first element of $\boldsymbol{X}$ corresponding to an intercept term, the model is formulated as

$$\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}}\beta + \boldsymbol{\varepsilon}^{\mathrm{T}}, \tag{5.1.5}$$

where β is some $(p+1) \times k$ matrix of coefficients, $\boldsymbol{\varepsilon} \sim N_k(\boldsymbol{0}, \Sigma_{\boldsymbol{\varepsilon}})$, and $\Sigma_{\boldsymbol{\varepsilon}}$ is a symmetric $k \times k$ matrix.

Parameter estimation from training data

In practice, β and $\Sigma_{\boldsymbol{\varepsilon}}$ are estimated from data of the form

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nk} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

where each row of Y and X corresponds to some observation of variables $\boldsymbol{Y}$ and $\boldsymbol{X}$ respectively, with the first column of X representing the intercept term.

The sum of squared errors is sought to be simultaneously minimised in all $\boldsymbol{Y}$-directions. As such, we seek to minimise the objective function

$$RSS(\beta) = \sum_{l=1}^{k} \sum_{i=1}^{n} \left(y_{il} - \left(\beta_{0l} + \sum_{j=1}^{p} x_{ij} \beta_{jl} \right) \right)^2,$$

where β_{jl} is the $(j+1, l)$th entry of β.

Note that the residual sum of squres here is equivalent to that from the univariate case (Equation (5.1.2)), summed over all response variables.

In matrix form, this can be rewritten as

$$RSS(\beta) = \mathrm{tr}\left((Y - X\beta)^{\mathrm{T}} (Y - X\beta) \right).$$

Theorem 5.1.1. *The residual sum of squares* $RSS(\beta)$ *is uniquely minimised at* $\hat{\beta} := (X^T X)^{-1} X^T Y$.

Proof. Consider $\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Y$.

$$\mathrm{tr}\left((Y - X\beta)^{\mathrm{T}} (Y - X\beta) \right)$$

$$= \mathrm{tr}\left((Y - X\hat{\beta} + X\hat{\beta} - X\beta)^{\mathrm{T}} (Y - X\hat{\beta} + X\hat{\beta} - X\beta) \right)$$

$$= \mathrm{tr}\left((Y - X\hat{\beta})^{\mathrm{T}} (Y - X\hat{\beta}) \right) + \mathrm{tr}\left((X\hat{\beta} - X\beta)^{\mathrm{T}} (X\hat{\beta} - X\beta) \right)$$

$$+ \operatorname{tr}\left((Y - X\hat{\beta})^{\mathrm{T}}(X\hat{\beta} - X\beta)\right) + \operatorname{tr}\left((X\hat{\beta} - X\beta)^{\mathrm{T}}(Y - X\hat{\beta})\right)$$

$$= \operatorname{tr}\left((Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})\right) + \operatorname{tr}\left((X\hat{\beta} - X\beta)^{\mathrm{T}}(X\hat{\beta} - X\beta)\right)$$

$$+ 2\operatorname{tr}\left((Y - X\hat{\beta})^{\mathrm{T}}(X\hat{\beta} - X\beta)\right), \quad \text{since } \operatorname{tr}(A^{\mathrm{T}}) = \operatorname{tr}(A).$$

We can then expand the cross-term:

$$\operatorname{tr}\left((Y - X\hat{\beta})^{\mathrm{T}}(X\hat{\beta} - X\beta)\right)$$

$$= \operatorname{tr}\left(Y^{\mathrm{T}}(X\hat{\beta} - X\beta) - (X\hat{\beta})^{\mathrm{T}}(X\hat{\beta} - X\beta)\right)$$

$$= \operatorname{tr}\left(Y^{\mathrm{T}}X(\hat{\beta} - \beta) - (X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y)^{\mathrm{T}}X(\hat{\beta} - \beta)\right)$$

$$= \operatorname{tr}\left(Y^{\mathrm{T}}X(\hat{\beta} - \beta) - Y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(\hat{\beta} - \beta)\right)$$

$$= \operatorname{tr}\left(Y^{\mathrm{T}}X(\hat{\beta} - \beta) - Y^{\mathrm{T}}X(\hat{\beta} - \beta)\right)$$

$$= 0.$$

We thus obtain

$$RSS(\beta) = \operatorname{tr}\left((Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})\right) + \operatorname{tr}\left((X\hat{\beta} - X\beta)^{\mathrm{T}}(X\hat{\beta} - X\beta)\right).$$

$$\geq \operatorname{tr}\left((Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})\right),$$

with equality only at $\beta = \hat{\beta}$, meaning $\hat{\beta}$ uniquely minimises $RSS(\beta)$. $\qquad\square$

Multivariate regression thus gives conditional expectations of the response variables given by $\hat{\boldsymbol{y}_0} = \boldsymbol{x_0}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$. Recall from Section 5.1.1 that in univariate regression, we have conditional expectations $\hat{y}_j = \boldsymbol{x_0}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y_j^\star}$, for responses $j = 1, \ldots, k$, and $\boldsymbol{y_j^\star}$ denoting the jth *column* of Y. Multivariate regression thus gives the same conditional expectations $\hat{\boldsymbol{Y}}$ as concatenating those obtained by univariate regression.

The key difference between the univariate and multivariate approaches then comes in the estimation of Σ_ε. An unbiased estimate of the residual covariance matrix is given by

$$\hat{\Sigma}_{\boldsymbol{\varepsilon}} = \frac{1}{n - (p+1)} \sum_{i=1}^{n} \left(\boldsymbol{y_i^{\mathrm{T}}} - \boldsymbol{x_i^{\mathrm{T}}}\hat{\beta}\right)\left(\boldsymbol{y_i^{\mathrm{T}}} - \boldsymbol{x_i^{\mathrm{T}}}\hat{\beta}\right)^{\mathrm{T}}, \qquad (5.1.6)$$

where $\boldsymbol{y_i}$ is the ith observation of $\boldsymbol{Y}$ and $\boldsymbol{x_i^{\mathrm{T}}}$ is the corresponding row of X. That this estimator is unbiased is shown below.

Theorem 5.1.2. $\hat{\Sigma}_{\boldsymbol{\varepsilon}}$ *is unbiased for* $\Sigma_{\boldsymbol{\varepsilon}}$.

Proof. First define $\boldsymbol{y_i^\star}$ to be the ith column of Y, $\boldsymbol{\hat{\beta}_i}$ to be the ith column of $\hat{\beta}$, and $\boldsymbol{\beta_i}$ to be the ith column of β. Note that for $[\hat{\Sigma}_\varepsilon]_{ij}$ the (i,j)th element of $\hat{\Sigma}_\varepsilon$,

$$[\hat{\Sigma}_\varepsilon]_{ij} = \frac{1}{n-(p+1)}(\boldsymbol{y_i^\star} - X\boldsymbol{\hat{\beta}_i})^{\mathrm{T}}(\boldsymbol{y_j^\star} - X\boldsymbol{\hat{\beta}_j}).$$

Defining $P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, we can write

$$\begin{aligned}
(n-(p+1))[\hat{\Sigma}_\varepsilon]_{ij} &= (\boldsymbol{y_i^\star} - X\boldsymbol{\hat{\beta}_i})^{\mathrm{T}}(\boldsymbol{y_j^\star} - X\boldsymbol{\hat{\beta}_j}) \\
&= (\boldsymbol{y_i^\star} - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y_i^\star})^{\mathrm{T}}(\boldsymbol{y_j^\star} - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y_j^\star}) \\
&= ((I-P)\boldsymbol{y_i^\star})^{\mathrm{T}}((I-P)\boldsymbol{y_j^\star}).
\end{aligned}$$

Observing that $PX\boldsymbol{\beta_i} = X\boldsymbol{\beta_i}$, we obtain $(I-P)X\boldsymbol{\beta_i} = 0$, and so

$$\begin{aligned}
(n-(p+1))[\hat{\Sigma}_\varepsilon]_{ij} &= ((I-P)(\boldsymbol{y_i^\star} - X\boldsymbol{\beta_i}))^{\mathrm{T}}((I-P)(\boldsymbol{y_j^\star} - X\boldsymbol{\beta_j})) \\
&= (\boldsymbol{y_i^\star} - X\boldsymbol{\beta_i})^{\mathrm{T}}(I-P)^{\mathrm{T}}(I-P)(\boldsymbol{y_j^\star} - X\boldsymbol{\beta_j}).
\end{aligned}$$

Noting that P is both symmetric and idempotent, it is observed that $I-P$ is symmetric and idempotent, and so

$$(n-(p+1))[\hat{\Sigma}_\varepsilon]_{ij} = (\boldsymbol{y_i^\star} - X\boldsymbol{\beta_i})^{\mathrm{T}}(I-P)(\boldsymbol{y_j^\star} - X\boldsymbol{\beta_j}).$$

Since each element of $\hat{\Sigma}_\varepsilon$ is a scalar, we can apply the properties of the trace operator:

$$\begin{aligned}
(n-(p+1))[\hat{\Sigma}_\varepsilon]_{ij} &= \mathrm{tr}((\boldsymbol{y_i^\star} - X\boldsymbol{\beta_i})^{\mathrm{T}}(I-P)(\boldsymbol{y_j^\star} - X\boldsymbol{\beta_j})) \\
&= \mathrm{tr}((I-P)(\boldsymbol{y_j^\star} - X\boldsymbol{\beta_j})(\boldsymbol{y_i^\star} - X\boldsymbol{\beta_i})^{\mathrm{T}}).
\end{aligned}$$

Taking expectation, and noting that observations are independent of each other, and Σ_ε is independent of X,

$$\begin{aligned}
(n-(p+1))E[[\hat{\Sigma}_\varepsilon]_{ij}] &= \mathrm{tr}\left((I-P)E\left[(\boldsymbol{y_j^\star} - X\boldsymbol{\beta_j})(\boldsymbol{y_i^\star} - X\boldsymbol{\beta_i})^{\mathrm{T}}\right]\right) \\
&= \mathrm{tr}((I-P)\,\mathrm{Cov}(\varepsilon_i,\varepsilon_j)I) \\
&= \mathrm{tr}((I-P)[\Sigma_\varepsilon]_{ij}I) \\
&= [\Sigma_\varepsilon]_{ij}\,\mathrm{tr}(I-P).
\end{aligned}$$

Now note that $\mathrm{tr}(I-P) = n-(p+1)$:

$$\begin{aligned}
\mathrm{tr}(I-P) &= (\mathrm{tr}(I_{n\times n}) - \mathrm{tr}(P)) \\
&= n - \mathrm{tr}(X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}) \\
&= n - \mathrm{tr}((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X) \\
&= n - \mathrm{tr}(I_{p+1\times p+1}) \\
&= n - (p+1). \qquad\qquad\qquad (5.1.7)
\end{aligned}$$

Finally, we can thus write

$$\begin{aligned}
(n-(p+1))E[[\hat{\Sigma}_\varepsilon]_{ij}] &= (n-(p+1))[\Sigma_\varepsilon]_{ij} \\
\Rightarrow E[[\hat{\Sigma}_\varepsilon]_{ij}] &= [\Sigma_\varepsilon]_{ij}.
\end{aligned}$$

$\square$

The above argument shows that $\hat{\Sigma}_{\varepsilon}$ is unbiased for Σ_{ε}. $\hat{\Sigma}_{\varepsilon}$ is not the maximum likelihood estimate for Σ_{ε}, though; an argument that the MLE for Σ_{ε} is $\frac{(n-p-1)}{n}\hat{\Sigma}_{\varepsilon}$ can be found in Anderson (1958) [5].

Since the estimated variance-covariance matrix $\hat{\Sigma}_{\varepsilon}$ is unbiased for Σ_{ε}, and not just its diagonal terms, the multivariate regression model is able to account for covariances between residuals. This is the key difference between the univariate and multivariate regression models; univariate regression does not directly account for linear relationships in the residual structure. As seen in the above proof, the ability to account for *linear* relationships between the residuals relied on the assumption that the residuals were distributed independently of each other and of X. This is reflected in the assumptions of the multivariate model, and is something that will be of great significance in this thesis; the multivariate model is only adequate when relationships between the residuals are linear.

Assumptions of multivariate regression

The central assumption of multivariate regression is that the residuals ε are normally distributed with mean $\mathbf{0}$ and variance matrix Σ_{ε}, independently of the predictors $\boldsymbol{X}$. This is usually broken down into the following requirements, which together form a necessary but not sufficient condition for ε's distributional form:

1. Linearity—That the expectation of ε is independent of each element of the predictor variables $\boldsymbol{X}$, marginally.

2. Homoscedasticity—That the variance of ε is independent of each element of $\boldsymbol{X}$, marginally.

3. Normality—That ε has a multivariate normal distribution.

Linearity and homoscedasticity are necessary to ensure estimates are unbiased, and have appropriate error measures defined. Normality is necessary to ensure that the shape of prediction regions is representative of the shape taken by error in practice. It is important at this point to note that marginal normality does not imply multivariate normality, and that in fact, almost all sets of marginally normal distributions do not have a joint normal distribution. As such, tests for multivariate normality must be more extensive than the cursory glance at a quantile-quantile plot often undertaken in univariate analyses.

Independence of observations is also assumed, as this is necessary to ensure $\hat{\Sigma}_{\varepsilon}$ is unbiased for Σ_{ε} (see the proof of Theorem 5.1.2).

Assessing multivariate normality

One way of testing for multivariate normality is to assess whether the distribution of the residuals shares important features of the multivariate normal distribution. For many analyses, this is sufficient; so long as the residuals are distributed with similar behaviour to a normal distribution, it may not matter if the distribution is genuinely normal. Such feature-focused approaches include:

- Quantile-quantile plot assessment of marginal normality.

- Tests for non-linear relationships between the residuals.

It is clear that these features flow necessarily from the multivariate normal distribution, but are not sufficient for multivariate normality. Marginal normality does not imply multivariate normality. Similarly, while the multivariate normal distribution does not contain non-linear relationships between its marginals, being characterised entirely by its mean and covariance matrix, not all distributions containing no non-linear relationships between its marginals is normal (for example, the multivariate t-distribution).

In this section, we discuss the feature-focused approach of testing for non-linear relationships between residuals, as well as a more holistic test for normality, the energy test, introduced by Skèkely and Rizzo [166]. In doing so, we review existing literature on tests for multivariate normality. We find that both the feature-focused approaches and the energy test have a place in assumption-checking for multivariate regression.

Testing for non-linear relationships between residuals A simple test for non-linear relationships between residuals is to fit a multiple regression of marginal residuals upon each other, with non-linear basis expansions; if the data is multivariate normal, there should be no significant non-linear relationship. For example, if a quadratic term is significant in a regression of residuals in one direction upon residuals in another direction, then there is evidence that there is a non-linear relationship between the residuals and hence the data cannot be multivariate normal.

An issue with using this as an indicator of normality is the complexity of these regressions. Firstly, the choice of non-linear term to test is somewhat ambiguous; not all non-linear relationships are polynomial, so there are a large number of functional forms to consider. This means that where multiple tests for non-linearity are undertaken, p-value adjustments need to be made to account for the associated increased Type I error. For this reason, testing for non-linear relationships should only be undertaken where the shape of potential non-linearities is known in advance.

Another issue with this approach is that complexity increases exponentially as more response variables are added in the multivariate regression model, since comparisons between residuals are made pairwise. Still, for a small number of response variables, and where the shape of potential non-linearities is known, non-linear regression of residuals provides a simple way of assessing a key feature of the assumption of multivariate normality.

The energy test for multivariate normality A more holistic test for multivariate normality is the energy test, described by Szèkely and Rizzo [166]. The test relies upon the idea that the structure of Euclidean distances between independently-drawn deviates from a given distribution is unique to that distribution. In other words, if two distributions X and Y are different, the expected distance between one point from X and one point from Y should be greater than the mean of: (1)

the expected distance between two points from $\boldsymbol{X}$; and (2) the expected distance between two points from $\boldsymbol{Y}$.

To this end, Szèkely and Rizzo demonstrate [166] that for $\boldsymbol{X}$ and $\boldsymbol{X}'$ independent and identically distributed, independently of $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ i.i.d.,

$$2E\left\|\boldsymbol{X} - \boldsymbol{Y}\right\| - E\left\|\boldsymbol{Y} - \boldsymbol{Y}'\right\| - E\left\|\boldsymbol{X} - \boldsymbol{X}'\right\| \geq 0, \qquad (5.1.8)$$

with equality only when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed.

For observed deviates from $\boldsymbol{X}$, and known distribution $\boldsymbol{Y}$, one can estimate the inequality in Equation (5.1.8); if the equality is substantially greater than zero, it is unlikely that $\boldsymbol{X}$ is distributed according to $\boldsymbol{Y}$.

In the case of testing for multivariate normality, the observed data are first transformed so as to have mean $\boldsymbol{0}$ and variance I, yielding the observations from $\boldsymbol{X}$, denoted $\{\boldsymbol{x_i} : i = 1, \ldots, n\}$. The energy statistic for multivariate normality is then defined to be

$$\mathscr{E} = n\left(\frac{2}{n}\sum_{i=1}^{n} E\left\|\boldsymbol{x_i} - \boldsymbol{Z}\right\| - E\left\|\boldsymbol{Z} - \boldsymbol{Z}'\right\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\boldsymbol{x_i} - \boldsymbol{x_j}\right\|\right), \qquad (5.1.9)$$

where $\boldsymbol{Z}, \boldsymbol{Z}' \overset{iid}{\sim} N(\boldsymbol{0}, I)$ (Equation 6 in Skèkely and Rizzo [166]; the article also gives formulas for calculating the expected distances). P-values arising from the energy statistic are estimated using simulation under the null hypothesis $\boldsymbol{X} \sim N(\boldsymbol{0}, I)$.

Considering the form of Equation (5.1.9), one might draw analogy between the energy statistic and an ANOVA F-statistic. In an ANOVA, one can test whether there is a difference in means between two sampled groups by comparing the differences between observations within the same group, and the differences between observations in different groups. Similarly, the energy test statistic tests for a difference between distributions by comparing the mean distance between observations from the same distribution, and the mean distance between observations in different distributions. Both statistics cannot be less than zero, and are large when a difference between groups exists.

Key advantages of the energy test are its affine invariance and consistency [166]. In other words, the energy statistic for some observed data is identical for some affine transformation of that data, and for $n \to \infty$, normality of non-normal samples will be rejected by the test almost surely [166]. This contrasts with alternative tests for multivariate normality, which may not be affine invariant (see, e.g., the Kankainen–Taskinen–Oja kurtosis test, as per Joenssen and Vogel [99]), or may not be consistent (see, eg., Mardia's skewness and kurtosis tests, as per Mecklin and Mundfrom [126]).

Notably, the energy test also performs well against alternative methods. As opposed to feature-focused distributional tests, which consider specific properties such as skewness and kurtosis to determine whether the shape of a distribution is similar to that desired, the energy test makes a direct comparison between the alignment of points in two distributions, meaning it is able to identify any property of a distribution that renders the assumption of normality invalid. For this reason, the energy test is referred to as an 'omnibus test' [166].

The energy test also performs better than other omnibus tests in terms of power. In 2012, Joenssson and Vogel performed an assessment of the performance of nine tests for multivariate normality implemented in R [99], including Mardia's skewness and kurtosis tests; multivariate generalisations of the Shapiro-Wilk and Kolmogorov-Smirnov tests; and the energy test, among others. For deviates from almost all non-normal distributions considered, the energy test was more powerful than all others in rejecting normality, and the test did not perform poorly in any circumstances [99].

Prediction regions under the multivariate regression model

Conditional expectations of the response variables for the multivariate model are given by
$$\hat{\boldsymbol{Y}}^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}}\hat{\beta}.$$

Given the construction of $\hat{\beta}$, these are identical to those produced by separate univariate regressions for each response (Equation (5.1.4)).

Prediction regions are obtained by taking contours of the density of the distribution around $\hat{\boldsymbol{Y}}$ given by $\boldsymbol{\varepsilon}$. Since these contours are given by a multivariate distribution with marginals dependent on each other in a linear way, these can take any hyperellipsoidal shape; that is, they can take any shape obtainable by affine transformation of the unit hypersphere. This is more general than the prediction region shape obtainable from concatenating univariate regressions and assuming independence, which is limited to hyperellipsoids obtainable by scaling of the unit sphere.

Theorem 5.1.3. *Prediction regions arising from the multivariate regression model will be hyperellipsoids.*

Proof. Under the multivariate regression model, each row of Y is distributed independently in the form
$$\boldsymbol{y_i}^{\mathrm{T}} \overset{iid}{\sim} N_k(\boldsymbol{x_i}^{\mathrm{T}}\beta, \Sigma).$$

Stacking these rows, and using their independence, we have

$$Y \sim N_{n\times k}(X\beta, I_{n\times n}, \Sigma),$$

where $N_{m\times n}(\mu, A, B)$ denotes the matrix normal distribution with mean matrix μ, with $m \times m$ variance matrix A among the rows, and $n \times n$ variance matrix B among the columns.

A property of the matrix normal distribution is that for $a \times m$ matrix C, and $W \sim N_{m\times n}(\mu, A, B)$,
$$CW \sim N_{a\times n}(C\mu, CAC^{\mathrm{T}}, B). \tag{5.1.10}$$

Applying this for $C = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ and $W = Y$, and noting that $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$, we obtain

$$\hat{\beta} \sim N_{p\times k}((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\beta, (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}I((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}}, \Sigma)$$

$$\Rightarrow \hat{\beta} \sim N_{p \times k}(\beta, (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}, \Sigma)$$
$$\Rightarrow \hat{\beta} \sim N_{p \times k}(\beta, (X^{\mathrm{T}}X)^{-1}, \Sigma).$$

For new observation $\boldsymbol{x_0}$ and associated response $\boldsymbol{y_0}$, we estimate $\boldsymbol{y_0^{\mathrm{T}}}$ by $\boldsymbol{x_0^{\mathrm{T}}}\hat{\beta}$. Using the property in Equation (5.1.10),

$$\boldsymbol{x_0^{\mathrm{T}}}\hat{\beta} \sim N_{1 \times k}(\boldsymbol{x_0^{\mathrm{T}}}\beta, \boldsymbol{x_0^{\mathrm{T}}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0}, \Sigma).$$

Since this is just the expression of a vector normal distribution,

$$\boldsymbol{x_0^{\mathrm{T}}}\hat{\beta} \sim N_k(\boldsymbol{x_0^{\mathrm{T}}}\beta, \boldsymbol{x_0^{\mathrm{T}}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0}\Sigma).$$

Combining this with the fact that $\boldsymbol{y_0^{\mathrm{T}}} \sim N_k(\boldsymbol{x_0^{\mathrm{T}}}\beta, \Sigma)$ independently of $\hat{\beta}$, we obtain

$$\boldsymbol{y_0^{\mathrm{T}}} - \boldsymbol{x_0^{\mathrm{T}}}\hat{\beta} \sim N_k(0, (1 + \boldsymbol{x_0^{\mathrm{T}}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0})\Sigma),$$

and thus that

$$\frac{\boldsymbol{y_0^{\mathrm{T}}} - \boldsymbol{x_0^{\mathrm{T}}}\hat{\beta}}{\sqrt{1 + \boldsymbol{x_0^{\mathrm{T}}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0}}} \sim N_k(0, \Sigma). \tag{5.1.11}$$

We now seek the distribution of $\hat{\Sigma}_{\boldsymbol{\varepsilon}}$. We rearrange the definition of $\hat{\Sigma}_{\boldsymbol{\varepsilon}}$ given in Equation (5.1.6) to obtain

$$\begin{aligned}
(n - (p+1))\hat{\Sigma}_{\boldsymbol{\varepsilon}} &= \sum_{i=1}^{n} \left(\boldsymbol{y_i^{\mathrm{T}}} - \boldsymbol{x_i^{\mathrm{T}}}\hat{\beta}\right)\left(\boldsymbol{y_i^{\mathrm{T}}} - \boldsymbol{x_i^{\mathrm{T}}}\hat{\beta}\right)^{\mathrm{T}} \\
&= (Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta}) \\
&= ((I - P)Y)^{\mathrm{T}}((I - P)Y),
\end{aligned}$$

for $P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$. Noting, as in the proof of Theorem 5.1.2, that $(I-P)X\beta = 0$,

$$\begin{aligned}
(n - (p+1))\hat{\Sigma}_{\boldsymbol{\varepsilon}} &= ((I - P)(Y - X\beta))^{\mathrm{T}}(I - P)(Y - X\beta) \\
&= (Y - X\beta)^{\mathrm{T}}(I - P)^{\mathrm{T}}(I - P)(Y - X\beta),
\end{aligned}$$

and since P and hence I-P are symmetric and idempotent,

$$(n - (p+1))\hat{\Sigma}_{\boldsymbol{\varepsilon}} = (Y - X\beta)^{\mathrm{T}}(I - P)(Y - X\beta). \tag{5.1.12}$$

Theorem 3.4.4 in Mardia *et al.* (1979) [117] states that for some matrix of observations $Z_{n \times p}$ with rows $\boldsymbol{z_i} \overset{iid}{\sim} N_p(0, \Sigma)$, and $C_{n \times n}$ is a symmetric and idempotent matrix with trace r,

$$Z^{\mathrm{T}}CZ \sim \mathrm{W}_p(\Sigma, r),$$

where $W_p(\Sigma, r)$ denotes the $p \times p$ Wishart distribution with scale matrix Σ and degrees of freedom r. The Wishart distribution is a matrix extension to the chi-squared distribution, whose properties are discussed in more detail in Mardia *et al.* [117]; here only its relevant properties are used.

Here we have that $I - P$ is symmetric and idempotent, with trace $n - (p + 1)$ (see Equation (5.1.7)). Under the model assumptions, $\boldsymbol{y_i}^{\mathrm{T}} \stackrel{iid}{\sim} N_k(\boldsymbol{x_i}^{\mathrm{T}}\beta, \Sigma)$ and hence $\boldsymbol{y_i}^{\mathrm{T}} - \boldsymbol{x_i}^{\mathrm{T}}\beta \stackrel{iid}{\sim} N_k(0, \Sigma)$. Applying Theorem 3.4.4 from Mardia *et al.* [117], we can obtain

$$(Y - X\beta)^{\mathrm{T}}(I - P)(Y - X\beta) \sim \mathrm{W}_k(\Sigma, n - (p + 1))$$
$$\Rightarrow (n - (p + 1))\hat{\Sigma}_{\boldsymbol{\varepsilon}} \sim \mathrm{W}_k(\Sigma, n - (p + 1)). \tag{5.1.13}$$

Theorem 17.11 in Arnold (1981) [11] provides that for $Z \sim N_k(0, \Sigma)$, and $W \sim W_k(\Sigma, r)$, with $r \geq k$, Σ non-zero and Z independent of W,

$$\frac{r - k + 1}{k}Z^{\mathrm{T}}W^{-1}Z \sim F_{k,r-k+1}.$$

Applying this to $W = ((n - (p + 1))\hat{\Sigma}_{\boldsymbol{\varepsilon}})^{-1}$ and

$$Z^{\mathrm{T}} = \frac{\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta}}{\sqrt{1 + \boldsymbol{x_0}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0}}},$$

we obtain

$$\frac{n - p - k}{k(n - p - 1)} \frac{(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}\hat{\beta})\hat{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})^{\mathrm{T}}}{1 + \boldsymbol{x_0}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0}} \sim F_{k,n-p-k}. \tag{5.1.14}$$

The statistic

$$\frac{(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})\hat{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})^{\mathrm{T}}}{1 + \boldsymbol{x_0}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0}}$$

is often called Hotelling's T^2 statistic (see, e.g., Hooper and Zellner (1961) [91]).

From Equation (5.1.14), to obtain a prediction interval for some significance level α, we can now simply take all values of $\boldsymbol{y_0}$ such that

$$\frac{n - p - k}{k(n - p - 1)} \frac{(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})\hat{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})^{\mathrm{T}}}{1 + \boldsymbol{x_0}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0}} \leq F_{k,n-p-k}(\alpha)$$

for $F_{k,n-p-k}(\alpha)$ the $(1 - \alpha)$th quantile of the F distribution on k and $n - p - k$ degrees of freedom. This can be rewritten as

$$(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})\hat{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})^{\mathrm{T}} \leq \frac{k(n - p - 1)}{n - p - k}(1 + \boldsymbol{x_0}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x_0})F_{k,n-p-k}(\alpha).$$

Since the right hand side of this equation is a constant value at any given value of $\boldsymbol{x_0}$, and is independent of $\boldsymbol{y_0}$, we can denote it by a constant d_0. The confidence region is thus given by

$$\{\boldsymbol{y_0} : (\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})\hat{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y_0}^{\mathrm{T}} - \boldsymbol{x_0}^{\mathrm{T}}\hat{\beta})^{\mathrm{T}} \leq d_0\}; \tag{5.1.15}$$

in other words, the set of all points $\boldsymbol{y_0}$ whose estimated Mahalanobis distance from $\boldsymbol{x_0}^{\mathrm{T}}\hat{\beta}$ is less than d_0.

From Equation (5.1.15), it is clear that the shape of the prediction region arising from the multivariate regression model is a hyperellipsoid centred at $\boldsymbol{x_0}^{\mathrm{T}}\hat{\beta}$ (For example, see the definition of an ellipsoid given on page 483 of Anderson [5]). $\quad\square$

5.2   Application of multivariate regression

In Chapter 4, dimension reduction techniques called Multiple Correspondence Analysis and Factorial Analysis of Mixed Data were applied, respectively, to demographic and political variables measured by the 2016 Australian Election Study. These techniques produced a set of dimension-reduced demographic variables, and a set of dimension-reduced political variables. As discussed in detail in Chapter 4, the first two dimension-reduced political variables characterise, among other issues, levels of social inclusivity and trust in authority respectively. We now seek a model to, in the first instance, relate these two political variables to some subset of the dimension-reduced demographic variables.

The multivariate regression model is chosen due to its increased flexibility as compared with undertaking separate univariate regressions; as seen in Chapter 5.1, the multivariate model does not assume independence of residuals. In order for the multivariate model to be appropriate, multivariate normality of residuals must be assumed, with residuals distributed independently of the predictors and of the fitted values (Chapter 5.1.2). Indications of the reasonableness of this assumption are obtained by checking:

- linearity of residuals in each direction against each predictor, and against fitted values;

- homoscedascity of residuals in each direction against each predictor, and against fitted values; and

- multivariate normality of the residuals.

Models will be fit with an aim of satisfying these assumptions. Where a model does not meet the assumptions of multivariate regression, it should not be used, since the model will produce inaccurate confidence and prediction regions.

Whether the assumptions can be justified will also be discussed with reference to the relevant political science literature in Section 5.2.3; since the response variables are produced in such a manner that they can be physically interpreted, relationships between them have been discussed in political science literature [127].

5.2.1   A linear model in the demographics

In the first instance, the first eight dimension-reduced demographic variables will be used as predictors for the first two dimension-reduced political variables, with no basis expansions. Labelling the two political variables Y_1 and Y_2, and the demographic variables $X_1, X_2, \ldots, X_8$, this approach yields the model

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 0 & 3.09 & -1.28 & -0.32 & -0.61 & 4.70 & -0.20 & -1.99 & 0.39 \\ 0 & 0.46 & 3.00 & -0.83 & -1.94 & -0.66 & -1.10 & 1.91 & -0.79 \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_8 \end{bmatrix} + \varepsilon,$$

where $\varepsilon \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4.11^2 & 0 \\ 0 & 3.25^2 \end{bmatrix} \right)$.

Recalling that the first two predictors characterise, in a rough sense, levels of education and stage of life respectively, we can interpret this model for mean subjects. The $(1, 2)$ entry of the $\hat{\beta}$ matrix represents that there is a positive relationship between level of education and social inclusivity; in other words, people who have experienced more extensive education are more likely to be more socially inclusive. For a one unit increase in education level, there is a 3.09 unit increase in mean level of social inclusivity. Similar results include that:

- Older people are less likely to be socially inclusive. A one unit increase in 'stage of life' is associated with an average 1.28 unit decrease in social inclusivity.

- More extensively-educated people are slightly more likely to trust powerful institutions. A one unit increase in the education level variable is associated with an average 0.46 unit increase in the 'trust in authority' variable.

- Age is a much stronger indicator of trust in authority than education level; a one unit increase in 'stage of life' is accompanied by an average 3.00 unit increase in the 'trust in authority' variable.

It is important to note in interpreting such a model that the units of each variable do not have a direct physical meaning. Each dimension-reduced variable is scaled by the amount of variation in the original data set it 'captures' (See Chapter 3.1.2). For example, if a dimension-reduced variable has variance 5 units, this means that the dimension-reduced variable accounts for the variance of five of the unreduced variables. For the purpose of this initial linear model, a more direct physical interpretation of the relationships exposed is thus not explored.

Tests for linearity and homoscedasticity

Scatter plots of the residuals of the linear model in each response direction Y_1 and Y_2 against fitted values and all predictors can be found in Figures 5.1 and 5.2, and Figures 5.3 and 5.4, respectively.

For the assumption of linearity in the residuals to be reasonable, there should ideally be no discernible trend in the residuals. This entails the plots in Figures 5.1, 5.2, 5.3 and 5.4 having a linear shape around the horizontal axis. As can be seen in these figures, there is not a clear trend in any of the plots, making it likely that the assumption $E[\varepsilon | \boldsymbol{X}, \hat{\boldsymbol{Y}}] = \boldsymbol{0}$ is reasonable. There is an unusual horizontal spread of data in these plots, but there is not any non-linearity present. The horizontal spread arises from the dimension-reduced predictor variables being produced from categorical observed variables, meaning there are discrete clusters in the predictor variables. This does not affect the assumption of linearity, which is justified on the grounds that there is not a trend in the residuals dependent on any of the predictors.

There should also be constant variance in the residuals, that is, we require $\text{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}, \hat{\boldsymbol{Y}}) = \Sigma$ for some covariance matrix Σ. This is to ensure that estimates can be made in a consistent manner, with clearly defined error bounds. For the assumption to be reasonable, spread should be even throughout Figures 5.1, 5.2, 5.3 and 5.4. These plots demonstrate that this assumption is likely reasonable; while there is apparent lesser spread at the extreme values of each predictor, this can be explained by the smaller number of data points in these regions.
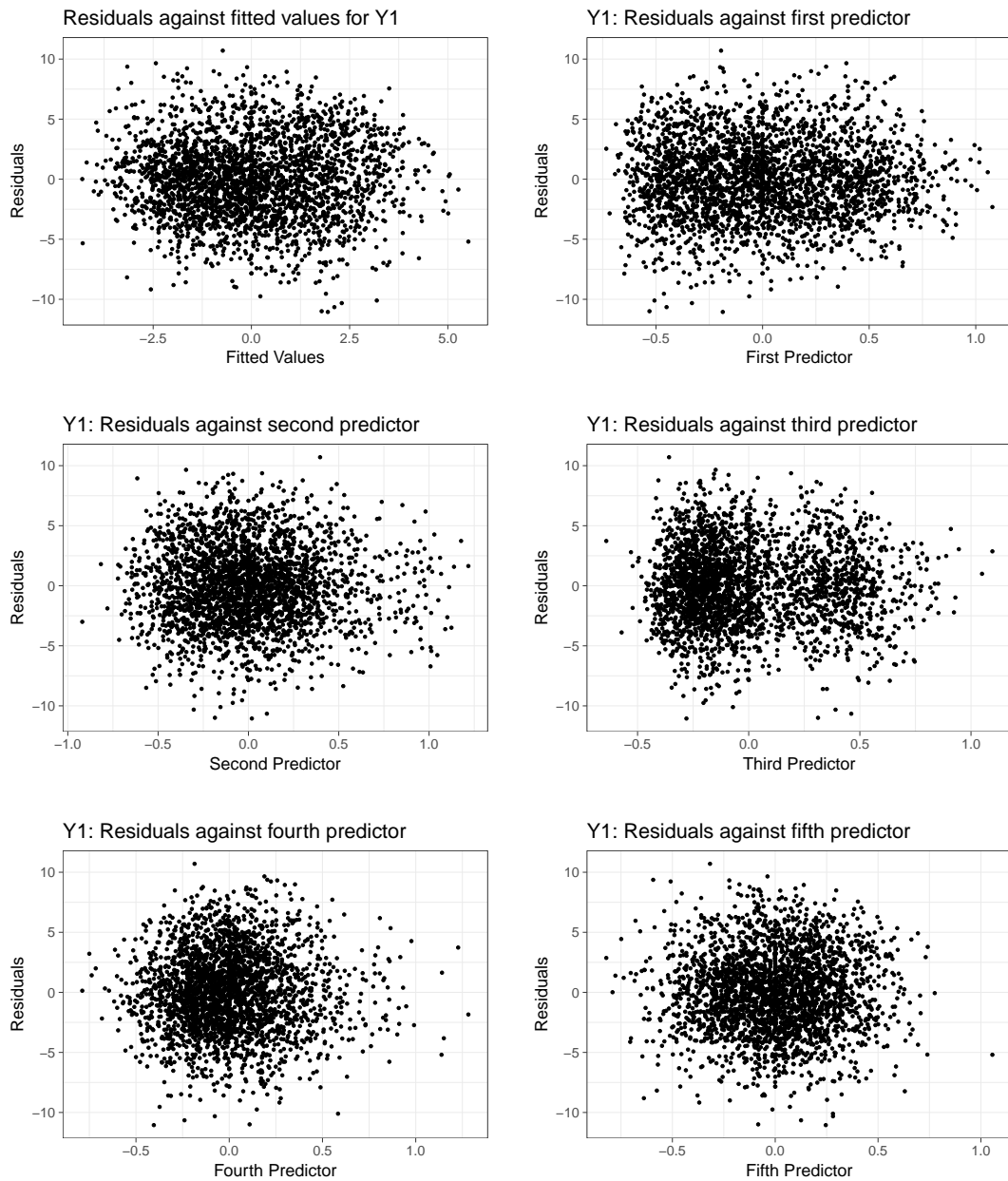
Figure 5.1: Residuals against fitted, and residuals against predictor, plots for response variable Y_1, and predictors $X_1, X_2, \ldots, X_5$. There does not appear to be any trend, or uneven spread, on any of these plots, that is related to the residual distribution; all unusual shape is due to the clustered distributions of the predictor variables. This shape does not violate the assumptions of linearity and homoscedasticity, which are reasonable here.
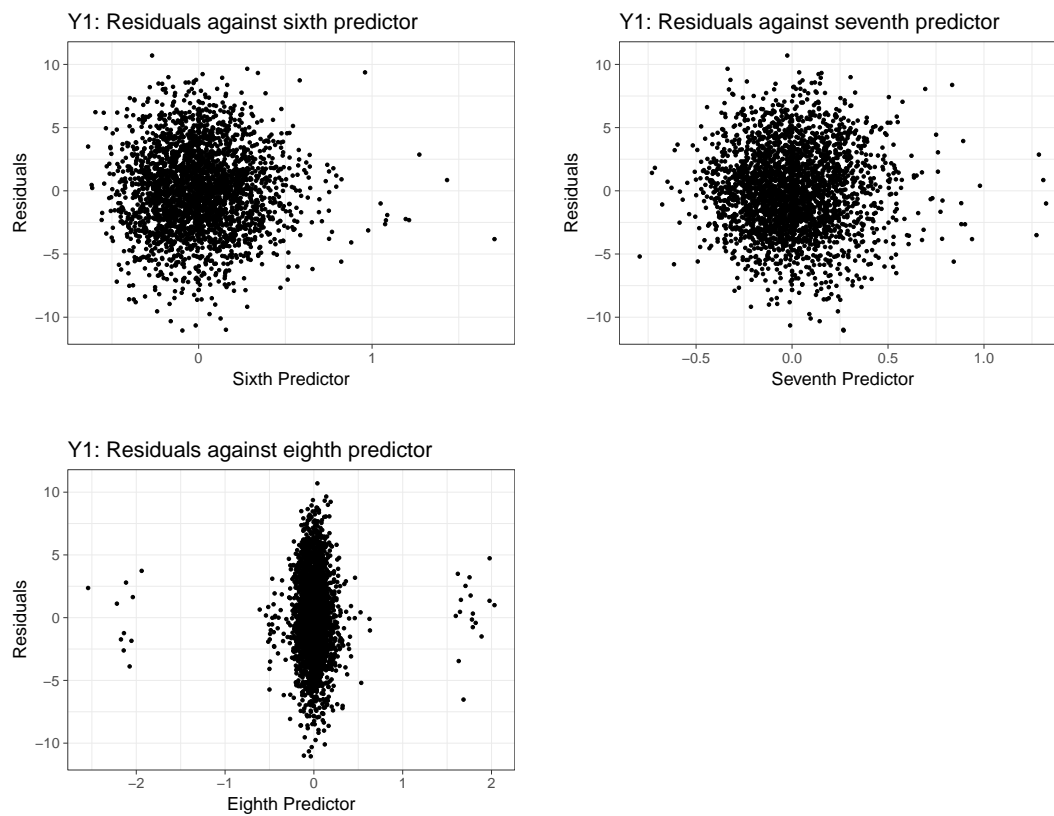
Figure 5.2: Residuals against predictor plots for response variable Y_1, and predictors X_6, X_7 and X_8. There does not appear to be any trend, or uneven spread, on any of these plots, that is related to the residual distribution; all unusual shape is due to the clustered distributions of the predictor variables. This shape does not violate the assumptions of linearity and homoscedasticity, which are reasonable here.
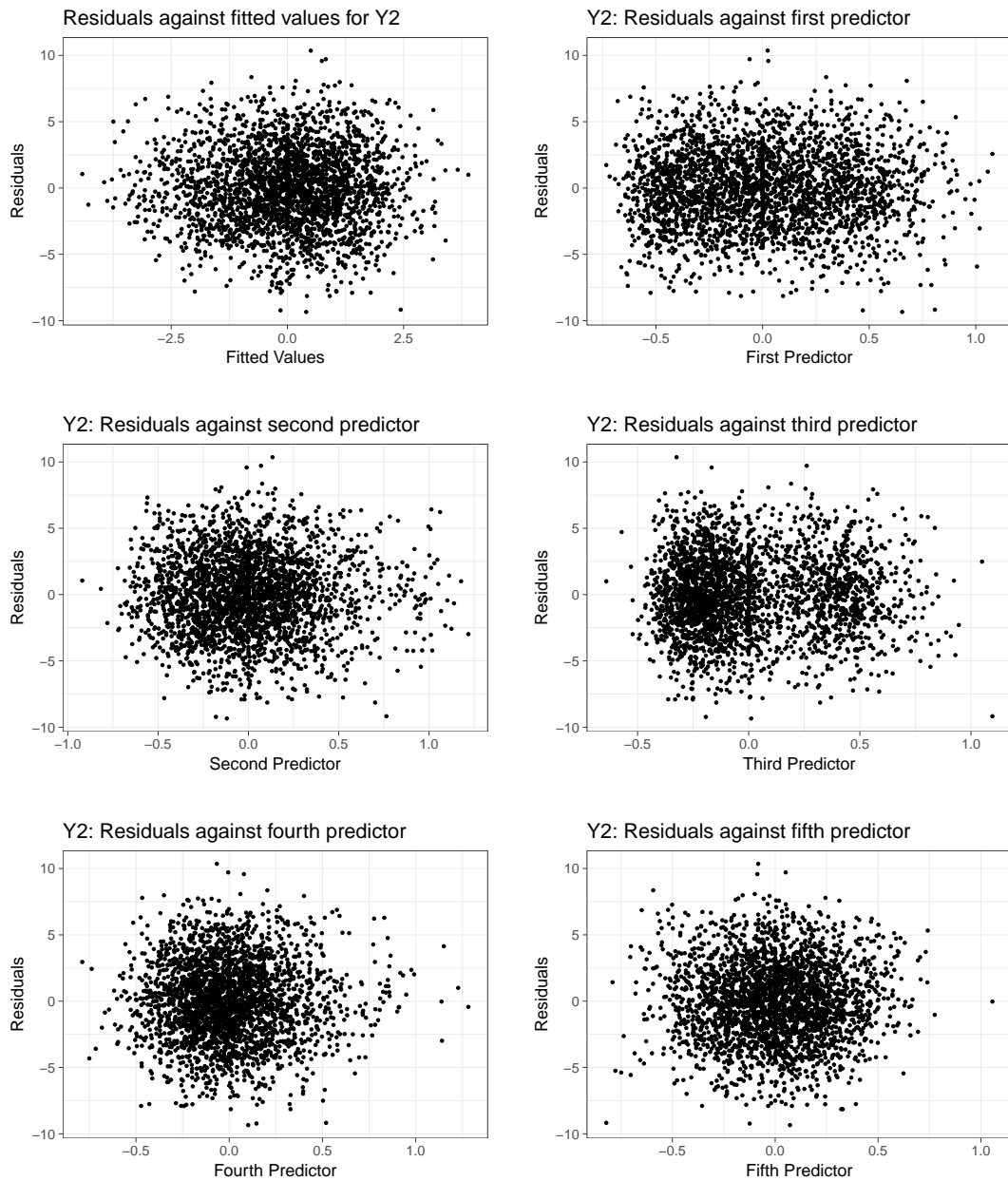
Figure 5.3: Residuals against fitted, and residuals against predictor, plots for response variable Y_2, and predictors $X_1, X_2, \ldots, X_5$. There does not appear to be any trend, or uneven spread, on any of these plots, that is related to the residual distribution; all unusual shape is due to the clustered distributions of the predictor variables. This shape does not violate the assumptions of linearity and homoscedasticity, which are reasonable here.
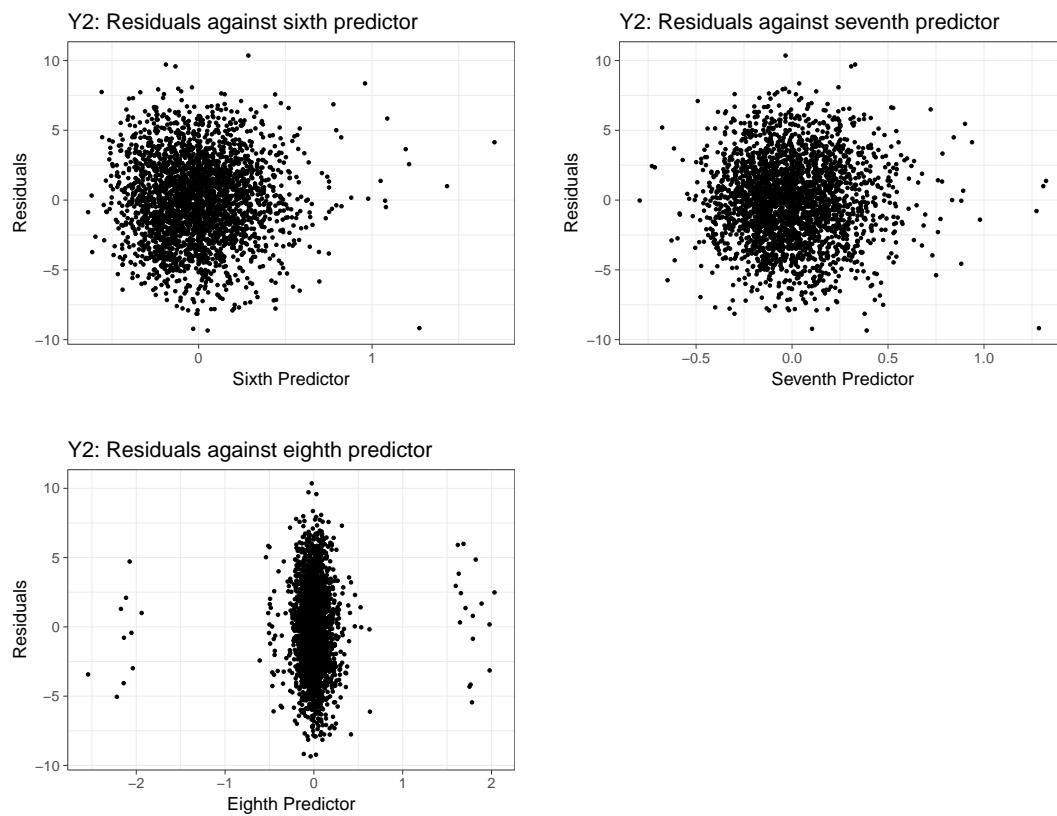
Figure 5.4: Residuals against predictor plots for response variable Y_2, and predictors X_6, X_7 and X_8. There does not appear to be any trend, or uneven spread, on any of these plots, that is related to the residual distribution; all unusual shape is due to the clustered distributions of the predictor variables. This shape does not violate the assumptions of linearity and homoscedasticity, which are reasonable here.
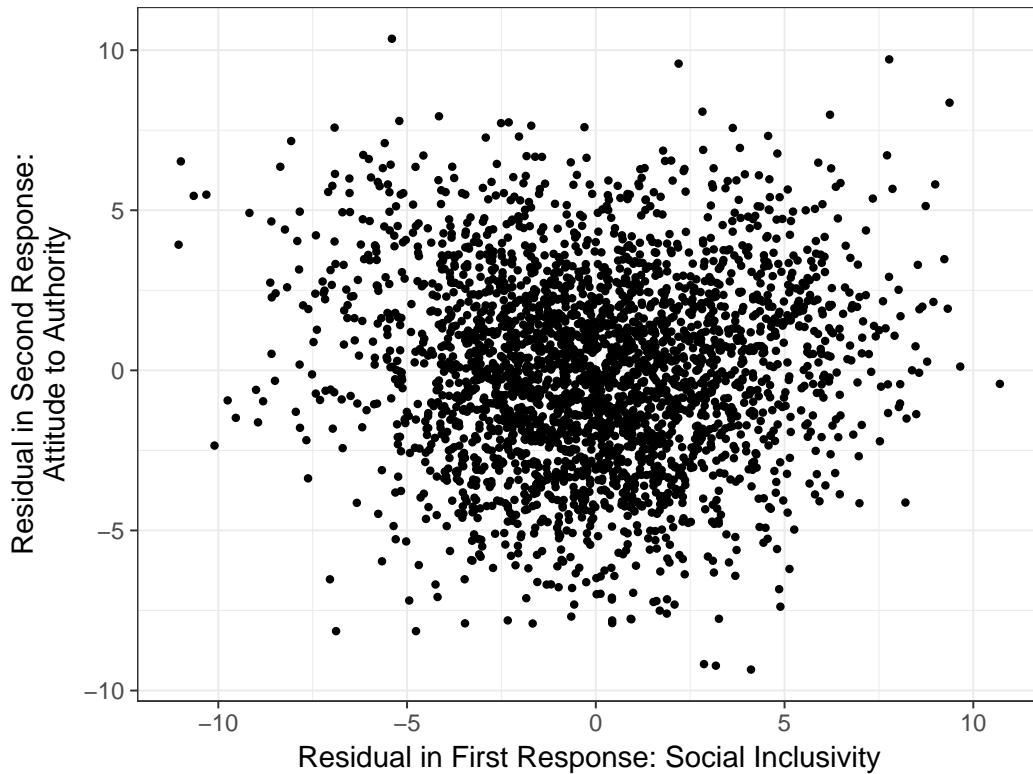
Figure 5.5: Scatter plot of residuals in the Y_2 direction against residuals in the Y_1 direction, from a model with eight linear predictors. There appears to be some heart-shaped dependence structure between the residuals, indicating that the assumption of multivariate normality cannot be justified.

Tests for normality

In testing the assumption of multivariate normality, visual diagnostics are first considered. If there is some non-linear dependence structure between the residuals, they cannot be multivariate Gaussian, since the normal distribution can only account for linear relationships between variables. Figure 5.5 contains a scatter plot of the residuals in the Y_1 and Y_2 directions. It can be seen that there appears to be a non-linear association between the response variables' residuals; some curvature exists in the scatter plot. Adding a LOESS line appears to reinforce this (Figure 5.6), with substantial curvature apparent. This is further reinforced by constructing a regression model with the second residual dependent on a quadratic basis expansion of the first; the reported p-value for inclusion of a quadratic term ε_1^2 in predicting ε_2 is 7.99×10^{-27}.

Assessment of normality can also be undertaken using the energy test described by Székely and Rizzo [166]. As described in Section 5.1.2, the energy test uses the pairwise expected distances between the observed points and points from a normal distribution to determine how dissimilar the observed values are to normally distributed data. The test can be likened to ANOVA; the distances be-
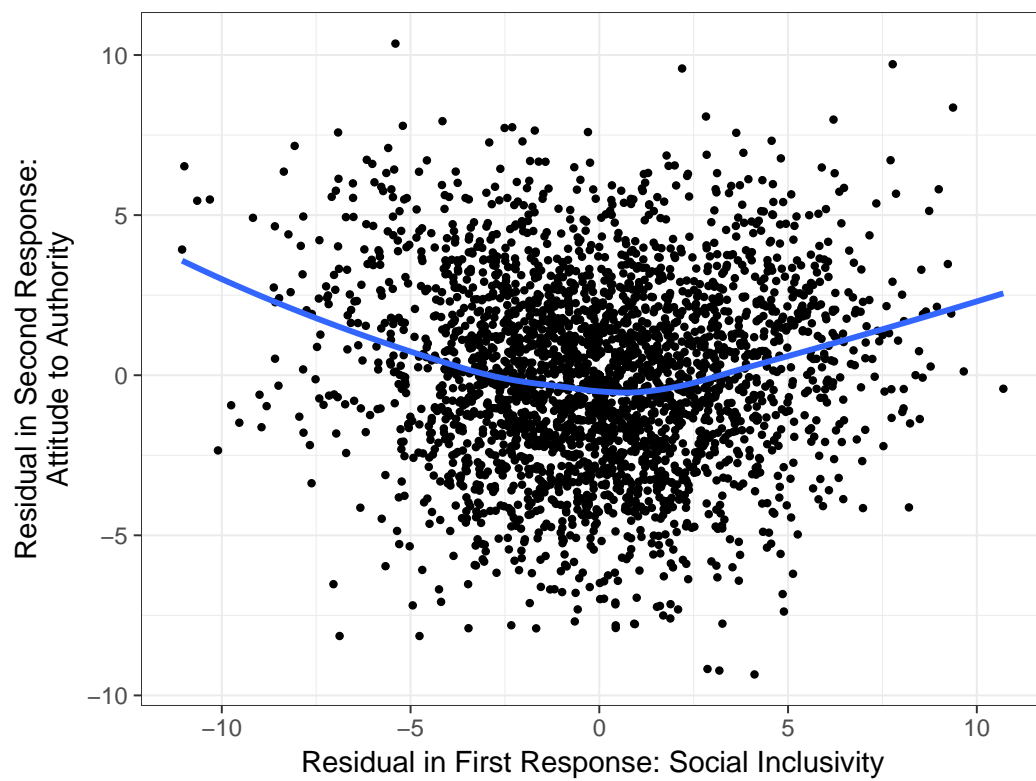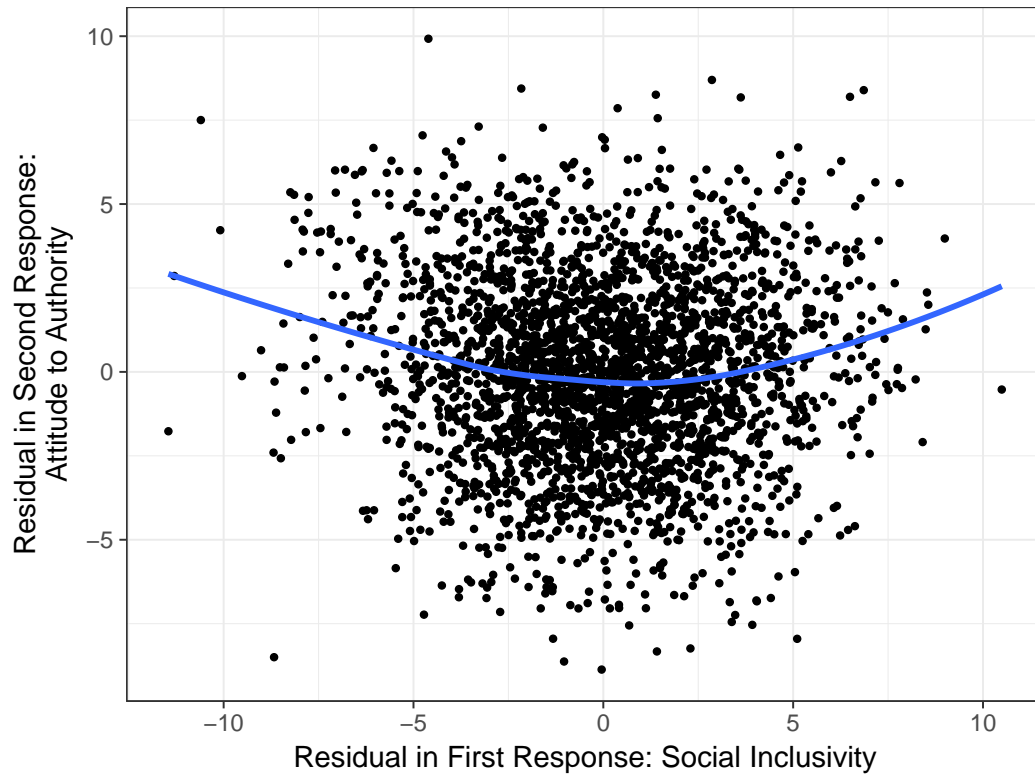
Figure 5.6: Scatter plot of residuals in the Y_2 direction against residuals in the Y_1 direction, from a model with eight linear predictors, with a LOESS line added for perspective. There is clear curvature in the relationship between residuals, meaning that the residuals are highly unlikely to have been produced as deviates of a multivariate normal distribution.

tween distributions are compared with the distances within distributions. If the between-distribution distances are substantially higher than the within-distribution distances, the distributions are likely different. The energy test is chosen since it is among the more consistently powerful tests for multivariate normality implemented in R whose reported type I error rate is accurate [99]. Here, the energy test gives a p-value for multivariate normality which is much smaller than 0.001, indicating that it is highly unlikely that a multivariate normal distribution would produce deviates as unusual as these.

As a result of the output from the visual and numerical tests, the assumption of multivariate normality cannot be justified. The model estimating Y_1 and Y_2 using just the first eight dimension-reduced demographic variables is therefore inappropriate, since there is a non-linear dependence structure in the residuals.

As will be discussed in Section 5.2.3, this non-linear dependence structure is also justified from a social science standpoint, since it represents that unexpectedly extreme views with respect to social inclusivity are associated with a lack of trust in authority.

It may be that some more complex model is able to explain away the non-linear dependence structure in the residuals. If this were true, it would be possible to build a more complex model such that the assumption of multivariate normality is justified. To determine whether it is indeed possible to build such a model, additional predictors will be added to the existing model and a thorough exploration of basis expansions will be undertaken in subsequent sections.

Introducing additional predictors

We have displayed that a non-linear relationship exists between the residuals from a model estimating responses Y_1 and Y_2 using the first eight dimension-reduced demographic variables $X_1, \ldots, X_8$. This renders the multivariate model unusable, since the assumption of multivariate normality is unjustifiable. In an attempt to account for this relationship and fulfil the model assumptions, a larger subset of the dimension-reduced demographic variables will be included as predictors.

If it is possible to explain away the non-linear association between the residuals using some combination of the dimension-reduced demographic variables, then a model incorporating all of these dimension-reduced variables should have the desired residual structure. In other words, if a full model of all linear predictors is insufficient, then no model comprising a subset of these predictors will be sufficient. We fit a model of $(Y_1, Y_2)^{\mathrm{T}}$ using 100 dimension-reduced demographic variables, $\{X_1, \ldots, X_{100}\}$. The full output of the model is not provided here, as there are 202 parameters in β.

To test for normality of residuals, a plot of residuals in the first and second response direction is included in Figure 5.7, with a LOESS line to aid interpretation. It is clear in this plot that there is a non-linear association between the residuals, meaning they cannot be normally distributed. This is reinforced by regressing ϵ_2 upon ϵ_1^2, which yields a p-value of 9.02×10^{-16} for the inclusion of the quadratic term, demonstrating this non-linear structure. The energy test for multivariate

Figure 5.7: Scatter plot of residuals in the Y_2 direction against residuals in the Y_1 direction from a model with 100 linear predictors, with a LOESS line added for perspective. There is clear curvature in the relationship between residuals.

normality here again outputs a p-value smaller than 0.001, meaning the assumption of multivariate normality is not reasonable.

It appears that some relationship between the response variables exists that cannot be explained by a linear function of the demographic factors alone. The following section will thus consider a number of basis expansions to determine whether demographic models can explain away this dependence structure in the residuals. These basis expansions will include the introduction of polynomial, interaction, exponential, logarithmic and square root terms.

5.2.2   Basis expansions

As can be seen in Figure 5.7, in a linear model estimating the first two dimension-reduced political variables using the dimension-reduced demographic variables, there is a non-linear relationship between the residuals in the Y_1 and Y_2 directions. This means the assumption of multivariate normality of residuals (Section 5.1.2) is unreasonable. To circumvent this issue, we seek a model that explains away the non-linear relationship in the residuals. In this section, models are produced using additional predictor terms in the form of basis expansions of the dimension-reduced demographic variables. These basis expansions include polynomial, interaction,

exponential, logarithmic and square root terms.

It is hypothesised that if these basis expansions are unable to explain the dependence structure in the residuals, then the non-linear shape seen in Figure 5.7 cannot be explained using the demographic variables alone, meaning the assumption of multivariate normality of residuals cannot be justified and therefore that the multivariate regression model should not be favoured.

Polynomial terms

Using all 100 predictors, quadratic, cubic and quartic polynomial regression models were then fit to determine whether these could explain the non-linear dependence structure in the residuals. Due to the principle of marginality, the quadratic, cubic and quartic models have 402, 602 and 802 parameters within β respectively.

To test the assumption of multivariate normality, plots of the residuals in the Y_2 direction against the residuals in the Y_1 direction are provided for all three of these models in Figure 5.8. Curvature is apparent in all three plots. While the LOESS line from the residuals in the quartic model in Figure 5.8 does not appear to indicate a quadratic relationship, this is due to the influence of a single data point, and a non-linear relationship exists regardless. Since all three plots demonstrate non-linear relationships between the residuals, the assumption of multivariate normality is not reasonable.

This is confirmed by modelling the residuals in each of the models against each other. As in the linear model case, a quadratic relationship between ϵ_2 and ϵ_1^2 exists in all three models, with a p-value for inclusion of a quadratic term in a regression of ϵ_2 upon ϵ_1^2 being 1.28×10^{-12}, 3.71×10^{-10} and 1.29×10^{-8} for the quadratic, cubic and quartic models respectively. These increasing, but still strongly significant, p-values are much more likely the result of spurious relationships brought about by vast overfitting of the model, as opposed to a more complete explaining away of the residual dependence structure, due to the high degrees of freedom in the polynomial models in all 100 predictors.

The lack of justifiability of an assumption of multivariate normality in the residuals for the polynomial models is reinforced by the Skèkely and Rizzo's energy test for multivariate normality [166], which gives p-values smaller than 0.001 for all three models. We thus reject the assumption of normality in all of these polynomial models, meaning that these models are unable to explain away the non-linear dependence between residuals.

Interaction terms

A model was then fit with all possible interaction terms between the predictors, these being all products $\{X_i X_j : i, j = 1, 2, \ldots, 50, j \leq i\}$. Note that this includes interaction terms as well as quadratic terms in each predictor. Here a quadratic relationship in the residuals persisted, despite how substantially overfit the model has become, with 2652 parameters in β accounting for the principal of marginality, and just 2818 data points. The p-value for inclusion of the quadratic term in the model $\epsilon_2 \sim \epsilon_1^2 + \epsilon_1$ is 3.21×10^{-3}, meaning there remains a non-linear dependence structure
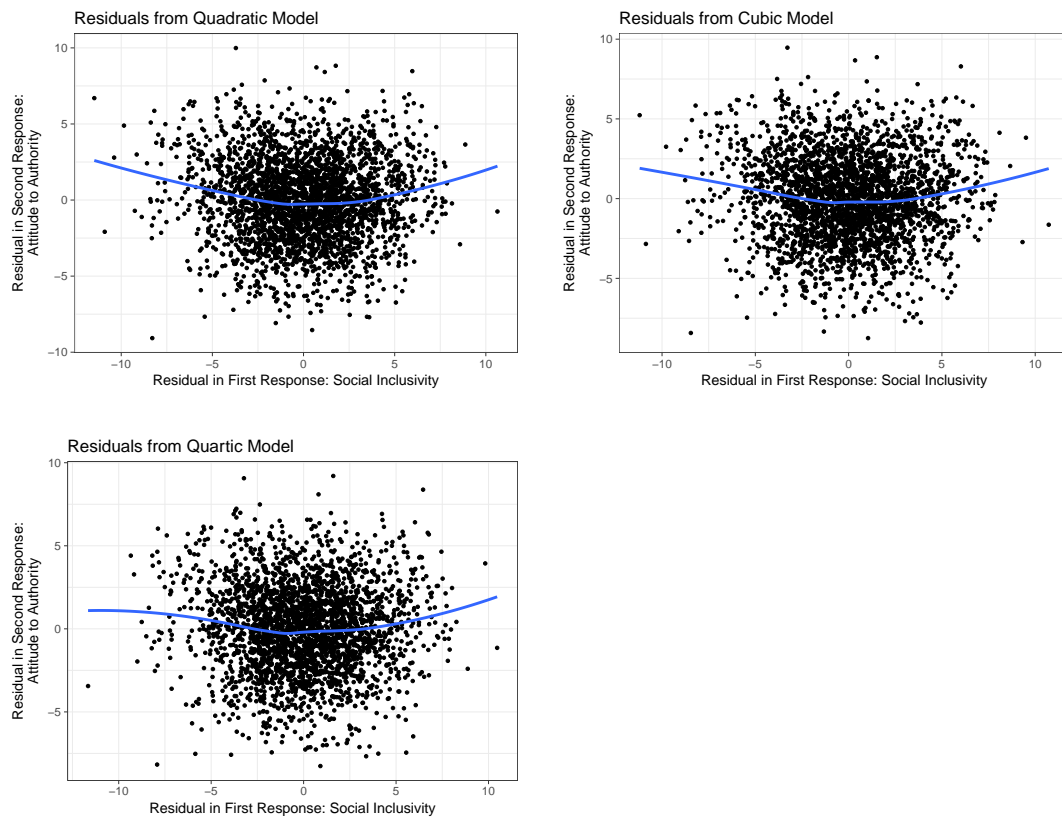
Figure 5.8: Scatter plots of residuals in the Y_2 direction against residuals in the Y_1 direction from models incorporating quadratic, cubic and quartic basis expansions of predictors, respectively. LOESS lines are added for perspective. There is clear curvature in the relationship between residuals in all three plots, indicating that the assumption of multivariate normality of residuals cannot be justified for any of these multivariate polynomial regression models.

| | P-value | |
Model	Quadratic term inclusion	Energy test
Logarithmic	7.18×10^{-10}	< 0.001
Exponential	1.02×10^{-10}	< 0.001
Square root	1.58×10^{-12}	< 0.001

Table 5.1: P-values for hypothesis tests whose null hypothesis is true if the residuals are multivariate normal. The 'Quadratic' column contains tests for a quadratic relationship between ϵ_2 and ϵ_1, where there should be no relationship for normally distributed data. The 'Energy' column refers to the energy test for multivariate normality described by Skèkely and Rizzo [166]. It can be seen that all p-values are more than sufficient to reject the null hypothesis of multivariate normality.

in the residuals. This is reinforced by considering the energy test for multivariate normality, which gives a p-value smaller than 0.001. Multivariate normality is still not a justifiable assumption in the case with quadratic and interaction terms.

Further functions

To explore whether other functional forms might better explain the non-linear dependence structure in the residuals, models with logarithmic, exponential and square root functions of the predictors were considered. When taking logarithms and square roots, predictors were shifted upwards by a constant value, such that the minimum value of each was one, to prevent undefined values. For each model, normality was tested in two ways: by testing for significance of a quadratic term in a linear model of ϵ_2 against ϵ_1^2, and by using the energy test for multivariate normality. As can be seen in Table 5.1, all tests are strongly significant, meaning that normality should be rejected. It can thus be concluded that multivariate normality of residuals is an invalid assumption for any of these basis expanded models.

5.2.3 Conclusions on the use of multivariate regression

A non-linear relationship exists between the first and second response variables, which remained in the residuals of a model using the first eight demographic predictors. This non-linear relationship precludes the residuals from being multivariate Gaussian. Additional predictors were added to try and explain the non-linear relationship in the residuals, but these were unsuccessful. Polynomial basis expansions, including quadratic, cubic and quartic regressions, were also tried, but again the relationship in the residuals persisted. Interaction terms were added, to build a vastly overfit model, and this too was insufficient. Finally, more involved basis expansions were chosen in the form of logarithmic, exponential and square root functions. These too were unable to remove the non-linear dependence structure in the residuals. As such, none of the models tried satisfied the multivariate regression

assumption of normality. This means that none of the multivariate regression models described in Sections 5.2.1-5.2.2 have justified assumptions, so none should be used.

The argument that this model structure is insufficient is bolstered by the fact that the non-linear relationship between the residuals in the Y_1 and Y_2 directions makes physical sense, given what the dimensions represent. The dependence structure indicates a convex relationship between the residuals in the direction labelled "social inclusivity" and the residuals in the direction labelled "attitudes to authority" (see Section 4.4). The "social inclusivity" axis is dominated by questions about immigration and Aboriginal and Torres Strait Islander affairs, and the "attitude to authority" axis is characterised by questions regarding trust in government institutions, corporations and unions. Thus we can describe the non-linear dependence structure in the following way: when subjects are much more, or much less, socially inclusive than a regression model would expect them to be, these subjects are more likely to have high levels of distrust in authority. This is a relationship that makes physical sense; those with unexpectedly extreme social views are likely to take a dim view of mainstream institutions [127]. For example, one who takes an extreme view of issues relating to immigration is likely to seek some sort of social change, and so is unlikely to have faith in institutions like the Federal Parliament and Courts, who represent the status quo. As such, the non-linear relationship between the residuals in the first and second dimension-reduced response variable directions looks to be justified from a social science standpoint.

Both an empirical view of possible regression models, and an understanding of the social science foundations for the models, suggest that the complex relationship between the first and second dimension-reduced political variables may exist independently of the available demographic variables. This non-linear dependence structure in the residuals being justified, it might be more useful to attempt to build a model able to include the unusual error, as opposed to building more and more complex models with a view to explaining away the error structure.

Chapter 6 will look at three alternative models. The first of these models' advantage lies in being built sequentially, rather than simultaneously. In parameter estimation for the multivariate regression model, all elements of the β matrix are estimated simultaneously, followed by all elements of $\hat{\Sigma}_\varepsilon$. This means that, for example, estimates of Y_2 cannot respond to the model for Y_1. Section 6.1 suggests and applies a model-building structure in which the error in each direction is recursively captured by the model, so the model can respond to the unusual residual structure exhibited in this Chapter. Sections 6.2 and 6.3 then describe two further models built to respond to the residual structure, using different distributional assumptions on the models' error term than those of the multivariate regression model.

Chapter 6

Alternatives to the multivariate regression model

Chapter 5 explored the foundational linear model for multidimensional responses, namely the multivariate regression model. This approach models a set of response variables $\boldsymbol{Y}$ as the sum of an affine transformation of predictor variables $\boldsymbol{X}$, and some multivariate normal random variable, denoted $\boldsymbol{\varepsilon}$. Section 5.2 demonstrated that such a model is insufficient to describe the relationships between demography and political opinion as reflected in the 2016 Australian Election Study, since the assumption of multivariate normality of the residuals does not hold. This is due to the presence of a non-linear relationship between the residual axes, which cannot be explained by the normally-distributed error term in the multivariate regression model.

This chapter proposes three alternative models to the multivariate regression model. All of the three models are extensions to simpler regression models, built so as to be able to encapsulate non-linear relationships between residuals in different directions. The first, the recursive multivariate model (Section 6.1), is built around the construction of successive dependent univariate regression models in each response variable. The second, the mixture model (Section 6.2), is expressed identically to the multivariate regression model, save that the error term $\boldsymbol{\varepsilon}$ is replaced by a Gaussian mixture of multivariate normal random variables denoted $\boldsymbol{\varepsilon_{mix}}$. Similarly, the third model, the kernel density model (Section 6.3), replaces the multivariate regression error term $\boldsymbol{\varepsilon}$ with a kernel density estimate on the residuals denoted $\boldsymbol{\varepsilon_{kde}}$. This chapter explains each of the three models, and the derivations of their parameters, before they are compared in Chapter 7.

6.1 A recursive model

The recursive model is constructed with an eye to respond directly to non-linear relationships between residuals, by modelling the residual terms as functions of one another. This allows for the creation of more complex, and non-normal, error structures in observed data to be reflected in a model.

Suppose we have some set of k response variables $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_k]^{\mathrm{T}}$, modelled

on some set of p predictor variables $\boldsymbol{X} = [1, X_1, X_2, \ldots, X_p]^{\mathrm{T}}$, with the first element of $\boldsymbol{X}$ corresponding to an intercept term. The recursive model building process begins by fitting Y_1 against $\boldsymbol{X}$, using a univariate regression model (see Section 5.1.1). Y_2 is then fit against both $\boldsymbol{X}$ and some non-linear function of ε_1, the error term in the model for Y_1. This continues with Y_3 fit against $\boldsymbol{X}$ and non-linear functions of ε_1 and ε_2, and so on until the univariate model of Y_k is constructed from $\boldsymbol{X}$, and non-linear functions of $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{k-1}\}$. We can thus express the most general form of the recursive model as follows:

$$Y_1 = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta_1} + \varepsilon_1$$
$$Y_2 = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta_2} + f_{2,1}(\varepsilon_1) + \varepsilon_2$$
$$\vdots$$
$$Y_k = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta_k} + f_{k,1}(\varepsilon_1) + f_{k,2}(\varepsilon_2) + \cdots + f_{k,k-1}(\varepsilon_{k-1}) + \varepsilon_k,$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$ for $i = 1, 2, \ldots, k$ independently, and $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots \boldsymbol{\beta_k}$ are $(p+1)\times 1$ vectors of coefficients. Fitting this model also requires the estimation of the parameters of the $\frac{k(k-1)}{2}$ functions $f_{i,j}(\varepsilon_j)$ for $i = 2, \ldots, k$, $j = 1, \ldots, i-1$. In order for this model to be easily fit in a univariate regression framework, using least squares, the functions $f_{i,j}(\varepsilon_j)$ are usually chosen to be linear in basis expansions of ε_j.

In the 2016 Australian Election Study application, the recursive multivariate model is chosen to respond to what appears to be a quadratic relationship between ε_1 and ε_2 (See Figure 5.6), as well as what appears to be quadratic relationships between errors in other directions. As such, the model in the first three response variables is constructed as follows:

$$Y_1 = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta_1} + \varepsilon_1$$
$$Y_2 = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta_2} + \alpha_{20} + \alpha_{21}\varepsilon_1^2 + \alpha_{22}\varepsilon_1 + \varepsilon_2$$
$$Y_3 = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta_3} + \alpha_{30} + \alpha_{31}\varepsilon_1^2 + \alpha_{32}\varepsilon_1 + \alpha_{33}\varepsilon_2^2 + \alpha_{34}\varepsilon_2 + \varepsilon_3, \qquad (6.1.1)$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$ for $i = 1, 2, 3$ independently, $\alpha_{20}, \alpha_{21}, \alpha_{22}, \alpha_{30}, \alpha_{31}, \alpha_{32}, \alpha_{33}$, and α_{34} are constants, and $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ are $(p+1) \times 1$ vectors of coefficients. While the existence of two intercept terms appears redundant, these are included for ease of fitting the model, as will be explained in Section 6.1.1.

6.1.1 Fitting the model

A danger in building a model with this large number of interrelated error terms is that the model might become dominated by these errors, impeding the accuracy of conditional expectations in the name of better explaining the error structure. For example, suppose we have response variables Y_1 and Y_2 such that $Y_2 = Y_1^2 + \zeta_1$ for small error ζ_1, and predictor $X = Y_2 + \zeta_2$ for small error ζ_2, and so $X = Y_1^2 + \zeta_1 + \zeta_2$. If Y_1 is mean-centred at zero, there is no linear relationship between X and Y_1, since the relationship $X = Y_1^2 + \zeta_1 + \zeta_2$ is purely quadratic in Y_1. This means $\beta_1 = 0$, $\hat{\beta}_1 \simeq 0$, and so for the construction in Equation (6.1.1), $\varepsilon_1 \simeq Y_1$. Looking to Y_2, we can note that $Y_2 \simeq X$ and $Y_2 \simeq \varepsilon_1^2$. Since both X and ε_1^2 are predictors in the model

for Y_2, it may be that either predictor dominates the model for Y_2. However, when it comes to calculating the conditional expectation of Y_1 and Y_2 at some new value of X, X is observed but ε_1 is not. Thus by fitting the model in the observable and unobservable predictors simultaneously, there is potential to reduce the accuracy of the conditional expectations by assigning equal value to the unobservable predictors. As such, more generally, the recursive model should be explicitly designed to fit the predictor set $\boldsymbol{X}$ in preference to the errors ε_i.

To achieve this goal, the parameters $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_k}$ for the recursive model are estimated first by ordinary least squares, without regard to any of the error terms $\varepsilon_1, \ldots, \varepsilon_k$. The parameters α_{20}, α_{21} and α_{22} are then estimated with $\hat{\boldsymbol{\beta}}_1$ fixed, parameters $\alpha_{30}, \alpha_{31}, \alpha_{32}, \alpha_{33}$, and α_{34} estimated with $\hat{\boldsymbol{\beta}}_2$ fixed, and so on. In other words, the process for estimating the parameters of the recursive multivariate model are as follows:

1. Estimate $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_k$ by ordinary least squares, that is,

$$\hat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta_i}}{\operatorname{argmin}} \left\| \boldsymbol{y_i} - X^{\mathrm{T}} \boldsymbol{\beta_i} \right\|^2 = (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} \boldsymbol{y_i},$$

   for $i = 1, 2, \ldots, k$, $\boldsymbol{y_i}$ an $n \times 1$ vector of observations of Y_i, and X a $n \times (p+1)$ matrix of observed predictors, with a leading column of ones.

2. Estimate α_{20}, α_{21} and α_{22} by ordinary least squares with $\hat{\boldsymbol{\beta}}_1$ fixed, that is,

$$\{\alpha_{20}, \alpha_{21}, \alpha_{22}\} = \underset{\{\alpha_{20}, \alpha_{21}, \alpha_{22}\}}{\operatorname{argmin}} \left\| (\boldsymbol{y_1} - X^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1) - (\alpha_{20}\mathbf{1} + \alpha_{21}\boldsymbol{\varepsilon_1^2} + \alpha_{22}\boldsymbol{\varepsilon_1}) \right\|^2,$$

   where $\boldsymbol{\varepsilon_1}$ is the vector of the residuals in the Y_1 direction and $\boldsymbol{\varepsilon_1^2}$ is the vector of the squared residuals in the Y_1 direction.

3. Estimate $\alpha_3, \alpha_{31}, \alpha_{32}, \alpha_{33}$, and α_{34} by ordinary least squares with $\hat{\boldsymbol{\beta}}_2$ fixed, that is,

$$\{\alpha_{30}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \alpha_{34}\} =$$

$$\underset{\{\alpha_{30}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \alpha_{34}\}}{\operatorname{argmin}} \left\| (\boldsymbol{y_2} - X^{\mathrm{T}}\hat{\boldsymbol{\beta}}_2) - (\alpha_{30}\mathbf{1} + \alpha_{31}\boldsymbol{\varepsilon_1^2} + \alpha_{32}\boldsymbol{\varepsilon_1} + \alpha_{33}\boldsymbol{\varepsilon_2^2} + \alpha_{34}\boldsymbol{\varepsilon_2}) \right\|^2.$$

4. Continue in like fashion for the remaining α_{ij} terms.

It appears from the above that the conditional expectations of the response variables of the recursive multivariate model are similar to those both in univariate, and multivariate, regression, since the estimates of the $\boldsymbol{\beta_i}$s are identical. We now prove that for some new observation $\boldsymbol{x_0}$, $\hat{Y}_2|\boldsymbol{x_0}$ is identical under both the recursive multivariate model and the multivariate regression model.

Theorem 6.1.1. *The multivariate regression model produces the same conditional expectation of the response variables as under the recursive multivariate model; that is,*

$$\hat{Y}_2|\boldsymbol{x_0}, Multi = \hat{Y}_2|\boldsymbol{x_0}, Recursive$$

for some new observation of predictors $\boldsymbol{x_0}$ and fitted models Multi and Recursive.

Proof. Note first that $\hat{\boldsymbol{\beta}}_2^{multi} = \hat{\boldsymbol{\beta}}_2^{recursive} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y_2}$. Thus,

$$\hat{Y}_2|\boldsymbol{x_0}, Multi = \boldsymbol{x_0}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_2^{multi}$$
$$:= \boldsymbol{x_0}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_2. \tag{6.1.2}$$

Now for the recursive model,

$$\hat{Y}_2|\boldsymbol{x_0}, Recursive = \boldsymbol{x_0}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_2^{recursive} + \hat{\alpha}_{20} + \hat{\alpha}_{21}\hat{E}[\varepsilon_1^2] + \hat{\alpha}_{22}\hat{E}[\varepsilon_1] + \hat{E}[\varepsilon_2].$$

Now in the model for Y_1, ε_1 was estimated to have distribution $N(0, \sum_{i=1}^n \varepsilon_{1i}^2)$, meaning $\hat{E}[\varepsilon_1] = 0$ and $\hat{E}[\varepsilon_1^2] = \frac{1}{n}\sum_{i=1}^n \varepsilon_{1i}^2$. We also have that $\hat{\boldsymbol{\beta}}_2^{recursive} = \hat{\boldsymbol{\beta}}_2$ and that $\hat{E}[\varepsilon_2] = 0$:

$$\hat{Y}_2|\boldsymbol{x_0}, Recursive = \boldsymbol{x_0}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_2 + \hat{\alpha}_{20} + \hat{\alpha}_{21}\frac{1}{n}\sum_{i=1}^n \varepsilon_{1i}^2. \tag{6.1.3}$$

Estimates $\hat{\alpha}_{20}$ and $\hat{\alpha}_{21}$ are now sought. Using the construction

$$A = \begin{pmatrix} 1 & \varepsilon_{11}^2 & \varepsilon_{11} \\ 1 & \varepsilon_{12}^2 & \varepsilon_{12} \\ \vdots & \vdots & \vdots \\ 1 & \varepsilon_{1n}^2 & \varepsilon_{1n} \end{pmatrix},$$

The least squares construction of the α_{2j} estimates is

$$\begin{pmatrix} \hat{\alpha}_{20} \\ \hat{\alpha}_{21} \\ \hat{\alpha}_{22} \end{pmatrix} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}(\boldsymbol{y_2} - X^{\mathrm{T}}\hat{\boldsymbol{\beta}}_2).$$

Noting that

$$A^{\mathrm{T}}A = \begin{pmatrix} n & \sum_{i=1}^n \varepsilon_{1i}^2 & \sum_{i=1}^n \varepsilon_{1i} \\ \sum_{i=1}^n \varepsilon_{1i}^2 & \sum_{i=1}^n \varepsilon_{1i}^4 & \sum_{i=1}^n \varepsilon_{1i}^3 \\ \sum_{i=1}^n \varepsilon_{1i} & \sum_{i=1}^n \varepsilon_{1i}^3 & \sum_{i=1}^n \varepsilon_{1i}^2 \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n \varepsilon_{1i}^2 & 0 \\ \sum_{i=1}^n \varepsilon_{1i}^2 & \sum_{i=1}^n \varepsilon_{1i}^4 & \sum_{i=1}^n \varepsilon_{1i}^3 \\ 0 & \sum_{i=1}^n \varepsilon_{1i}^3 & \sum_{i=1}^n \varepsilon_{1i}^2 \end{pmatrix}, \qquad \left(\text{since } \sum_{i=1}^n \varepsilon_{1i} = 0\right)$$

We can use the well-known formula for inversion of a 3×3 matrix (e.g. [177]) to obtain

$$(A^{\mathrm{T}}A)^{-1} =$$

$$\frac{1}{|A|}\begin{pmatrix} \left(\sum_{i=1}^{n}\varepsilon_{1i}^4\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)-\left(\sum_{i=1}^{n}\varepsilon_{1i}^3\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^3\right) & -\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right) & \left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^3\right) \\ -\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)+\left(\sum_{i=1}^{n}\varepsilon_{1i}\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^3\right) & n\sum_{i=1}^{n}\varepsilon_{1i}^2 & -n\sum_{i=1}^{n}\varepsilon_{1i}^3 \\ \left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^3\right) & -n\sum_{i=1}^{n}\varepsilon_{1i}^3 & n\sum_{i=1}^{n}\varepsilon_{1i}^4-\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right) \end{pmatrix}.$$

Now since $A^{\mathrm{T}}(\boldsymbol{y_2} - X^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})$

$$= \begin{pmatrix} \sum_{i=1}^{n}(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}}) \\ \sum_{i=1}^{n}\varepsilon_{1i}^2(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}}) \\ \sum_{i=1}^{n}\varepsilon_{1i}(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}}) \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ \sum_{i=1}^{n}\varepsilon_{1i}^2(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}}) \\ \sum_{i=1}^{n}\varepsilon_{1i}(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}}) \end{pmatrix},$$

$$(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}(\boldsymbol{y_2} - X^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})$$

$$= \frac{1}{|A|}\begin{pmatrix} -\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})\right)+\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^3\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})\right) \\ n\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})\right)-n\left(\sum_{i=1}^{n}\varepsilon_{1i}^3\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})\right) \\ -n\left(\sum_{i=1}^{n}\varepsilon_{1i}^3\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})\right)+\left(n\sum_{i=1}^{n}\varepsilon_{1i}^4-\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})\right) \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\alpha}_{20} \\ \hat{\alpha}_{21} \\ \hat{\alpha}_{22} \end{pmatrix}.$$

Thus,

$$\hat{\alpha}_{20} = \frac{1}{|A|}\left(-\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2\right)\left(\sum_{i=1}^{n}\varepsilon_{1i}^2(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}}\hat{\boldsymbol{\beta_2}})\right)\right)$$

$$+ \left(\sum_{i=1}^{n} \varepsilon_{1i}^2 \right) \left(\sum_{i=1}^{n} \varepsilon_{1i}^3 \right) \left(\sum_{i=1}^{n} \varepsilon_{1i}(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}} \hat{\boldsymbol{\beta_2}}) \right) \Bigg) , \text{ and}$$

$$\hat{\alpha}_{21} = \frac{1}{|A|} \left(n \left(\sum_{i=1}^{n} \varepsilon_{1i}^2 \right) \left(\sum_{i=1}^{n} \varepsilon_{1i}^2(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}} \hat{\boldsymbol{\beta_2}}) \right) \right.$$

$$\left. - n \left(\sum_{i=1}^{n} \varepsilon_{1i}^3 \right) \left(\sum_{i=1}^{n} \varepsilon_{1i}(y_{2i} - \boldsymbol{x_i}^{\mathrm{T}} \hat{\boldsymbol{\beta_2}}) \right) \right) .$$

$$\Rightarrow \hat{\alpha}_{20} = - \hat{\alpha}_{21} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{1i}^2.$$

Using Equation (6.1.3), we then obtain

$$\hat{Y}_2 | \boldsymbol{x_0}, Recursive = \boldsymbol{x_0}^{\mathrm{T}} \hat{\boldsymbol{\beta_2}} + \hat{\alpha}_{20} + \hat{\alpha}_{21} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{1i}^2$$

$$= \boldsymbol{x_0}^{\mathrm{T}} \hat{\boldsymbol{\beta_2}} - \hat{\alpha}_{21} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{1i}^2 + \hat{\alpha}_{21} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{1i}^2$$

$$= \boldsymbol{x_0}^{\mathrm{T}} \hat{\boldsymbol{\beta_2}},$$

and since, using Equation (6.1.2), this is identical to the conditional expectations from the multivariate model,

$$\hat{Y}_2 | \boldsymbol{x_0}, Recursive = \hat{Y}_2 | \boldsymbol{x_0}, Multi.$$

$\square$

Theorem 6.1.1 demonstrates that while non-centrality of error terms is induced by the ε_1^2 term in the sequential regressions, for the purpose of conditional expectations, in which estimated expected values of ε_1^2 are taken, this non-centrality is absorbed by the model's intercept.

6.1.2   Construction of prediction regions

While the conditional expectations under the multivariate regression model and of the recursive multivariate model are identical, prediction regions of the recursive model are not simply hyperellipsoids centred around these conditional expectations, as they are for the multivariate model (see Theorem 5.1.3). Rather, since error terms are designed to be related non-linearly, non-elliptical regions are produced.

In terms of the construction of intervals, due to the complex compounding of error terms in successive stages of the model, it is infeasible to derive analytic prediction intervals for the recursive model. Approximate $100(1 - \gamma)\%$ prediction regions for some new observation $\boldsymbol{x_0}$ can, however, be calculated by simulation in the following manner:

1. Simulate N observations from each of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$, gathered in the vector form $\{\boldsymbol{\varepsilon_i} : i = 1, 2, \ldots, N\}$.

2. Take the $100N(1-\gamma)\%$ observations for which $\sum_{j=1}^{k}\frac{\varepsilon_{ij}^2}{\sigma_j}$ is smallest, denoted $\{\boldsymbol{\varepsilon}_i^{accepted}\}$. This cutoff is chosen because it corresponds to the $100N(1-\gamma)\%$ observations with the largest densities; a sum of marginal Mahalanobis distances [115] can be used as the ε_j variables are independent.

3. The $100(1-\gamma)\%$ prediction region is an alpha shape around the accepted points $\{\boldsymbol{Y}|\boldsymbol{x_0}, \boldsymbol{\varepsilon}_i^{accepted}\}$.

An alpha shape is a generalisation of the notion of a complex hull [1, 54], and is used here to encompass the set of points $\{\boldsymbol{Y}|\boldsymbol{x_0}, \boldsymbol{\varepsilon}_i^{accepted}\}$ without forcing the prediction region to be convex. Allowing for non-convex prediction regions is necessary in this instance since, for example, a quadratic relationship between error terms should not necessitate the prediction region including the area between the tails of the parabola.

The performance of the alpha shape prediction region, in terms of sensitivity and power, for the recursive model relative to other candidate models, is discussed in Section 7.4.6.

6.1.3 Order of response variables and choice of relation functions

As has been discussed in this section, a key advantage of the recursive multivariate model is that it is able to capture more complex dependence structures between error terms by building models in successive response variables to incorporate the error structure of previous response variables. This allows for the capture of, say, a quadratic relationship between response variables Y_1 and Y_2 by including a ε_1^2 term in the model for Y_2.

However, it is not always immediately clear in which order the model's hierarchies should be constructed. In the above example, this might amount to a difficulty establishing whether to first construct a model for Y_1, and then include a ε_1^2 term in the model for Y_2, or construct a model for Y_2, and then include a ε_2^2 term in the model for Y_1. Certainly, it is possible to choose between these two models; bases for doing so, and methods for selection on these bases, are explained in Chapter 7. However, for k response variables, there are $k!$ possible orderings of the response variables, substantially complicating the model building process. Without expert or prior knowledge, there does not appear to be any obvious heuristic for reducing the number of candidate orderings, so model comparison might need to be undertaken between all $k!$ possible orderings.

A similar issue arises in the choice of candidate relation functions between the error terms, that is, the choice of function of the error in estimating Y_1 that occurs in Y_2, and so on. This is an arbitrary choice in that there is no restriction on the choice of functions, so long as its parameters are estimable. The following guidelines may be useful in restricting the space of relation functions:

1. Relation functions should adhere to the principle of parsimony. Arbitrarily complex functions may better fit the training data, but are less likely to be

generalisable [84].

2. Relation functions should be, in some way, explicable by reference to the domain in which the model is built. The further from the modelling domain relation functions stray, the less useful an explanatory model becomes [160].

3. Relation functions should be linear in basis expansions of previous error terms, so the functions' parameters can be estimated by ordinary least squares.

Despite these general guidelines, the somewhat arbitrary decisions relating both order of response variables and choice of relation functions present a drawback of the recursive model relative to a simpler model like the multivariate regression model. As such, the recursive multivariate model should not be used unless the simpler multivariate regression model proves inadequate. In the instance of this research, Section 5.2 demonstrated that the multivariate regression model is inadequate, so forms of the recursive multivariate model will be considered as candidates for explaining the complex error structure in estimating political beliefs from demographic variables.

The following section (Section 6.1.4) describes the application of the recursive model to the 2016 Australian Election Study data set. This includes feature selection, model diagnostics and assumption checking. This will be followed by *prima facie* assessment of the recursive model's suitability to describe the relationship between demographic and political identity, as expressed in the 2016 AES dataset.

6.1.4   Application of the recursive model to the Australian Election Study data

In applying the recursive model to the Australian Election Study data set, the number of predictors, and the number of response variables, needs to be selected. In Section 4.4, it was determined that the political spectrum comprises three key underlying variables, so these three variables will constitute the set of response variables. These three response variables, in descending order of relevance, are 'social inclusivity', 'attitudes to authority', and 'spending priorities'.

Feature selection, which entails choosing the appropriate set of predictors, is a somewhat more involved task. A reduced candidate set of predictors will be defined to be the set of all reduced demographic variables, which are associated with an eigenvalue greater than the mean eigenvalue of all reduced demographic variables; in other words, the set of all reduced demographic variables able to explain a greater proportion of the original data set's variance than is explained, on average, by a single level of the un-reduced data set. In Section 4.3, it was found that this is a set of 50 orthogonal variables. Handily for the feature selection process, these are orthogonal, so collinearity problems will not arise.

At each stage of the recursive model fit, stepwise feature selection will be undertaken, with a model sought to maximise the Bayesian Information Criterion ('BIC'). BIC is chosen as the selection criterion because it places a larger penalty

on additional predictors; a larger penalty is valuable due to the large number of candidate predictors, and the associated increased risk of overfitting.

The following paragraphs outline the outputs from recursively building a model in the following order:

1. A model of 'social inclusivity', denoted Y_1, using the set of reduced demographic variables $\boldsymbol{X} = [X_1, \ldots, X_{50}]$ as predictors; then

2. A model of 'attitudes to authority', denoted Y_2, using $\boldsymbol{X}$ as well as a quadratic function of the error in estimating Y_1, denoted ε_1, as predictors; and finally

3. A model of 'spending priorities', denoted Y_3, using $\boldsymbol{X}$ and quadratic functions of ε_1 and ε_2 as predictors, while observing the principle of marginality.

The first response variable Using both forward and backward stepwise selection, the following model was produced in the first response variable:

$$Y_1 = 2.79X_1 - 1.52X_2 + 4.62X_5 + 2.19X_7 - 1.40X_9$$
$$- 2.37X_{10} - 0.84X_{11} - 1.55X_{20} - 1.05X_{48} + \varepsilon_1,$$

where $\varepsilon_1 \sim N(0, 3.174^2)$.

Reinforcing the decision to undertake dimension reduction, the model in the first response variable heavily favours predictors with a lower index. The model contains six of the first 10 dimension-reduced demographic variables, but only three of the next 40, meaning that the variables most representative of demographic identity are also the best predictors of the first dimension of political identity. This is not always true when regressing on principal components (or some variation thereof) [100, 165], and represents that not a great deal of information was lost in undertaking dimension reduction and removing the least prominent components, since the information most relevant to political identity is contained in the most prominent components of demographic identity.

The second response variable A model in the second response variable was built using a set of candidate predictors comprising the first 50 dimension-reduced demographic variables, and a quadratic term in ε_1. Using both forward and backward stepwise selection, the following model was produced:

$$Y_2 = -0.30 - 3.07X_2 - 0.91X_3 - 1.91X_4 - 1.25X_5 - 0.87X_6$$
$$+ 1.41X_7 - 0.78X_8 - 0.85X_9 - 1.65X_{11} + 0.99X_{14}$$
$$+ 0.83X_{17} + 0.030\varepsilon_1^2 - 0.035\varepsilon_1 + \varepsilon_2,$$

where $\varepsilon_2 \sim N(0, 2.732^2)$. That the BIC-driven procedure selects the ε_1^2 feature goes some way to justify the selection of the recursive model; this demonstrates, as does Section 5.2, that a quadratic relationship exists between residuals in the Y_2 and Y_1 directions.

The third response variable   A model in the third response variable was built using a set of candidate predictors comprising the first 50 dimension-reduced demographic variables, and quadratic functions of ε_1 and ε_2. Using both forward and backward stepwise selection, the following model was produced:

$$
\begin{aligned}
Y_3 =& 0.11 - 2.02X_1 - 0.39X_2 - 0.44X_3 - 1.75X_4 + 0.57X_9 \\
& + 0.73X_{10} + 0.82X_{12} + 0.75X_{16} - 0.96X_{35} + 0.81X_{39} \\
& + 0.93X_{47} - 0.011\varepsilon_1^2 + 0.093\varepsilon_1 - 0.046\varepsilon_2 + \varepsilon_3,
\end{aligned}
$$

where $\varepsilon_3 \sim N(0, 2.215^2)$. Again, there is a quadratic relationship between the error in the Y_3 and the Y_1 direction. No such relationship reveals itself between the Y_3 and Y_2 directions.

Marginal model diagnostics

Looking at the marginal model in each political direction separately, each should uphold the assumptions of univariate regression (see Section 5.1.1). That is, the residuals should be normally distributed, and linearity and homoscedasticity should be observed in plots both of the residuals against the fitted values, and of the residuals against each predictor.

Normality   Normal quantile-quantile plots for the residuals, for each of the three stages of the recursive model, can be found in Figure 6.1. Since each exhibits a roughly linear trend, the assumption of normality appears justified for each marginal model.

Linearity and homoscedasticity   In Figures 6.2, 6.3 and 6.4 respectively, plots of residuals against each predictor, and against fitted values, are given for the first, second and third marginal univariate regression in the recursive model-building method. Since all plots appear to not exhibit any curvature, and appear to have constant spread, the assumptions of linearity and homoscedasticity appear justified.

   While some plots look to have unusual distributions in the x-direction, this does not invalidate the assumptions checked here. For example, Figures 6.3 and 6.4 contain plots of residuals against ε_1^2. Since ε_1^2 is a squared normal term, it has a chi-squared distribution, meaning the density of points is highest at zero, and decreases as ε_1^2 increases. As a result, there are fewer points in the plot for higher values of ε_1^2, which gives the appearance of heteroscedasticity as the points thin out. This does not mean that the assumption of homoscedasticity should be rejected; the variance does not substantially change between different values of ε_1^2, just the prevalence of points in the sample.

Joint model diagnostics

For simultaneous estimation of Y_1, Y_2 and Y_3, the assumption is made that the error terms in each of the three models are independent (see Section 6.1). This
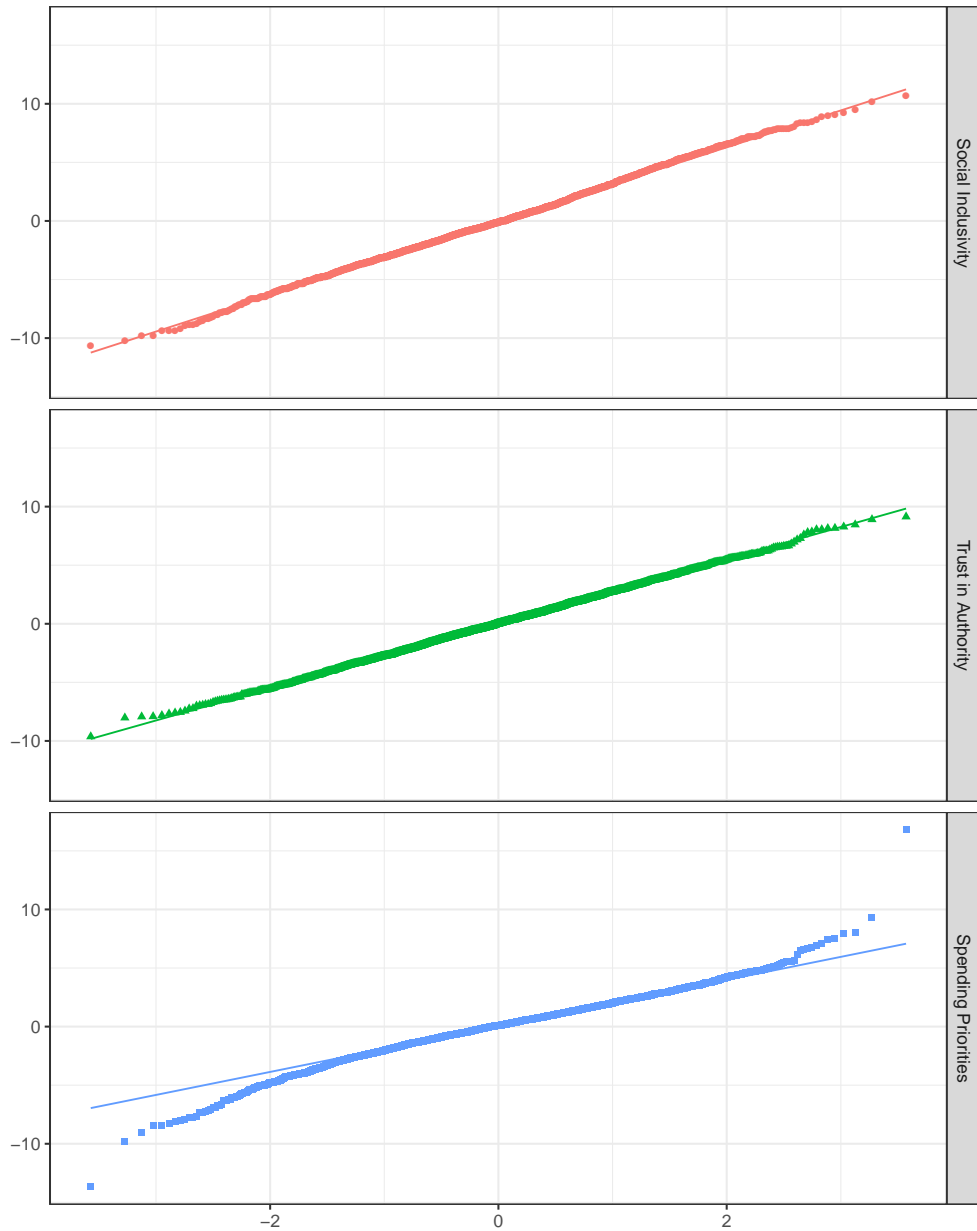
Figure 6.1: Normal quantile-quantile plots for the models in each of the first three response directions, for the recursive multivariate models. The assumption of normality appears justified in each. While the error term in the third direction appears slightly heavy-tailed, this is not sufficiently substantial to reject the assumption of marginal normality.
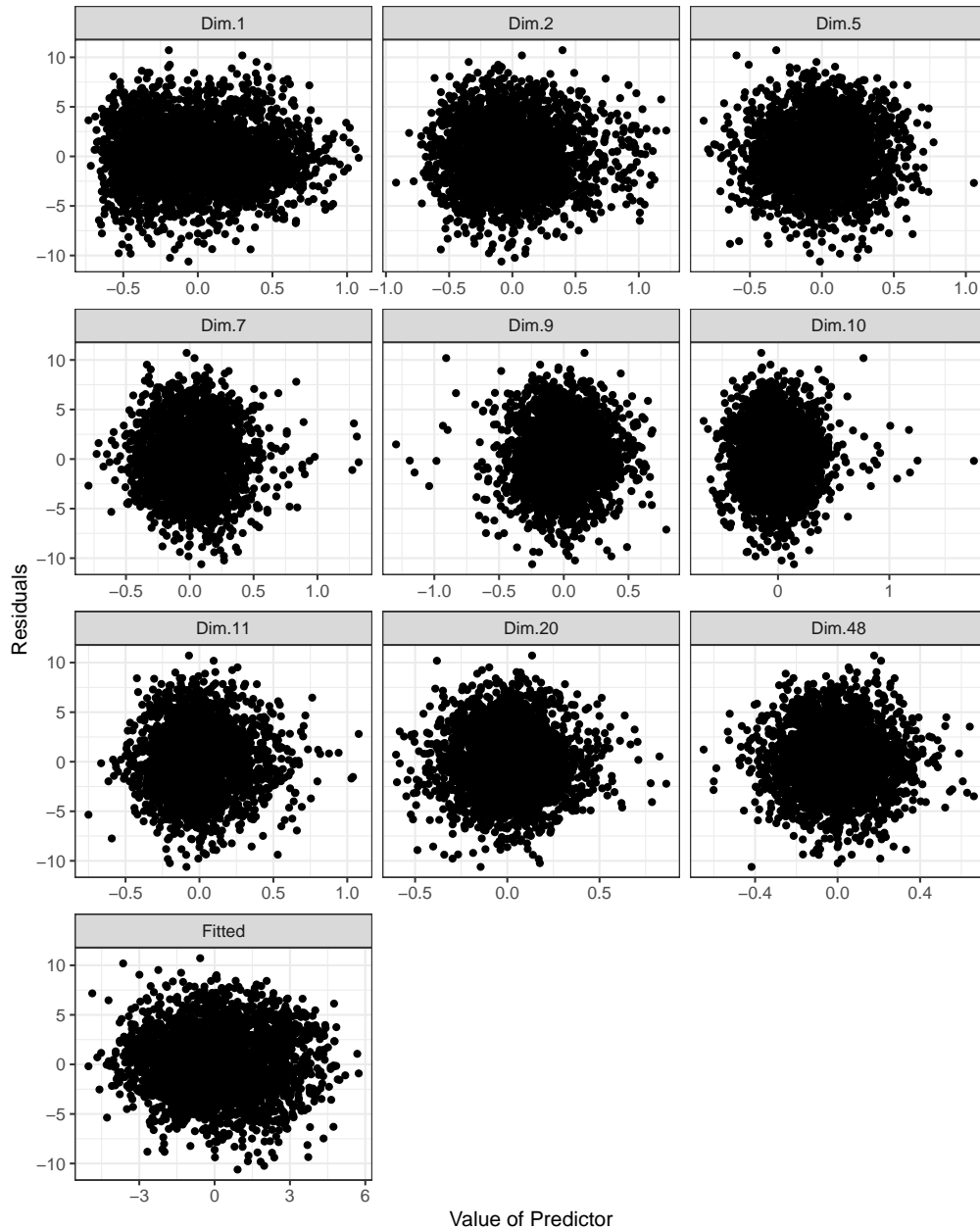
Figure 6.2: Plots of residuals against fitted, values, and against each included predictor, for Y_1, the first model in the three-stage recursive model building process. All plots look sufficiently linear, and with sufficiently constant spread, to not reject the assumptions of linearity and homoscedasticity respectively, as required by the marginal univariate regression model. While the plots often look to thin at the tails, this is due to less data in these regions, rather than heteroscadasticity.
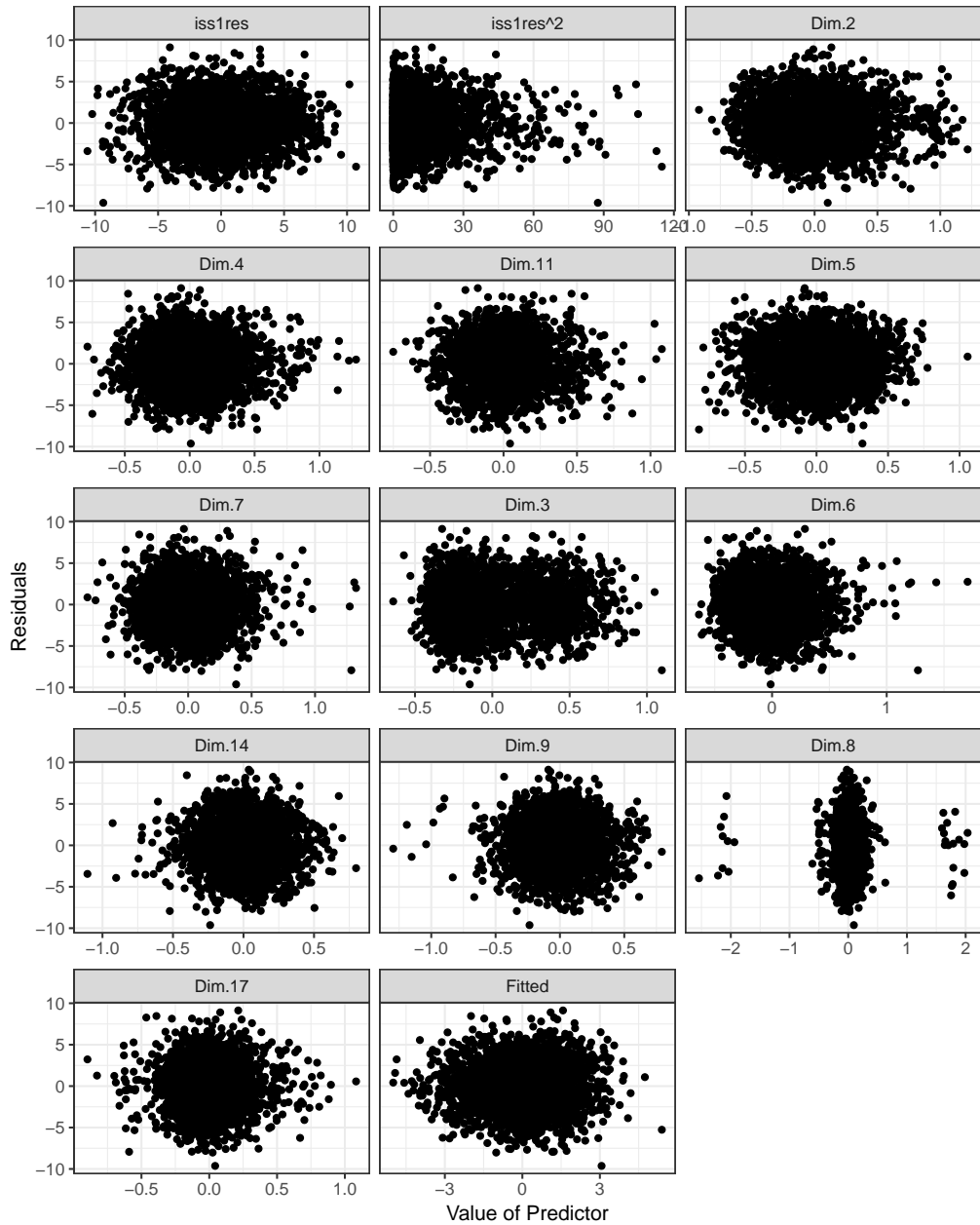
Figure 6.3: Plots of residuals against fitted, values, and against each included predictor, for Y_2, the second model in the three-stage recursive model building process. All plots look sufficiently linear, and with sufficiently constant spread, to not reject the assumptions of linearity and homoscedasticity respectively, as required by the marginal univariate regression model. The presense of lesser spread in the residuals against ε_1^2, moving from left to right, is due to fewer high values of this predictor, a consequence of squaring the normally distributed residuals in the first response direction. The unusual shape around the eighth demographic dimension is also a result of its unusual distribution, rather than heteroscedasticity.
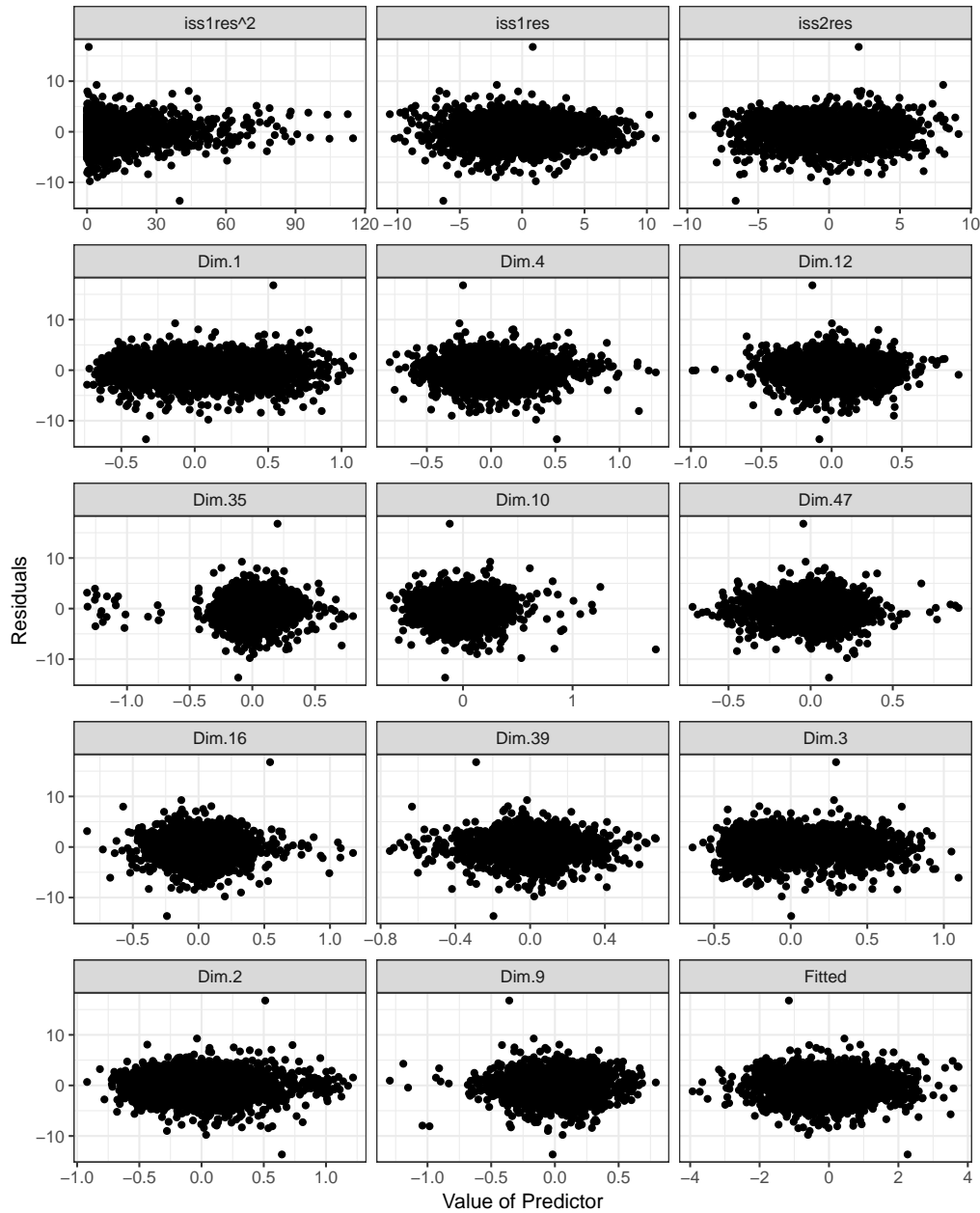
Figure 6.4: Plots of residuals against fitted, values, and against each included predictor, for Y_3, the third model in the three-stage recursive model building process. All plots look sufficiently linear, and with sufficiently constant spread, to not reject the assumptions of linearity and homoscedasticity respectively, as required by the marginal univariate regression model. The presense of lesser spread in the residuals against ε_1^2, moving from left to right, is due to fewer high values of this predictor, a consequence of squaring the normally distributed residuals in the first response direction. The unusual shape around the 35th demographic dimension is also a result of its unusual distribution, rather than heteroscedasticity.

would mean that $[\varepsilon_1, \varepsilon_2, \varepsilon_3] \sim N(\mathbf{0}, \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2))$. As explained in Section 5.1.2, whether these error terms form a joint multivariate normal distribution can be assessed using the energy test described by Székely and Rizzo [166]. The energy test uses the pairwise expected distances between the observed points and points from a normal distribution, to determine how dissimilar the observed values are to normally distributed data. To do this, the energy test involves synthesising data from the reference distribution, and a calculation of the difference between mean distances within each distribution, and mean distances between each distribution (see Section 5.1.2 for more details). This energy test statistic can be compared to theoretical values obtained from synthetic data from the reference distribution, to produce a non-parametric p-value. Using this p-value, we can assess the hypothesis that the residuals are deviates from the desired normal distribution.

The test was applied to the residuals in the 2016 AES data set, using 100 sets of simulated data from the $N(\mathbf{0}, \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2))$ for comparison, and 500 replicates for the generation of each non-parametric p-value. The synthetic data sets are of the same size of the Australian Election Study ($n = 2818$), and this will be held constant in comparing multiple models, since the energy test statistic is sensitive to the size of synthetic samples. The mean p-value produced was 0.172. This indicates that there is insufficient evidence at the 5% significance level to reject the assumption that the recursive model leads to a independent error terms.

As such, the assumptions of the recursive model appear justified in the AES 2016 instance. After other candidate models have also been assessed in Sections 6.2 and 6.3, the models will be compared in Section 7.4.6 to determine if the recursive model should be favoured over other alternatives.

6.2 A mixture model on residuals

As established in Section 5.2, a major drawback of the multivariate regression model is the assertion that the model's error can be explained by a single multivariate normal distribution. This means the model is unable to capture any tendencies of error terms to co-depend upon one another non-linearly, to cluster together, or to exhibit non-normal scedasticity. A natural extension to the multivariate regression model might be, then, to replace the multivariate normal error term with a more complex multivariate distribution. A simple candidate, which is able to exhibit clustering, non-linear co-dependence, and non-normal scedasticity, is a Gaussian mixture. A Gaussian mixture, being a weighted sum of multivariate normal distributions, can explain clusters of residuals. The mixture of normal distributions can also relate complicated relationships within clusters of points, akin to using piecewise functions to account for a non-differentiable structure.

With p predictor variables and an intercept, $\boldsymbol{X} = [1, X_1, \ldots, X_p]^{\mathrm{T}}$, and k response variables $\boldsymbol{Y} = [Y_1, \ldots, Y_k]^{\mathrm{T}}$, the mixture model can be expressed in the form

$$\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}}\beta + \boldsymbol{\varepsilon}_{\boldsymbol{mix}}^{\mathrm{T}},$$

where β is a $(p+1) \times k$ matrix of coefficients, $\boldsymbol{\varepsilon_{mix}} \sim \sum_{m=1}^{M} \pi_m N(\boldsymbol{\mu}_m, \Sigma_m)$, and M is the number of components of the Gaussian mixture. The proportions π_m for $m = 1, \ldots, M$ are constrained such that $\sum_{m=1}^{M} \pi_m = 1$. The coefficient matrix β is identical to that in the multivariate regression model. It is assumed that the p predictor variables $\boldsymbol{X}$ are independent of the error term $\boldsymbol{\varepsilon_{mix}}$.

6.2.1   Fitting the model

Firstly, the estimate $\hat{\beta}$ of the coefficient matrix β is constructed using ordinary least squares, that is, such that it satisfies

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \operatorname{tr} \left((Y - X\beta)^{\mathrm{T}} (Y - X\beta) \right),$$

where X is an $n \times (p+1)$ matrix of observed predictors with an intercept column, and Y is an $n \times k$ matrix of observed responses (see Section 5.1.2 for more details on this notation).

The parameters of the Gaussian mixture term $\boldsymbol{\varepsilon_{mix}}$ then need to be estimated. The Expectation Maximisation (EM) algorithm is chosen for parameter estimation due to its simplicity and its ubiquity [183]. To estimate the set of parameters, denoted $\boldsymbol{\theta} := \{\boldsymbol{\mu}_m, \Sigma_m, \pi_m : i = 1, 2, \ldots, M\}$, using the EM algorithm, we augment the known data $\{\boldsymbol{\varepsilon_i} : i = 1, 2, \ldots, n\}$ with component indicators

$$R_{im} = \begin{cases} 1 & \text{if subject } i \text{ is in component } m \\ 0 & \text{otherwise,} \end{cases}$$

for $m = 1, 2, \ldots, M$. The R_{im}s are treated as missing data.

E-Step

During the lth iteration, we must calculate the expected full log-likelihood, given the observed data and current parameter estimates $\boldsymbol{\theta}^{(l)}$. The log-likelihood of a Gaussian mixture is easily derived as a product of normal distributions, with component indicator:

$$l(\boldsymbol{\theta}|\{\boldsymbol{\varepsilon_i}, R_{im}\}) = \log \left[\prod_{i=1}^{n} \prod_{m=1}^{M} \pi_m^{R_{im}} \left(\frac{1}{\sqrt{2\pi\sigma_m^2}} \exp \left(-\frac{1}{2\sigma_m^2} (\boldsymbol{\varepsilon_i} - \mu_m)^2 \right) \right)^{R_{im}} \right]$$

$$= \sum_{i=1}^{n} \sum_{m=1}^{M} R_{im} \left(\log \pi_m - \frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_m^2 - \frac{1}{2\sigma_m^2} (\boldsymbol{\varepsilon_i} - \mu_m)^2 \right).$$

Since l is linear in R_{im}, $E[l(\boldsymbol{\theta}|\{\boldsymbol{\varepsilon_i}, R_{im}\})|\{\boldsymbol{\varepsilon_i}\}] =$

$$\sum_{i=1}^{n} \sum_{m=1}^{M} E[R_{im}|\{\boldsymbol{\varepsilon_i}\}] \left(\log \pi_m - \frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_m^2 - \frac{1}{2\sigma_m^2} (\boldsymbol{\varepsilon_i} - \mu_m)^2 \right). \quad (6.2.1)$$

We can also define $\gamma_{im} := E[R_{im}|\{\boldsymbol{\varepsilon_i}\}]$. Since the observations of $\boldsymbol{\varepsilon_i}$ are independent,

$$
\begin{aligned}
\gamma_{im} &= E[R_{im}|\boldsymbol{\varepsilon_i}] \\
&= P(R_{im} = 1|\boldsymbol{\varepsilon_i}) \\
&= \frac{f(R_{im} = 1, \boldsymbol{\varepsilon_i})}{\sum_{j=1}^{K} f(R_{ij} = 1, \boldsymbol{\varepsilon_i})} \\
&= \frac{f(\boldsymbol{\varepsilon_i}|R_{im} = 1)P(R_{im} = 1)}{\sum_{j=1}^{M} f(\boldsymbol{\varepsilon_i}|R_{ij} = 1)P(R_{ij} = 1)} \\
&= \frac{\pi_m f_m(\boldsymbol{\varepsilon_i})}{\sum_{j=1}^{M} \pi_j f_j(\boldsymbol{\varepsilon_i})},
\end{aligned}
\tag{6.2.2}
$$

for $f_m(\boldsymbol{\varepsilon_i})$ the density of the mth normally distributed component of the mixture.

M-Step

Given the estimates at the $(l-1)$th iteration $\mu_m^{(l-1)}, \sigma_m^{2\,(l-1)}$ and $\pi_m^{(l-1)}$, and using these to compute $\gamma_{im}^{(l-1)}$, we seek the lth estimates

- $\pi_m^{(l)} = \operatorname{argmax}_{\pi_m} E\left[l(\{\pi_m, \mu_m^{(l-1)}, \sigma_m^{2\,(l-1)}|\boldsymbol{\varepsilon_i}, \gamma_{im}^{(l-1)}\})|\{\boldsymbol{\varepsilon_i}\}\right]$, subject to $\sum_{m=1}^{M} \pi_m^{(l)} = 1$;

- $\mu_m^{(l)} = \operatorname{argmax}_{\mu_m} E\left[l(\{\pi_m^{(l-1)}, \mu_m, \sigma_m^{2\,(l-1)}|\boldsymbol{\varepsilon_i}, \gamma_{im}^{(l-1)}\})|\{\boldsymbol{\varepsilon_i}\}\right]$; and

- $\sigma_m^{2\,(l)} = \operatorname{argmax}_{\mu_m} E\left[l(\{\pi_m^{(l-1)}, \mu_m^{(l-1)}, \sigma_m^2|\boldsymbol{\varepsilon_i}, \gamma_{im}^{(l-1)}\})|\{\boldsymbol{\varepsilon_i}\}\right]$,

for each $m = 1, \ldots, M$. These parameter updates are derived as follows:

Deriving $\pi_m^{(l)}$:   Since we are maximising the expectation subject to the constraint $\sum_{m=1}^{M} \pi_m^{(l)} = 1$, we make use of a Lagrange multiplier. Using Equation (6.2.1), the Lagrangian is

$$
\begin{aligned}
L(\{\pi_m\}) = \sum_{i=1}^{n}\sum_{m=1}^{M} \gamma_{im}^{(l-1)} &\left(\log \pi_m - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma_m^{2\,(l-1)} \right. \\
&\left. - \frac{1}{2\sigma_m^{2\,(l-1)}}(\boldsymbol{\varepsilon_i} - \mu_m^{(l-1)})^2 \right) + \lambda\left(\sum_{m=1}^{M} \pi_m^{(l)} - 1 \right).
\end{aligned}
$$

$$
\Rightarrow \frac{d}{d\pi_m} L(\{\pi_m\}) = \sum_{i=1}^{n} \gamma_{im}^{(l-1)}\frac{1}{\pi_m} + \lambda.
\tag{6.2.3}
$$

Setting the derivative equal to zero leads to

$$
\pi_m = -\frac{1}{\lambda}\sum_{i=1}^{n} \gamma_{im}^{(l-1)}.
$$

Now since the sum of the π_ms is one,

$$1 = -\frac{1}{\lambda} \sum_{m=1}^{M} \sum_{i=1}^{n} \gamma_{im}^{(l-1)},$$

and since similarly, Equation (6.2.2) implies the sum over m of the γ_{im}s is 1,

$$\lambda = -\sum_{i=1}^{n} 1$$

$$= -n.$$

From Equation (6.2.3), this thus yields $\pi_m^{(l)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{im}^{(l-1)}.$

Deriving $\mu_m^{(l)}$:   We seek when the derivative of Equation (6.2.1) is zero:

$$E\left[l(\{\pi_m^{(l-1)}, \mu_m, \sigma_m^{2\,(l-1)} | \boldsymbol{\varepsilon_i}, \gamma_{im}^{(l-1)}\}) | \{\boldsymbol{\varepsilon_i}\} \right]$$

$$= \sum_{i=1}^{n} \sum_{m=1}^{M} \gamma_{im}^{(l-1)} \left(\log \pi_m^{(l-1)} - \frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_m^{2\,(l-1)} - \frac{1}{2\sigma_m^{2\,(l-1)}} (\boldsymbol{\varepsilon_i} - \mu_m)^2 \right).$$

Taking the derivative,

$$\frac{d}{d\mu_m} E\left[l(\{\pi_m^{(l-1)}, \mu_m, \sigma_m^{2\,(l-1)} | \boldsymbol{\varepsilon_i}, \gamma_{im}^{(l-1)}\}) | \{\boldsymbol{\varepsilon_i}\} \right] = \sum_{i=1}^{n} \gamma_{im}^{(l-1)} \left(\frac{1}{\sigma_m^{2\,(l-1)}} (\boldsymbol{\varepsilon_i} - \mu_m) \right),$$

which is equal to zero when

$$\sum_{i=1}^{n} \gamma_{im}^{(l-1)} (\boldsymbol{\varepsilon_i} - \mu_m) = 0$$

$$\sum_{i=1}^{n} \gamma_{im}^{(l-1)} \mu_m = \sum_{i=1}^{n} \gamma_{im}^{(l-1)} \boldsymbol{\varepsilon_i}$$

$$\Rightarrow \mu_m^{(l)} = \frac{\sum_{i=1}^{n} \gamma_{im}^{(l-1)} \boldsymbol{\varepsilon_i}}{\sum_{i=1}^{n} \gamma_{im}^{(l-1)}}.$$

Deriving $\sigma_m^{2\,(l)}$:   We seek when the derivative of Equation (6.2.1) is zero:

$$E\left[l(\{\pi_m^{(l-1)}, \mu_m^{(l-1)}, \sigma_m^2; | \boldsymbol{\varepsilon_i}, \gamma_{im}^{(l-1)}\}) | \{\boldsymbol{\varepsilon_i}\} \right]$$

$$= \sum_{i=1}^{n} \sum_{m=1}^{M} \gamma_{im}^{(l-1)} \left(\log \pi_m^{(l-1)} - \frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_m^2 - \frac{1}{2\sigma_m^2} (\boldsymbol{\varepsilon_i} - \mu_m^{(l-1)})^2 \right).$$

Taking the derivative,

$$\frac{d}{d\sigma_m^2} E\left[l(\{\pi_m^{(l-1)}, \mu_m, \sigma_m^{2\,(l-1)} | \varepsilon_i, \gamma_{im}^{(l-1)}\}) | \{\varepsilon_i\}\right] = \sum_{i=1}^n \gamma_{im}^{(l-1)} \left(-\frac{1}{2\sigma_m^2} + \frac{1}{2(\sigma_m^2)^2}(\varepsilon_i - \mu_m^{(l-1)})^2\right),$$

which is equal to zero when

$$\sum_{i=1}^n \gamma_{im}^{(l-1)} \left(-\frac{1}{2\sigma_m^2} + \frac{1}{2(\sigma_m^2)^2}(\varepsilon_i - \mu_m^{(l-1)})^2\right) = 0$$

$$\Rightarrow \sum_{i=1}^n \gamma_{im}^{(l-1)} \frac{1}{2\sigma_m^2} = \sum_{i=1}^n \gamma_{im}^{(l-1)} \frac{1}{2(\sigma_m^2)^2}(\varepsilon_i - \mu_m^{(l-1)})^2$$

$$\Rightarrow \sigma_m^2 \sum_{i=1}^n \gamma_{im}^{(l-1)} = \sum_{i=1}^n \gamma_{im}^{(l-1)}(\varepsilon_i - \mu_m^{(l-1)})^2$$

$$\Rightarrow \sigma_m^{2\,(l)} = \frac{\sum_{i=1}^n \gamma_{im}^{(l-1)}(\varepsilon_i - \mu_m^{(l-1)})^2}{\sum_{i=1}^n \gamma_{im}^{(l-1)}}.$$

Obtaining $\gamma_{im}^{(l)}$: After the above are calculated, new values of γ_{im} are obtained using Equation (6.2.2):

$$\gamma_{im}^{(l)} = \frac{\pi_m^{(l)} f_m(\varepsilon_i)}{\sum_{j=1}^M \pi_j^{(l)} f_j(\varepsilon_i)}.$$

We thus have parameter updates for π_m, μ_m, σ_m^2 and γ_{im}. By repeating the above E- and M-steps until some convergence criteria have been met, the parameters for the Gaussian mixture model are obtained.

6.2.2 Construction of prediction regions

A difficulty in calculating $100(1 - \gamma)\%$ prediction regions for the Gaussian mixture model arises as a result of the overlapping densities of components, meaning analytically selecting the smallest region whose cumulative density exceeds $100(1 - \gamma)\%$ is much more difficult than it appears at first glance. As such, as in the recursive multivariate model case, prediction regions are produced using simulation.

The method for simulating prediction regions in the mixture model case differs somewhat from the recursive multivariate case. In both, the goal is to take some proportion of simulated observations whose density is greatest. While in the recursive multivariate case, this could be done by comparing the Mahalanobis distances of residuals, here this must be calculated more directly. Approximate $100(1 - \gamma)\%$ prediction regions for some new observation x_0 can, be calculated by simulation in the following manner:

1. Simulate N observations from ε_{mix}, denoted $\{\varepsilon_i : i = 1, 2, \ldots, N\}$. Also indicate which component of the mixture each component comes from, denoted $\{R_{im} : i = 1, 2, \ldots, N, m = 1, 2, \ldots, M\}$.

2. Take the $100N(1-\gamma)\%$ observations for which $F(\boldsymbol{\varepsilon_i})$ is greatest, where $F()$ is the density function of the Gaussian mixture. Denote this $\{\boldsymbol{\varepsilon_i^{accepted}}\}$. Record also $\{R_{im}^{accepted}\}$.

3. For each component $m = 1, 2, \ldots, M$, construct an alpha shape [1, 54] around all points $\{\boldsymbol{\varepsilon_i^{accepted}} : R_{im}^{accepted} = 1\}$. Denote these regions P_m: $m = 1, 2, \ldots, M$.

4. The prediction region is $\displaystyle\bigcup_{m=1}^{M} P_m$.

In the above method, separate regions are constructed for each component of the Gaussian mixture. This is done to allow for disjoint clusters to be produced, if necessary.

6.2.3   Choosing the number of components

The choice of the number of components, M for the Gaussian mixture model is not a straightforward one. The addition of more components will never decrease the likelihood of observations, biasing a purely maximum-likelihood based approach like expectation maximisation in favour of larger M.

This means that a bias-variance trade off emerges; as M increases, each component's mean and variance is dependent upon a smaller total weight of observed points, making parameter estimates more variable. Meanwhile, bias emerges at smaller M than is necessary, since a small-component mixture model cannot fully approximate a larger-component Gaussian mixture. As such, some component-selection approach that seeks to reduce both bias and variance should be considered.

In the more restrictive case of a k-means clustering model, rather than a Gaussian mixture, Hastie, Tibshirani and Friedman [84] suggest a visual method to balance bias and variance in the selection of M: plotting likelihood against the number of components, and select the number after which likelihood increases substantially decline. This is described as looking for a 'kink' in the plot [84]. While the visual approach could easily also be used for a Gaussian mixture model, this is a highly subjective technique; reasonable people may differ in deciding at what point in a plot a 'kink' occurs.

A numerical alternative is suggested—measuring likelihood, penalised against the number of components. For example, since an increased number of components is associated with an increase in the number of parameters, AIC or BIC can be used [110]. This approach is implemented in the R package mclust [158] for both k-means clustering and Gaussian mixture models.

Another alternative approach is the use of cross-validated density, or prediction, regions. This entails the creation of prediction regions as in Section 6.2.2 from different training sets within the observed data, and the selection of the number of components for which the coverage of the prediction regions, as applied to different test sets within the observed data, is most accurate. Accuracy of prediction regions is discussed in more detail in Section 7.2.1.

In this work, the choice of M is undertaken using the cross-validation approach, as a result of a focus here on a model's ability to maintain accurate prediction intervals.

6.2.4 Application of the mixture model to the Australian Election Study data

In applying the mixture model to the Australian Election Study data set, the set of predictors, the number of response variables, and the number of mixture components, needs to be selected. In Section 4.4, it was determined that the political spectrum comprises three key underlying variables, so these three variables will constitute the set of response variables. These three response variables, in descending order of relevance, are 'social inclusivity', 'attitudes to authority', and 'spending priorities'.

The set of predictors will be chosen as they would have been done for multivariate regression; that is, by choosing the set that minimises the Bayesian Information Criterion ('BIC') for the multivariate regression model. As for the recursive model, BIC is the chosen goodness-of-fit metric in feature selection due to the large number of candidate predictors, and the desire for a more parsimonious model; BIC places a larger penalty on the addition of predictors than does competing criteria like Akaike's Information Criterion ('AIC').

Using both forward and backward BIC-driven step-wise selection, the following model was estimated:

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} -2.79 & 1.52 & 0.27 & 0.31 & 4.62 & 2.19 & -1.39 & -2.37 & -0.84 & -1.55 \\ 0.31 & 3.03 & -0.93 & -1.82 & -1.21 & 1.58 & -0.98 & -0.27 & -1.66 & -0.03 \\ -2.03 & -0.40 & 0.41 & 1.76 & 0.03 & 0.19 & -0.62 & -0.78 & 0.40 & -0.44 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_7 \\ X_9 \\ X_{10} \\ X_{11} \\ X_{20} \end{bmatrix} + \boldsymbol{\varepsilon_{mix}}.
$$

These parameters are the same as that in a multivariate model with the same predictor set, and will be the same as in the application of the kernel density model introduced in Section 6.3. It is notable that the stepwise selection procedure does not include an intercept term, since the mean of all response variables, and all predictors, is zero.

The parameters of $\boldsymbol{\varepsilon_{mix}}$ must also be estimated. This involves fitting the parameters of a Gaussian mixture model on the residuals, for various numbers of mixture components. The model chosen will be the one with a number of components that produces the most accurate prediction regions as measured by cross-validation (see Section 7.2.1). Essentially, a model is sought whose prediction

regions, for some set of new observations, contain the proportion of observations they purport to contain. If prediction regions are inaccurate, a model will either overstate or understate the level of confidence in its estimates. In this thesis, we use a k-fold cross validation procedure for assessing accuracy of $100(1-\alpha)\%$ prediction regions. This procedure can be expressed as follows:

1. Randomly partition the n residuals into k groups.

2. For $i = 1, 2, \ldots, k$:

   - Fit the mixture model to all but the ith partition.

   - Using the fitted model, produce a $100(1-\alpha)\%$ prediction region for each point in the ith partition.

   - Calculate the proportion of observations in the ith partition that fall outside their respective prediction regions. Denote this proportion $p_{\alpha,i}$.

3. Calculate the mean proportion of observations not covered by the prediction regions, $\overline{p_\alpha} := \frac{1}{k}\sum_{i=1}^{k} p_{\alpha,i}$.

4. The accuracy of the prediction regions can be measured as the absolute difference between the purported and the estimated empirical coverage of the regions, in other words, $|\alpha - \overline{p_\alpha}|$.

The accuracies $|\alpha - \overline{p_\alpha}|$ can be calculated and compared for all numbers of components, and for a variety of relevant values of α, and the smallest number of components, which does not substantially under-perform in accuracy relative to the other candidates, should be preferred.

Using this process, the number of components selected is 3, as the addition of extra components does not substantially improve the performance of prediction regions. Using the Expectation Maximisation procedure outlined in Section 6.2.1, the following parameters were estimated for the Gaussian mixture model with 3 components:

$$\boldsymbol{\varepsilon_{mix}} \sim \sum_{m=1}^{M} \pi_m N(\boldsymbol{\mu}_m, \Sigma_m), \text{ where}$$

$$\pi_1 = 0.382;$$
$$\pi_2 = 0.209;$$
$$\pi_3 = 0.409;$$
$$\boldsymbol{\mu}_1 = \begin{bmatrix} -1.096 \\ -0.643 \\ -0.145 \end{bmatrix};$$
$$\boldsymbol{\mu}_2 = \begin{bmatrix} -1.908 \\ 0.561 \\ 0.468 \end{bmatrix};$$

$$\boldsymbol{\mu}_3 = \begin{bmatrix} 2.000 \\ 0.315 \\ -0.104 \end{bmatrix} ;$$

$$\Sigma_1 = \begin{bmatrix} 5.096 & -0.301 & -0.827 \\ -0.301 & 7.212 & -1.227 \\ -0.827 & -1.227 & 3.687 \end{bmatrix} ;$$

$$\Sigma_2 = \begin{bmatrix} 9.464 & -3.980 & -2.555 \\ -3.980 & 10.523 & 3.053 \\ -2.555 & 3.053 & 12.192 \end{bmatrix} ; \text{ and}$$

$$\Sigma_3 = \begin{bmatrix} 8.008 & 0.907 & -0.550 \\ 0.907 & 6.196 & 0.828 \\ -0.550 & 0.828 & 2.670 \end{bmatrix} .$$

Model diagnostics

For the Gaussian mixture model on the residuals to be valid, the residual distribution $\boldsymbol{\varepsilon_{mix}} \sim \sum_{m=1}^{M} \pi_m N(\boldsymbol{\mu}_m, \Sigma_m)$ should closely approximate the distribution of the residuals. As was done in Section 6.1.4, the energy test for closeness of observed deviates to a proposed distribution of $\boldsymbol{\varepsilon_{mix}}$ was undertaken. A great advantage of this test is its flexibility; it can be applied to any proposed distribution. Here, as in testing the validity of the multivariate normal distribution for the recursive model, the test was applied to the residuals in the 2016 AES data set, using 100 sets of simulated data from the estimated distribution of $\boldsymbol{\varepsilon_{mix}}$ for comparison, and 500 replicates for the generation of each non-parametric p-value. The synthetic data sets are of the same size of the Australian Election Study ($n = 2818$). The mean p-value in testing the hypothesis that the data come from the proposed distribution of $\boldsymbol{\varepsilon_{mix}}$ was 0.675, indicating that there is insufficient evidence at the 5% significance level to suggest that the residuals do not belong to the Gaussian mixture model.

As such, the assumptions of the mixture model appear justified in the AES 2016 instance. Candidate models, including this one, will be compared in Section 7.4.6 to determine if the mixture model should be favoured over other alternatives.

6.3   Kernel density estimation

A third and final alternative to the multivariate regression model is centred on the most directly data-driven approach to responding to relationships between residuals—building a distribution from a weighted sum of small distributions around each of the residuals. The kernel density estimate model attempts to respond to the peculiarities of data it seeks to explain by responding to each data point separately. In doing so, provided the imputed distributions about each point have sufficient variance, a set of observed points is smoothed into a distribution able to capture their shape, while also capturing the unobserved probability density in the space between the points.

For the purposes of this work, a Gaussian kernel is used for the distributions about each observed point. While other kernels can be used, including the Epanechnikov and tri-cube kernels [84], the Gaussian kernel is used here due to the ease of making calculations on the basis of the normal distribution and the ease of its extension to the multivariate case.

With p predictor variables and an intercept, $\boldsymbol{X} = [1, X_1, \ldots, X_p]^{\mathrm{T}}$, and k response variables $\boldsymbol{Y} = [Y_1, \ldots, Y_k]^{\mathrm{T}}$, the kernel density model can be expressed in the form

$$\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}}\beta + \boldsymbol{\varepsilon}_{\boldsymbol{kde}}^{\mathrm{T}},$$

where β is a $(p+1) \times k$ matrix of coefficients, and $\boldsymbol{\varepsilon}_{\boldsymbol{kde}} \sim \sum_{i=1}^{n} \frac{1}{n} N(\boldsymbol{\varepsilon}_i, \Sigma_{kde})$, the kernel density estimate being calculated using n data points. $\boldsymbol{\varepsilon}_{\boldsymbol{i}}$ is the ith observed residual from ordinary least squares estimation of β, and Σ_{kde} is some symmetric $p \times p$ variance matrix independent of i; Σ_{kde} is called the 'bandwidth'.

It should be noted that while they look similar, the mixture model is not nested within the kernel density estimate model. A mixture model with $M = n$ components, as many as are in the kernel density estimate model, would have likelihood maximised at variance $\boldsymbol{\Sigma_m} = 0 \; \forall \; m$, while $\Sigma_{kde} \neq 0$.

6.3.1   Choice of bandwidth

The choice of 'bandwidth' in kernel density estimation is a widely-researched topic [80, 92, 174, 175]. In general, the aim of bandwidth selection is to balance the twin goals of producing a smooth distributional estimate, and of fidelity to the data producing the kernel density estimates. Hastie, Tibshirani and Friedman (2009) frame this in terms of the bias-variance trade-off [84]. This approach is favoured by both the 'plug-in' and the 'smoothed cross validation' bandwidth estimators, described below.

To balance bias and variance, bandwidth estimation procedures are usually designed to minimise the mean integrated squared error ('MISE'), a function similar to the mean squared error, but integrated across all possible observations, rather than just seeking to minimise error at the observed points $f(x)$. MISE is expressed as

$$MISE(\Sigma_{kde}) = E\left[\int_{-\infty}^{\infty} (\hat{f}_{\Sigma_{kde}}(\boldsymbol{\varepsilon}_{\boldsymbol{kde}}) - f(\boldsymbol{\varepsilon}_{\boldsymbol{kde}}))^2 d\boldsymbol{\varepsilon}_{\boldsymbol{kde}}\right].$$

In other words, the squared error of the kernel density estimate with certain bandwidth Σ_{kde} is integrated across all possible values of the error $\boldsymbol{\varepsilon}_{\boldsymbol{kde}}$. However, the true function $f(\boldsymbol{\varepsilon}_{\boldsymbol{kde}})$ is unknown, and the integral is not easy to calculate, so approximations are made by both the 'plug-in' and 'smoothed cross validation' bandwidth estimators to accommodate this. Details of their derivation can be found in [174] and [80] respectively, and both are implemented in the R package ks[53].

While the plug-in estimator focuses on reducing the error in estimating $f(\boldsymbol{\varepsilon}_{\boldsymbol{kde}})$ in small-sample situations [174], the smoothed cross-validation estimator estimates $f(\boldsymbol{\varepsilon}_{\boldsymbol{kde}})$ based upon leave-one-out cross-validation [80], to reduce a tendency for overconfidence in the data. Here, due to the relatively large sample of the 2016 Australian Election Study, the smoothed cross-validation estimator is used.

6.3.2   Relative complexity of the kernel density model

The kernel density model, with a normal distribution placed at every observed point, has $nk + \frac{k(k+1)}{2}$ parameters for k response variables, since n k-dimensional means are estimated, and one $k \times k$ variance matrix Σ_{kde}—the symmetric $(k \times k)$ common covariance matrix has $\frac{k(k+1)}{2}$ unique parameters. By comparison, the recursive multivariate model, as expressed in Equation (6.1.1), requires the estimation of $2k - 1$ parameters at each recursion, meaning just k^2 parameters are required. The mixture model requires the estimation of M k-dimensional means and M $k \times k$ variance matrices, meaning it has $Mk + M\frac{k(k+1)}{2}$ parameters. Clearly, the kernel density model requires far more parameters to be estimated than for alternative candidate models, provided $n >> k$, $n >> M$. As a result, if the kernel density model performs similarly to these more parsimonious competitors, the competitors should be preferred.

Due to its complexity, it should be said that the kernel density model, unlike its competitors, does not attempt to directly explain the domain-level scenario the model seeks to describe. For example, in the political modelling example, the existence of a quadratic term in the recursive model indicates that perhaps some quadratic relationship exists between error in estimating one aspect of political identity, and the error in estimating another. By contrast, the kernel density model does not suggest that the error in estimating political opinion from available demographic information is drawn from a mixture of some $n = 2818$ normal distributions. The usefulness of the kernel density model is in combining these normal distributions to produce a smooth representation of the domain-level scenario it models, while still responding to any unusual variation within the data. In this sense the kernel density model is regarded as a predictive model, rather than an explanatory model (see this distinction in Shmeuli (2010) [160]).

6.3.3   Application of the kernel density model to the Australian Election Study data

Application of the kernel density model to the Australian Election Study data set comprises two stages: fitting conditional expectations using least squares, and computing the kernel density estimate. The conditional expectations are identical to those in the mixture model case, since both rely on fitting *ad hoc* distributions to the residuals from a least squares linear model. Thus (as in Section 6.2.4), with response variables labelled such that

- Y_1 refers to 'social inclusivity';

- Y_2 refers to 'attitudes to authority'; and

- Y_2 refers to 'spending priorities',

the following model was formulated:

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} -2.79 & 1.52 & 0.27 & 0.31 & 4.62 & 2.19 & -1.39 & -2.37 & -0.84 & -1.55 \\ 0.31 & 3.03 & -0.93 & -1.82 & -1.21 & 1.58 & -0.98 & -0.27 & -1.66 & -0.03 \\ -2.03 & -0.40 & 0.41 & 1.76 & 0.03 & 0.19 & -0.62 & -0.78 & 0.40 & -0.44 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_7 \\ X_9 \\ X_{10} \\ X_{11} \\ X_{20} \end{bmatrix} + \varepsilon_{kde},
$$

for X_j the jth dimension-reduced demographic variable, and error term ε_{kde}.

The next stage consists of fitting the kernel density estimate model to the residuals. This model is in the form $\varepsilon_{kde} \sim \sum_{i=1}^{n} \frac{1}{n} N(\varepsilon_i, \Sigma_{kde})$,, where ε_i is the ith residual from the ordinary least squares procedure. With the ε_i's known, Σ_{kde} is chosen using the smoothed cross-validation bandwidth estimator outlined in Section 6.3.1 and implemented in the R package ks[53]. Using this package, we estimate

$$
\Sigma_{kde} = \begin{bmatrix} 0.727 & -0.0188 & -0.0597 \\ -0.0188 & 0.609 & 0.0234 \\ -0.0597 & 0.0234 & 0.215 \end{bmatrix}.
$$

To determine whether the kernel density model appropriately fits the residuals of the Australian Election Study data, we can check the assumption that the residuals are independent deviates from the proposed error distribution.

Model diagnostics

For the kernel density model to be valid, the model's estimated error distribution $\varepsilon_{kde} \sim \sum_{i=1}^{n} \frac{1}{n} N(\varepsilon_i, \Sigma_{kde})$ should closely approximate the distribution of the residuals. As was done in Sections 6.1.4 and 6.2.4, the energy test for closeness of observed deviates to a proposed distribution of ε_{kde} was undertaken. Here, as previously, the test was applied to the residuals in the 2016 AES data set, using 100 sets of simulated data from the estimated distribution of ε_{kde} for comparison, and 500 replicates for the generation of each non-parametric p-value. The synthetic data sets are of the same size as the Australian Election Study ($n = 2818$). The mean p-value in testing the hypothesis that the data come from the proposed distribution of ε_{kde} was 0.478, indicating that there is insufficient evidence at the 5% significance level to suggest that the residuals do not belong to the kernel density model.

6.4   Summary of candidate models

As described in Sections 6.1, 6.2 and 6.3, three candidate models have been introduced as potential extensions to the multivariate regression model, which account

for the unusual residual structure present in the data from the Australian Election Study. Each of these models responds to the shortcomings of the multivariate regression model in a different way:

- The residual model attempts to explicitly model the relationships between marginal error terms;

- The mixture model attempts to segment the error terms into a more flexible Gaussian mixture; and

- The kernel density model fits a new, highly sensitive density function over the residuals.

In applying each of these models to the Australian Election Study data (see Sections 6.1.4, 6.2.4 and 6.3.3), it appears *prima facie* that all of the assumptions of each of the models appear to have been met, and so none of the models can be immediately discounted on this ground. Since each of the models provides the same conditional expectations of the political variables, model selection cannot be so simple as choosing the model with the smallest residual sum of squares, while making each model more parsimonious would reduce it to the multivariate case, whose assumptions fail.

More sophisticated methods for model comparison must then be explored, in order to distinguish betwen the three candidate models. In Chapter 7, model selection techniques will be introduced, discussed, and applied. These techniques will include a new and highly flexible framework for comparing non-nested models, which clarifies a technique called model mimicry, and applies it in multi-model environments and with diverse selection criteria.

Chapter 7

Assumption-Driven Model Selection

7.1 Assumption fidelity as a model selection criterion

In Chapter 5, it was determined that a multivariate regression model for assessing relationships between the dimension-reduced political variables and the dimension-reduced demographic variables was insufficient. This was because the assumption of multivariate normality of the residuals was not justified, regardless of which basis expansion of the dimension-reduced demographic variables was chosen. That the assumption of normality is not justified is important in the context of political modelling for two key reasons, which will now be discussed in turn.

Firstly, in order to account for the residual structure, a multivariate regression model may produce larger prediction regions than desired. The fact that insufficient models may produce larger prediction regions is discussed in Section 7.2.

Secondly, the shape of the error distribution, if correctly captured by the chosen model, may itself be informative. For example, in Section 5.2.3, it was noted that the multivariate regression model's error demonstrates that on average, those with more extreme positions than would be anticipated by the model on the spectrum of "social inclusivity" are also likely to have low levels of trust in authority, regardless of whether they are socially inclusive, or socially non-inclusive. This finding, that the quality of "extremity" can present itself in a number of political areas, and that this quality is not explicable by demography, is itself of value. In an explanatory model, even the error structure can provide insight. Given that the multivariate regression model was insufficient, Chapter 6 introduced three candidate models, each with the goal of responding to the unusual residual structure in the multivariate model. These three models are the recursive model (Section 6.1), the mixture model (Section 6.2) and the kernel density model (Section 6.3). Each model produces the same conditional expectations of the response variables (see Sections 6.1, 6.2 and 6.3), so criteria other than their mean squared error must be used to distinguish them.

We note that the multivariate model is nested within both the recursive model and the mixture model; the recursive model is the same as the multivariate model

when all coefficients of non-linear terms of the error recursions are zero, while the Gaussian mixture model with only one component reduces to the multivariate model. However, it is not true for any other pairing of the candidate models that one model is nested in the other. This is because:

- The mixture model and the recursive model are equivalent only in the special cases in which each is reduced to the multivariate regression model;

- The kernel density model cannot be reduced to the multivariate or the recursive model, since it comprises a fixed number n of normal distributions of equal probability; and

- The mixture model cannot be extended to be equivalent to a kernel density model, since a mixture model as described in Section 6.2 with n components estimated from n data points would have $\Sigma_m = 0 \; \forall \; m$.

As such, comparing the four candidate models must be done with regard to the fact that they do not form a nested structure. This non-nested structure will be particularly relevant in Section 7.3, in which we note that raw goodness-of-fit statistics cannot be directly compared for non-nested models. The following section describes a number of methods for comparing non-nested models.

7.2   Bases for comparing non-nested models

Where candidate non-nested models produce identical conditional expectations of the response variables, some other criteria must be relied upon to distinguish them. This section describes four such bases for model comparison:

- Cross-validated accuracy of prediction regions (Section 7.2.1);

- Size of prediction regions (Section 7.2.1);

- Parsimony and relevance of the model to its domain (Section 7.2.2); and

- Goodness-of-fit to model assumptions (Section 7.2.3).

Clearly, all of these bases are indicative, and not determinative, of the choice of model. Depending on the purpose of the modelling exercise, different bases may be allocated greater weight than others, but a choice of model should be based upon a balanced consideration of all relevant factors.

7.2.1   Prediction regions

There are two main things to consider when using prediction regions as a basis for model comparison: accuracy and size.

Accuracy:

A usable model should produce prediction regions that, for some set of new observations, contain the proportion of observations they purport to contain. If they are inaccurate, a model will either overstate or understate the level of confidence in its estimates. In order to determine whether prediction regions are accurate, a cross-validation procedure might be undertaken. In Section 6.2.1, such a cross-validation procedure was outlined in order to select the number of components of a Gaussian mixture model and reduce overfitting. The procedure from Section 6.2.1 can also be used for comparing models, and will be applied in this chapter.

Size:

Even where models produce prediction regions of similar accuracy, one model may produce prediction regions larger than others. For example, suppose errors for a model fall into two disjoint clusters. A model which does not account for this fact might, for example, produce a prediction region in the shape of an ellipse surrounding both clusters, while a model that does account for this might produce a prediction region comprised of two smaller disjoint ellipses around the two clusters. The second model, which is able to better explain the error structure may have the same accuracy as a competitor, while producing smaller prediction regions.

This example demonstrates that a good model should also produce the smallest prediction regions possible, which are still accurate. In some sense, this corresponds to the power of model's estimates. Power is a statistical measure of the probability of correctly rejecting a false null hypothesis. In the context of prediction regions, a new observation would be 'rejected' by a model at significance level α if it falls outside a $100(1 - \alpha)\%$ prediction region for the observation. A more powerful prediction region will 'reject', or exclude, a greater number of observations, while still containing $100(1 - \alpha)\%$ of new observations; in other words, the prediction regions should be as small as possible, while still being accurate.

For this reason, one might compare models of similar accuracy $|\alpha - \bar{p}_\alpha|$ on the basis of the size of their prediction regions. To assess this, the volume of prediction regions at all observed predictor values can be calculated and averaged for each model; this accounts for the variation in the volume of prediction regions accompanying different predictor values. A model with smaller prediction regions may then be preferred.

Application to the candidate models of political opinion

Using the cross-validation procedure outlined in Section 6.2.1, the accuracies of each of the three candidate models for the Australian Election Study ('AES')—the multivariate model (see Chapter 5), the recursive model (Section 6.1), the mixture model (Section 6.2) and the kernel density model (Section 6.3)—were calculated. This was done for prediction regions of confidence 80%, 90% and 95%, and with 20-fold cross-validation. At the same time, the size of these prediction regions, in units[3], were calculated. A model might be preferred that produces the most

Confidence	Model	Coverage (%)	Accuracy	Volume (units^3)
95%	Multivariate regression	94.0	0.010	1823
	Recursive	93.7	0.013	1707
	Gaussian mixture	94.0	0.010	1782
	Kernel density	94.7	0.003	1590
90%	Multivariate regression	89.5	0.005	1304
	Recursive	88.7	0.013	1231
	Gaussian mixture	89.7	0.003	1253
	Kernel density	89.7	0.003	1129
80%	Multivariate regression	80.4	0.004	834
	Recursive	79.5	0.005	791
	Gaussian mixture	79.2	0.008	747
	Kernel density	79.9	0.001	722

Table 7.1: Performance of the prediction regions of the four candidate models in a 20-fold cross-validation setting. Coverage refers to the cross-validated proportion of observed points that fall into prediction regions produced by the model, of the given confidence level. A well-calibrated model will have coverage similar to the confidence level, and thus have a small value under 'accuracy', which measures the absolute difference between the stated confidence level and the observed coverage. A powerful model will have small prediction regions, so the average volume of the cross-validated prediction regions is also calculated. It can be seen that all models perform relatively well in terms of accuracy, with the kernel density model performing best. In terms of size, the multivariate model in general performs worst, while the kernel density model produces the smallest prediction regions. This table suggests a slight preference for the kernel density model, since it performs marginally better than its competitors in these respects, but is more complex.

accurate prediction regions, with the smallest prediction regions.

Table 7.1 contains the purported coverage, true coverage, accuracy, and size of the prediction regions calculated under each candidate model, and for each confidence level. It can be seen that all models produced relatively accurate prediction regions. Even the least accurate prediction regions, the 95% prediction regions for the recursive multivariate model, contained 93.7% of observations. This is not a substantial difference from the purported coverage of the regions—an error of 1.3%. While the mixture and kernel density models perform the best in this regard, no substantial differentiation between the models can be made on the basis of prediction region coverage, since all models perform quite well.

Greater differentiation between models can be based upon prediction region size. The size of prediction regions is essentially a comparison of the power of the error distributions; with smaller prediction regions, a model can be as confident

about where predictions may lie, while being more specific. In this instance, a smaller prediction region means a more specific range of possible political identities a person might hold, with the same level of confidence. The multivariate regression model, on the whole, presents the largest prediction regions, while the smallest prediction regions are produced by the kernel density model. This is somewhat unsurprising; as the complexity of the models increase, they become better able to explain the error structure in the residuals. The kernel density model, which is built to replicate a 'general' probability distribution, has the most flexibility to respond to the pecularities of the true error distribution, so can hone in on these and produce more powerful, smaller prediction regions.

As a side note, and a point of reference, all of the models' prediction regions perform substantially better than a model not informed by demographic information, further validating the use of demographic identity as a predictor. As an example, a mixture model on the residuals of an intercept-only linear model produces 80% prediction regions with a cross-validated average size of 1144 units[3], as compared with 747 units[3] from a mixture model on the data-informed residuals. Thus, the demographic information from the Australian Election Study substantially reduces error in estimating political opinion.

7.2.2 Parsimony and relevance

All other things being equal, a simpler model should be preferred to a more complex one. Parsimony is an important consideration, both in terms of a model's predictive and explanatory power. While it might have smaller mean squared error, a more complex model is in danger of overfitting the observed data, meaning it produces small error in the observed data at the expense of higher error in predicting new observations. This is an example of the bias-variance tradeoff [84]; a more complex model may have lower bias, since it has greater flexibility to more closely align with the data, but higher variance, since it is more sensitive to fluctuations in the data. This issue might occur even where predictive error is identical for two candidate models; a model with more complex assumptions on its error distributions is more likely to overfit the observed errors.

Parsimony is also vital for a model to be explicable in its domain. If a model is used with the goal of describing real-world phenomena of interest, a model should be chosen that can explain these phenomena in an approachable way. As Shmeuli (2010) writes [160],

> "Explanatory modelling requires interpretable statistical models [...] that are easily linked to the underlying theoretical model."

In many circumstances, a more complicated model will be more difficult to relate to the constructs the model seeks to represent. As such, regardless of whether a model seeks to explain phenomena, or make predictions relating to these phenomena, a simpler model should be preferred to a more complex one.

More generally, in explanatory modelling, a model should be chosen so as to be able explain the real-world constructs around which it is based. In the context of

political modelling, for example, the goal is often to determine what associations exist between different factors, and political opinions. A model which is too complicated or opaque to succinctly relate political opinion to the predictive factors should not be selected.

Application to the candidate models of political opinion

In terms of conditional expectations of the response variables, each model explains relationships between political ideology and demographic identity in an identical way—as a linear function of the dimension-reduced demographic variables. Thus, any model based on these conditional expectations is a valuable explanatory tool for two reasons:

- Linear models represent relationships in an interpretable way, since as one predictor increases, estimates of response variables change in a linear way; and

- The linear model weighs heavily on the more interpretable dimension-reduced demographic variables (see Section 4), meaning valuable connections can be drawn between demographic and political identities.

The models, however, are of varying complexity when it comes to explaining the errors in estimating political opinion using demographic identity. These vary from the simplest, the multivariate regression model, which treats the error as a multivariate normal distribution, thus allowing for linear dependencies in the error terms, through to the kernel density model, which treats the error as a mixture of n normal distributions. In between these in terms of complexity, the recursive model treats each error axis as a function of other error directions, while the mixture model treats the error as a mixture of normal distributions.

In terms of a model's interpretability, complexity is indicative, but not determinative, of a model's fit to the domain it seeks to explain. A more complex model might be perfectly explicable in its context, while a relatively simple model might make no sense, or be entirely arbitrary, within the context of its domain. To this end, the following paragraphs provide an explanation for each model's approach to explaining error, in the context of the model's socio-political domain. It will be seen that some models make more sense than others in terms of the way they relate the estimability of political opinion using demographic information.

Multivariate regression model: The multivariate regression model, in assuming that the error in estimating political opinion using demographic identity is normally distributed, allows for linear relationships in the error to be explained. For example, if underestimation of an individual's social inclusivity was associated with underestimation of an individual's distrust in authority, in a linear way, this could be explained by the multivariate regression model. Since this would represent a positive correlation between error in the conditional expectations of the first and second dimension-reduced political axes, "social inclusivity" being the first and "attitudes to authority" being the second, this would be reflected by a positive term

in the $[1, 2]$-entry (and $[2, 1]$-entry) of Σ_{ε}, the estimated covariance matrix of the multivariate regression error term ε.

It should be noted that in this instance, the response variables, the dimension-reduced political axes, are uncorrelated by design (see Chapter 3). However, this does not mean that the multivariate regression error term will have diagonal covariance matrix. If the part of the variation in the response variables that is able to be explained by the predictors exhibits some correlation, the unexplainable portion of the response variables' variation will need to be correlated in the opposite direction to ensure the response variables are, overall, uncorrelated. This means there may still be linear relationships in the error in estimating political ideology, despite political ideology being constructed to be uncorrelated, making multivariate regression a more useful explanatory model than separate univariate regressions.

Recursive model: The recursive model is designed to explain non-linear relationships between error terms. By treating successive error terms as functions of those prior, the recursive model allows for more sophisticated explanation of the predictability and unpredictability of political thought.

For example, in Section 6.1.4 the recursive model was applied to the 2016 AES data set. It was determined that the coefficient of a ε_1^2 term in estimating Y_2 was 0.030. This means that when the squared error in estimating Y_1 increases by one unit, we expect that Y_2 increases by 0.030. Noting that Y_1 refers to 'social inclusivity' and Y_2 refers to 'attitudes to authority', this means that individuals with a higher absolute error in estimating social inclusivity are more likely to have a higher value for 'attitudes to authority'. High values on the 'attitudes to authority' are associated with distrust of authority. More simply, individuals who are much more socially inclusive, or much less socially inclusive, than expected by the model are more likely than average to distrust authority—more extreme social views than expected are associated with distrust of authority. The recursive model demonstrates that the error in a model can itself be of explanatory value.

Using the coefficient of ε_1^2 in combination with the coefficient of ε_1 in estimating Y_2, we could even find the roots of the estimated function of Y_2 with respect to ε_1. This allows us to find on which end of the 'social inclusivity' scale this extremity is more likely to occur. From an explanatory perspective, the recursive model is thus a good fit for the political domain.

However, the recursive model suffers two small flaws when it comes to parsimony:

1. The selection of the order of response variables is largely arbitrary, meaning many possible orderings may need to be tried at the model selection stage. This adds to the number of parameters to be estimated over the course of fitting the model. This problem is more significant for higher numbers of response variables, so it is less relevant here, as there are just three response variables.

2. The number of coefficients of the error terms is k^2 for k response variables (see Equation (6.1.1); addition of the ith successive response variable adds $2i - 1$

parameters), meaning the model rapidly becomes very complex as the number of response variables increases.

The second of these issues is shared by all other candidate models, since each include the estimation of at least one $k \times k$ covariance matrix. As such, the recursive model is not overly complex, and is of substantial explanatory value.

Gaussian mixture model: Similarly to the recursive model, the Gaussian mixture model is chosen as a candidate model for its ability to reflect non-linear relationships between error terms. Unlike the recursive model, the Gaussian mixture model does this essentially using a clustering approach; rather than explicitly stating proposed relationships between the variables, as the recursive model does, the mixture model segments respondents into clusters of various sizes, each of which is treated as an independent multivariate normal distribution. While each component of the Gaussian mixture allows for only linear relationships between error directions (see the multivariate regression model), the mixture of components allows for more complicated relationships to be explained, making the Gaussian mixture model useful as an explanatory tool.

For example, in the 2016 Australian Election Study application of the mixture model in Section 6.2.4, three Gaussian components were identified as producing a balance between parsimony and coverage. These three components have means $[-1.096, -0.643, -0.145]$, $[-1.908, 0.561, 0.468]$, and $[2.000, 0.315, -0.104]$. Noting that the first axis represents 'social inclusivity', with high scores indicating socially inclusive people and low scores indicating more socially conservative people, and that the second axis represents 'attitudes to authority', with high scores indicating distrust in authority, and low scores indicating high trust in institutions, we can interpret these clusters of respondents:

- The cluster centred at $[-1.096, -0.643, -0.145]$ represents people who are slightly less socially inclusive than average, and are more likely to have higher trust in institutions;

- The cluster centred at $[-1.908, 0.561, 0.468]$ represents people who are substantially less socially inclusive than average, and are more likely to have lower trust in institutions; and

- The cluster centred at $[2.000, 0.315, -0.104]$ represents people who are substantially more socially inclusive than average, and are more likely to have lower trust in institutions.

From this, we can note, similarly to in the recursive case, that people with more extreme views, than their demography would indicate, with respect to social inclusivity are less likely to trust institutions, while those with more centrist social views are more likely to trust current structures of power. We can further use these results to answer the interpretation question posed in the recursive model instance—'on which side of the 'social inclusivity' scale is trust in authority more likely to be present?'. Since the only cluster centred around people with greater than average trust in

authority is on the less socially inclusive side of the first axis, it might be said that people who are more trusting in authority than their demographic identity would indicate are more likely to be slightly less socially inclusive than their demographic identity would indicate. More sophisticated analysis of this would also consider the covariance matrices of each of the mixture components; this is done in Chapter 8.

As seen above, the Gaussian mixture model provides ample opportunity to discuss the outcomes of the model in the context of trends in political thought. It can thus be said that the Gaussian mixture is a valuable explanatory model, since it is interpretable, and can be "easily linked to the underlying [theory]" (Shmeuli 2010 [160]).

In terms of parsimony, the Gaussian mixture model does not rely on a complicated model fitting procedure, like the recursive model, but it still involves the estimation of a not-insignificant number of parameters. The mixture model requires the estimation of $M - 1$ proportions, M k-dimensional means, and M ($k \times k$) symmetric covariance matrices, for M mixture components and k-dimensional response. In this instance, with $M = 3$ and $k = 3$, this totals 29 parameters. This is not too extreme in this instance, since our dimension-reduced data set has $n = 2818$ 3-dimensional observations, but indicates that the number of components should not be substantially higher. The number of parameters in describing the error term is greater than the 9 parameters of the recursive model and the 6 of the multivariate regression model, but much less than the 8460 of the kernel density model. Thus, much like the recursive model, the Gaussian mixture model is sufficiently parsimonious, and domain-relevant.

Kernel density model: Unlike other candidates, the kernel density model is not designed from the perspective of explaining relationships in the error, but rather with the goal of achieving the best representation of the distribution of the residuals as possible. The kernel density model, which places a normal distribution at each observed residual, provides the greatest flexibility of all candidate models, as it can respond to any arrangement of the residuals. However, it is exactly this flexibility that lessens the kernel density model's explanatory value. Using the kernel density model, densities at particular values of the error term can be estimated, and plots of the error distribution can be made, but it is harder to directly draw out information from the kernel density model; there are too many parameters to make meaningful inference. Quite simply, the kernel density model misses the forest for the trees.

The fact that the kernel density model is not domain-relevant is directly related to its complexity. Because the kernel density model is a Gaussian mixture with n components, and common covariance matrix Σ_{kde}, there are $nk + \frac{k(k+1)}{2}$ parameters of the kernel density model with k response variables—the symmetric ($k \times k$) common covariance matrix has $\frac{k(k+1)}{2}$ unique parameters, and the n k-dimensional mean vectors of the components have nk parameters. In the instance of the 2016 Australian Election Study, with $k = 3$ and $n = 2818$, we have 8460 parameters in total. It is precisely this complexity that makes the kernel density model difficult to interpret in its political science context.

It would be disingenuous, however, to suggest that the 8460 parameters of the

kernel density model must be interpreted independently of each other. Indeed, graphical representations of the kernel density can help to observe trends in the error term being modelled. For example, a plot of the kernel density estimates of the 2016 AES model's error distribution would demonstrate that there is a lesser estimated density of individuals with centrist 'social inclusivity' views, but with a high score on the 'attitudes to authority' axis. This backs up observations from the mixture and recursive models, that more extreme social views are associated with distrust of institutions. However, making this observation by looking at a plot of the kernel density estimate is just a step removed from making this observation by looking at a scatter plot of the errors themselves. Making conclusions about relationships in the error term by inspection is less testable and repeatable than by directly using the parameters from the mixture or the recursive model. Since it is difficult to justify results from plots of the kernel density estimates, and it is difficult to draw out any real inference from the kernel density model itself, the kernel density model is the weakest of all candidates described here when it comes to explanatory value.

Summary The multivariate regression model, recursive model and Gaussian mixture model are all sufficiently parsimonious to neither overfit the data, nor preclude the use of the model as an explanatory tool. Each of these three candidates feature a small number of error-term parameters relative to the number of data points, with 6 for the multivariate regression model, 9 for the recursive model, and 29 for the mixture model, compared to 2818 data points overall. This small number of parameters both reduces the risk of overfitting, and allows for the easy interpretation of parameters in context. For the multivariate regression model, linear relationships between error terms can be explored. More valuably, the recursive and mixture models could help to explain non-linear relationships between errors in conditional expectations of political opinion. This is especially useful in light of Section 5.2, which demonstrated the existence of non-linear relationships in the error in expectations of political opinion.

This section also demonstrates the main drawback of the kernel density model. The kernel density model, which places a multivariate normal distribution around each observed residual to produce an estimated distribution of the error, contains 8460 parameters. While this large number of parameters does not necessarily produce an overfitting problem, due to the smoothing effect of mixing the distributions, it does make it very difficult to interpret the kernel density model in context. As a result, the kernel density model is not a particularly useful explanatory model, meaning it should not be selected in this explanatory context unless it so outperforms other candidate models in all other respects as to displace this conclusion. On the issue of domain-relevance, the recursive and mixture models perform best.

Note that often, the balance between goodness-of-fit and parsimony is done using goodness-of-fit statistics such as Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). While these values can be computed for each of the candidate models used here, we note that this would be of little use as their

values cannot directly be compared in this instance; AIC and BIC should not be compared for non-nested models (see, e.g., [172]).

7.2.3 Goodness-of-fit to assumptions

Another basis upon which to compare non-nested models is how reasonable the models' assumptions are. Goodness-of-fit to model assumptions is not a simple binary question; some valid models are more well-founded than others in terms of their assumptions. For example, consider a simple linear regression model. One assumption is normality of the error term, which is usually assessed using a quantile-quantile plot of the residuals. The assumption of normality might be considered reasonable for a number of different quantile-quantile plots, but one might more readily accept some plots than others. This is one demonstration of the fact that the validity of an assumption is on a continuum. Where a model fits on this continuum for some distributional assumption may distinguish some candidate models from others.

Since fidelity to assumptions justifies the generalisability of a model, a model might be favoured if it satisfies its assumptions better than other candidates. However, it is not immediately obvious how one might compare the fidelity of two or more models to their respective distributional assumptions. For example, if two models assume different error distributions, comparing these assumptions would involve comparing the goodness-of-fit of one set of errors to one distribution, with the goodness-of-fit of another set of errors to a different distribution. Test statistics for distributional fit, like the Kolmogorov-Smirnoff or energy statistics (Section 5.1.2; [166]), are not measured on the same scale for all candidate distributions, and may be more sensitive to some kinds of lack-of-fit than to others. These raw goodness-of-fit statistics thus cannot be directly compared between models. The problem of comparing goodness-of-fit among such non-nested models is explored in detail in the following section (Section 7.3), with a method for comparing multiple non-nested functions on the basis of assumption fidelity described in Section 7.4.3.

7.3 Goodness-of-fit comparisons of non-nested models

7.3.1 Literature on non-nested model comparison

Likelihood-ratio test for nested hypotheses

The likelihood-ratio statistic (Wilks 1938) is a well-known basis for hypothesis tests comparing two nested models [179]. When models are nested, they come from the same parameterised family, so the hypothesis test consists of choosing between two sets of parameters for this family, often denoted θ_A and θ_B. The likelihood-ratio statistic is

$$\Lambda = \frac{\mathcal{L}(\theta_A | \boldsymbol{x})}{\mathcal{L}(\theta_B | \boldsymbol{x})},$$

for observed data $\boldsymbol{x}$ and common likelihood function $\mathcal{L}$. Significance levels for this statistic are easily determined, since for nested models,

$$-2\log(\Lambda) \xrightarrow{\mathcal{D}} \chi^2_{p_B - p_A}$$

as sample size $n \to \infty$, with p_B the dimensionality of θ_B and p_A the dimensionality of θ_A (Wilks 1938) [179].

This test, however, cannot be undertaken for non-nested hypotheses, since in this case, the likelihoods in the expression for Λ would be from different families of distribution, meaning the distribution of Λ does not necessarily follow Wilks' asymptotic expression.

Extension to non-nested hypotheses

In response to this shortcoming, Cox (1961) [46] extends the likelihood-ratio test to some cases with non-nested hypotheses. Suppose we have some realisations $\boldsymbol{x}$ of random variable X, and seek to compare two hypotheses:

$$H_A : X \sim A(\theta_A),$$
$$H_B : X \sim B(\theta_B),$$

where A and B are non-nested distributions, and H_A is the null hypothesis. Cox' test statistic compares the logarithm of the ratio of likelihoods under each hypothesis to the expected value of the log-ratio under the null hypothesis. Cox' statistic can be written as

$$T = \log\left(\frac{\mathcal{L}_A(\theta_A|\boldsymbol{x})}{\mathcal{L}_B(\theta_B|\boldsymbol{x})}\right) - E_{A(\theta_A)}\left[\log\left(\frac{\mathcal{L}_A(\theta_A|\boldsymbol{x})}{\mathcal{L}_B(\theta_B|\boldsymbol{x})}\right)\right],$$

where $\mathcal{L}_A$ and $\mathcal{L}_B$ are likelihood functions of distributions A and B respectively, and $E_{A(\theta_A)}$ denotes an expectation with respect to distribution A with parameters θ_A.

Cox demonstrates that T is asymptotically normally distributed, with mean zero. He notes that the variance of T, and the expectation term in the expression for T, is difficult to calculate, depending on the distributions A and B [46]. Derivations of these values for a limited number of pairs of non-nested hypotheses have been published (see, e.g., [47, 97, 173]).

A Monte Carlo approach to test statistic distributions

Due to the intractability of the asymptotic distribution of Cox' statistic T for some pairs of non-nested hypotheses, and the fact that T often converges to its asymptotic distribution slowly (Williams 1970) [181], Williams introduces a simulation approach to determining a distribution of the test statistic. Using an equivalent variation on the test statistic,

$$\lambda = \log\left(\frac{\mathcal{L}_A(\theta_A|\boldsymbol{x})}{\mathcal{L}_B(\theta_B|\boldsymbol{x})}\right),$$

Williams proposes simulating distributions for λ under both the null and alternative hypotheses, and then drawing a conclusion as to which hypothesis is to be favoured [180, 181]. More explicitly, his approach consists of the following steps:

- Estimate parameters θ_A and θ_B, for null distribution A and alternative distribution B respectively, from the n observations in observed data $\boldsymbol{x}$, and calculate an observed value of λ.

- Simulate R datasets of size n under $A(\theta_A)$, denoted $\boldsymbol{x}_A^r : r = 1, \ldots, R$, and R datasets of size n under $B(\theta_B)$, denoted $\boldsymbol{x}_B^r : r = 1, \ldots, R$.

- For each simulation $r = 1, \ldots, R$, calculate

$$\lambda_A^r = \log\left(\frac{\mathcal{L}_A(\theta_A|\boldsymbol{x}_A^r)}{\mathcal{L}_B(\theta_B|\boldsymbol{x}_A^r)}\right) \text{ and } \lambda_B^r = \log\left(\frac{\mathcal{L}_A(\theta_A|\boldsymbol{x}_B^r)}{\mathcal{L}_B(\theta_B|\boldsymbol{x}_B^r)}\right).$$

- Compare the observed value of λ to the distributions of the λ_A^rs and the λ_B^rs:

    The null hypothesis, that $X \sim A(\theta_A)$, is to be preferred if the observed value of λ is more likely under the distribution of λ_A^r than of λ_B^r. Otherwise, the alternative hypothesis, that $X \sim B(\theta_B)$, is to be preferred.

If the observed value of λ is unlikely under both distributions, it may be that neither model is preferable, and if the observed value of λ is likely under both models, there is insufficient evidence to prefer one model over another [180].

Williams proposes $R = 10$ due to the historical computing limitations, but larger values can now be chosen to give greater confidence in model choice [181].

Williams notes that a limitation of his approach is that the simulated distribution of the test statistic is strongly dependent upon the values of the estimated parameters θ_A and θ_B, and suggests further simulation may alleviate this [181]. He does not suggest a specific method for doing so. The model mimicry method, discussed in Section 7.4, accounts for this parameter uncertainty using a bootstrap.

7.4 The model mimicry method

7.4.1 The parametric bootstrap cross-fitting method for two models

As noted in Section 7.3.1, the approach of Williams in non-nested model selection is hindered by the fact that it does not account for parameter uncertainty [181, 172].

An additional limitation of his approach is that it is built around a likelihood ratio goodness-of-fit measure [172]. While this is suitable for a number of model comparison situations, popular measures like the Akaike information criterion (AIC) and the Bayes information criterion (BIC) might be sought to be compared across non-nested models.

Other goodness-of-fit measures may also be preferred in specific instances. For example, in the political modelling situation described in this work, a model is

sought whose distributional assumptions are best justified. A goodness-of-fit measure for multivariate distributions is thus more appropriate here than comparing likelihoods.

Accounting for these drawbacks of the approach of Williams (1970), Wagenmakers *et al.* (2004) [172] present a method for testing hypotheses of non-nested models which both accounts for uncertainty in parameter estimation, and is suited to general goodness-of-fit measures. The method consists of re-framing Williams' approach in terms of a generic goodness-of-fit measure, and adding the additional step of a non-parametric bootstrap prior to each simulation. The non-parametric bootstrap precludes the distribution of goodness-of-fit measures from relying heavily on a particular parameter estimate; instead, for a stable model, a variety of parameters will be used, drawn from the region of the parameter space inhabited by those estimated using the observed data.

Wagenmakers *et al.* call this approach "model mimicry" [172]. This is because the method leads to the selection of models that best replicate the observed data. The specific method proposed is called the "parametric bootstrap cross-fitting method" ('PBCM'), since the act of simulating data under each model is a parametric bootstrap, and each model is fit to both models' simulations.

The PBCM approach of Wagenmakers *et al.* is as follows [172]:

1. Apply the non-parametric bootstrap to the n observations in observed data $\boldsymbol{x}$; in other words, take a sample of size n from $\boldsymbol{x}$, sampling with replacement. Denote this $\boldsymbol{x}^r$.

2. Estimate parameters $\theta_A(\boldsymbol{x}^r)$ and $\theta_B(\boldsymbol{x}^r)$, for null distribution A and alternative distribution B respectively, from the n observations in bootstrap $\boldsymbol{x}^r$.

3. Simulate a dataset of size n under $\theta_A(\boldsymbol{x}^r)$, denoted $\boldsymbol{x}_A^r$, and a dataset of size n under $\theta_B(\boldsymbol{x}^r)$, denoted $\boldsymbol{x}_B^r$.

4. Fit both models to both sets of simulated data, and calculate goodness-of-fit ('GOF') measures for each of the models' fit. In other words:

   - Estimate parameters $\theta_A(\boldsymbol{x}_A^r)$ and $\theta_B(\boldsymbol{x}_A^r)$, for null distribution A and alternative distribution B respectively, from the n observations in $\boldsymbol{x}_A^r$, and calculate GOF measures $GOF_A(\boldsymbol{x}_A^r)$ and $GOF_B(\boldsymbol{x}_A^r)$; and

   - Estimate parameters $\theta_A(\boldsymbol{x}_B^r)$ and $\theta_B(\boldsymbol{x}_B^r)$, for null distribution A and alternative distribution B respectively, from the n observations in $\boldsymbol{x}_B^r$, and calculate GOF measures $GOF_A(\boldsymbol{x}_B^r)$ and $GOF_B(\boldsymbol{x}_B^r)$.

5. For the data generated from A, calculate the difference in the goodness-of-fit measures between the two models:

$$\Delta GOF_A^r = GOF_A(\boldsymbol{x}_A^r) - GOF_B(\boldsymbol{x}_A^r).$$

Do the same for the data generated from B:

$$\Delta GOF_B^r = GOF_A(\boldsymbol{x}_B^r) - GOF_B(\boldsymbol{x}_B^r).$$

6. Repeat steps 1-5 for $r = 1, \ldots, R$, yielding R observations from the distribution of ΔGOF_A and of ΔGOF_B.

7. Meanwhile, fit both models to the observed data $\boldsymbol{x}$, yielding $\theta_A(\boldsymbol{x})$ and $\theta_B(\boldsymbol{x})$. Calculate the goodness-of-fit of each model, $GOF_A(\boldsymbol{x})$ and $GOF_B(\boldsymbol{x})$, and the difference between these:

$$\Delta GOF_{obs} = GOF_A(\boldsymbol{x}) - GOF_B(\boldsymbol{x}).$$

8. Compare the observed value ΔGOF_{obs} to the distributions of ΔGOF_A and ΔGOF_B:

   - The null hypothesis, that $X \sim A(\theta_A)$, is to be preferred if the observed value ΔGOF_{obs} is more likely under the distribution of ΔGOF_A than of ΔGOF_B; otherwise, the alternative hypothesis, that $X \sim B(\theta_B)$, is to be preferred.

   - In other words, select model A if

   $$\frac{f(\Delta GOF_{obs} | A \text{ is true})}{f(\Delta GOF_{obs} | B \text{ is true})} \geq 1,$$

   for density function f.

A diagram outlining the model mimicry method can be found in Figure 7.1.

7.4.2 An example of model mimicry

Following the PBCM procedure of Wagenmakers *et al.*, this section contains a comparison of data from the normal and Cauchy distributions. Data was simulated from each of these distributions, and the PBCM was used to compare the ability of each model to replicate data from the other, using the energy statistic (Section 5.1.2) as a distributional goodness-of-fit measure. It should be noted here that traditional non-nested model selection techniques, based around asymptotic distributions of the Cox relative likelihood T (Section 7.3.1), presents an intractable problem in this case, since expectations and variances of the Cauchy distribution, and of functions thereof, are undefined. Use of the empirical energy statistic to measure goodness-of-fit is thus preferred due to its non-parametric nature.

Data from a normal distribution: First, 100 variates from the distribution $N(5, 2^2)$ were simulated. The parametric bootstrap cross-fitting method was then applied to this data, comparing the ability of a normal distribution and a Cauchy distribution to describe the data, with 500 replicates. This yielded 500 observations from a distribution of ΔGOF_{normal} and ΔGOF_{Cauchy}, the difference in Energy statistic between the normal and the Cauchy distributions when the true model is assumed to be normal and Cauchy respectively. Both models were also fit to the observed data, yielding the observed difference in Energy statistic between the two models (ΔGOF_{obs}).

(1) Simulate distributions of ΔGOF under the null and alternative hypotheses:
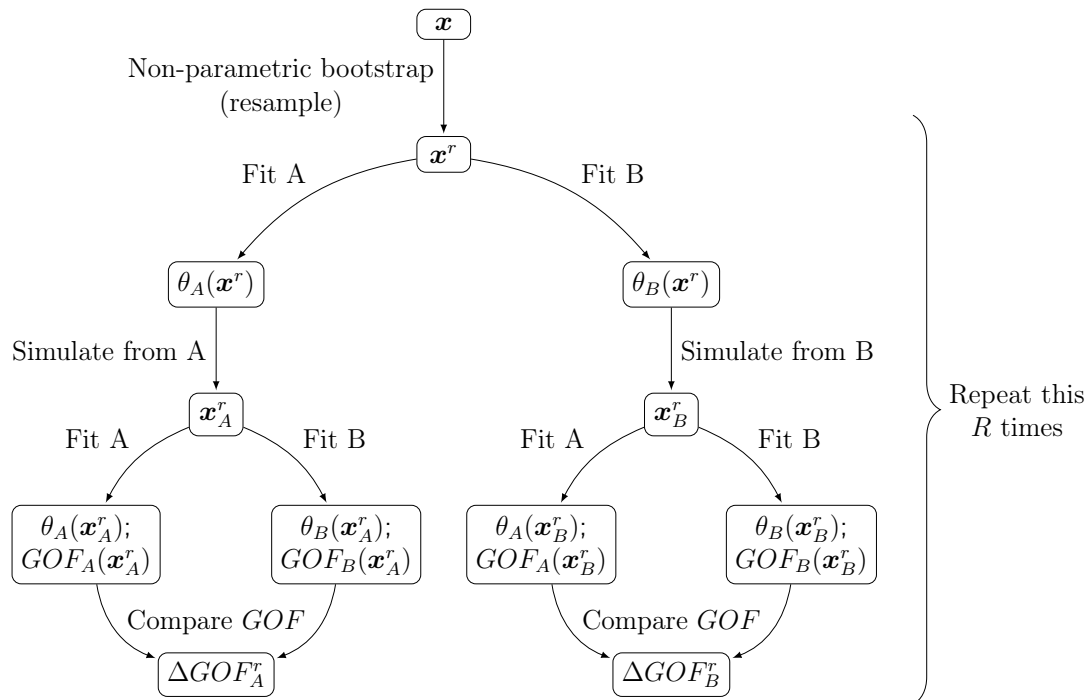
(2) Calculate the observed value of ΔGOF:
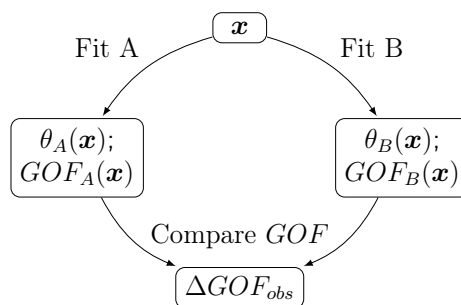
Figure 7.1: The parametric bootstrap cross-fitting method (PBCM) outlined by Wagenmakers *et al.* (2004). Simulated distributions of differences in goodness-of-fit under competing models are compared to the difference between the models' goodness-of-fit for the observed data.
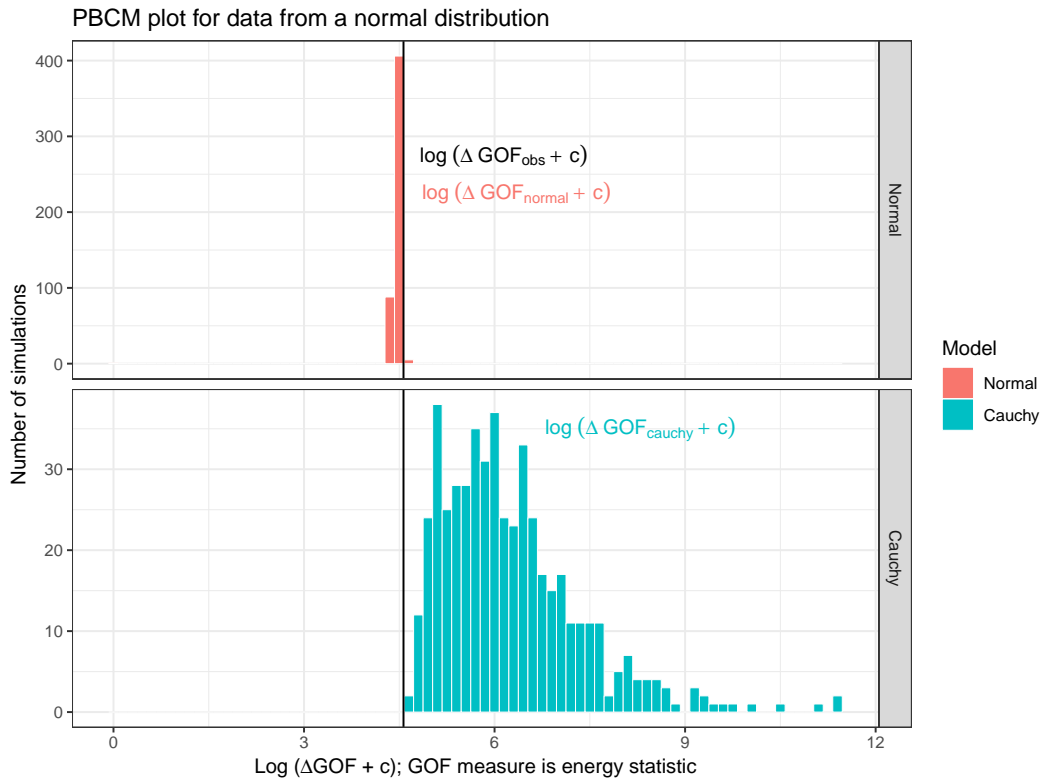
Figure 7.2: Plot comparing the fit of the normal and Cauchy distributions to data simulated from the normal distribution, using the parametric bootstrap cross-fitting method outlined by Wagenmakers *et al.*; the normal distribution appears to be the best fit to the data.

For a visual representation, a plot of logarithms of the two ΔGOF distributions, shifted by a constant c, with a vertical line for $\log(\Delta GOF_{obs} + c)$, can be found in Figure 7.2. Natural logarithms needed to be taken due to the very large variation in empirical energy statistics for the Cauchy distribution. The constant is added to ensure all values are positive, since logarithms can only be taken of positive values.

Figure 7.2 demonstrates that the normal distribution is a substantially better fit to the data than the Cauchy distribution, since

$$f(\Delta GOF_{obs}|X \sim N(\mu, \sigma^2)) > f(\Delta GOF_{obs}|X \sim \text{Cauchy}(x_0, \gamma)),$$

for model parameters μ, σ^2, x_0, and γ, and true data distribution X. This probability is invariant to the logarithmic transform in Figure 7.2.

Data from a Cauchy distribution: A comparison was also made between the fit of the normal and Cauchy distributions to data simulated according to a Cauchy distribution. Again, 100 variates were simulated, in this instance from the distribution Cauchy$(1, 5)$. The PBCM was applied, with 500 replicates, comparing the normal distribution to the Cauchy distribution. A histogram of the 500 observations
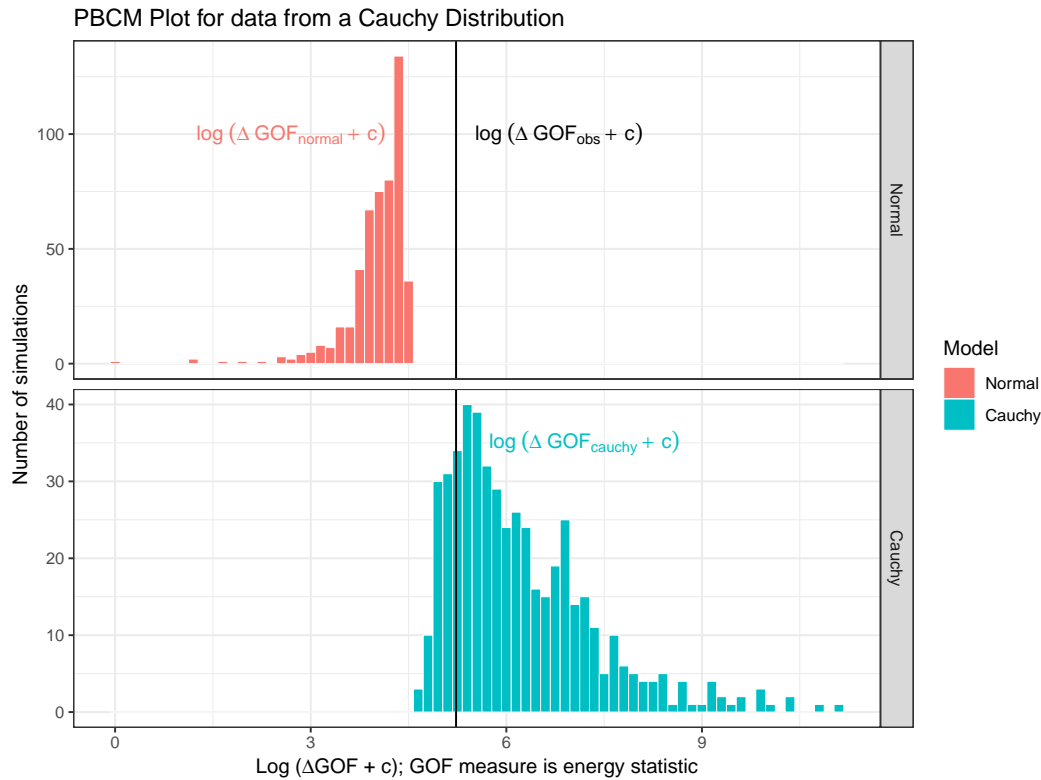
Figure 7.3: Plot comparing the fit of the normal and Cauchy distributions to data simulated from a Cauchy distribution, using the parametric bootstrap cross-fitting method outlined by Wagenmakers *et al.*; the Cauchy distribution appears to be the best fit to the data.

from the distributions of $\log(\Delta GOF_{normal}+c)$ and $\log(\Delta GOF_{Cauchy}+c)$ can be found in Figure 7.3. Once more, logarithms were taken due to large variation in energy statistics for the Cauchy distribution. It is clear from Figure 7.3 that a Cauchy distribution is a better fit to the Cauchy simulated data than a normal distribution, since

$$f(\Delta GOF_{obs}|X \sim \text{Cauchy}(x_0, \gamma)) > f(\Delta GOF_{obs}|X \sim N(\mu, \sigma^2)),$$

for model parameters μ, σ^2, x_0, and γ, and true data distribution X.

As seen in these examples, simulated distributions of goodness-of-fit measures provide a useful reference by which to compare the fit of two models. When the observed difference in goodness-of-fit between two models is more likely under one model than another, this model is to be preferred, regardless of whether the raw goodness-of-fit statistic is higher for one model than for another. The parametric bootstrap cross-fitting method, in being able to compare models for which goodness-of-fit statistics are not directly comparable, and being able to compare irregular models like the Cauchy distribution, presents a robust method for non-nested model comparison.

7.4.3 Extensions to more than two models

Wagenmakers *et al.*'s parametric bootstrap cross-fitting method ('PBCM') is limited in that it allows for the comparison of only two models. The method of comparing distributions of differences in goodness-of-fit does not easily extend to greater than two models, so when more than two non-nested candidate models are to be compared, another approach should be taken.

While more than two models can be compared on a pairwise basis, it should be noted that the PBCM assumes in each comparison that one of the competing models is true. The decision rule for model selection encouraged by the PBCM, that we should prefer model A over B if

$$\frac{f(\Delta GOF_{obs}|A \text{ is true})}{f(\Delta GOF_{obs}|B \text{ is true})} \geq 1,$$

is less useful when neither model A nor B is true.

For example, if comparing on a pairwise basis three models A, B and C, for which model A is true, one of the three comparisons will be undertaken on an incorrect assumption–that between model B and model C. This, however, will not be a fatal problem if model A is indeed true; the pairwise applications of the PBCM should reveal model A is preferable to both models B and C, rendering the comparison between models B and C unnecessary. If one is willing to accept the discomfort of undertaking particular pairwise analyses on flawed assumptions, the pairwise model selection procedure may be appropriate.

If one seeks to avoid this by comparing all models simultaneously, the similar approaches of Allcroft and Glasbey (2003) [3] and Schultheis and Naidu (2014) [156] may be preferred. Allcroft and Glasbey (2003) [3], one year before Wagenmakers *et al.* published their methodology [172], propose a technique similar to the PBCM, but with the capability to compare more than two models. The major point of difference between the Allcroft and Glasbey method and the PBCM, is that while $\Delta GOF = GOF_1 - GOF_2$ distributions are simulated in the PBCM, the Allcroft and Glasbey method uses raw GOF values to simulate multivariate distributions of $[GOF_1, GOF_2, \ldots, GOF_M]$ under each of models $1, 2, \ldots, M$. The observed value of $[GOF_1, GOF_2, \ldots, GOF_M]$ is then compared to the M simulated distributions to determine which hypothesis is most likely. After simulation, model selection then becomes a classification problem in an M-dimensional space.

In contrast to the PBCM, the Allcroft and Glasbey method omits the non-parametric bootstrap at each simulation, and does not re-estimate the parameters of each model for each simulation, instead using only the estimated parameters from the observed data throughout the procedure. These omissions are reversed in the work of Schultheis and Naidu (2014) [156], and their technique will be preferred here to reduce the procedure's sensitivity to parameter estimates.

The favoured method, of Schultheis and Naidu (2014), is thus as follows [156]:

1. Apply the non-parametric bootstrap to the n observations in observed data $\boldsymbol{x}$; in other words, take a sample of size n from $\boldsymbol{x}$, sampling with replacement. Denote this $\boldsymbol{x}^r$.

2. Estimate parameters $\theta_1(\boldsymbol{x}^r), \theta_2(\boldsymbol{x}^r), \ldots \theta_M(\boldsymbol{x}^r)$, for proposed distributions $1, 2, \ldots, M$ respectively, from the n observations in bootstrap $\boldsymbol{x}^r$.

3. Simulate dataset of size n under each of $\theta_1(\boldsymbol{x}^r), \theta_2(\boldsymbol{x}^r), \ldots \theta_M(\boldsymbol{x}^r)$, denoted, $\boldsymbol{x}_1^r, \boldsymbol{x}_2^r, \ldots, \boldsymbol{x}_M^r$ respectively.

4. Fit every model to every set of simulated data, and calculate goodness-of-fit ('GOF') measures for each of the models' fit. In other words:

   - Estimate parameters $\theta_1(\boldsymbol{x}_1^r), \theta_2(\boldsymbol{x}_1^r), \ldots, \theta_M(\boldsymbol{x}_1^r)$, for distributions $1, 2, \ldots, M$ respectively, from the n observations in $\boldsymbol{x}_1^r$, and calculate GOF measures $GOF_1(\boldsymbol{x}_1^r), GOF_2(\boldsymbol{x}_1^r), \ldots, GOF_M(\boldsymbol{x}_1^r)$;

   - Estimate parameters $\theta_1(\boldsymbol{x}_2^r), \theta_2(\boldsymbol{x}_2^r), \ldots, \theta_M(\boldsymbol{x}_2^r)$, for distributions $1, 2, \ldots, M$ respectively, from the n observations in $\boldsymbol{x}_2^r$, and calculate GOF measures $GOF_1(\boldsymbol{x}_2^r), GOF_2(\boldsymbol{x}_2^r), \ldots, GOF_M(\boldsymbol{x}_2^r)$;

$$\vdots$$

   - Estimate parameters $\theta_1(\boldsymbol{x}_M^r), \theta_2(\boldsymbol{x}_M^r), \ldots, \theta_M(\boldsymbol{x}_M^r)$, for distributions $1, 2, \ldots, M$ respectively, from the n observations in $\boldsymbol{x}_M^r$, and calculate GOF measures $GOF_1(\boldsymbol{x}_M^r), GOF_2(\boldsymbol{x}_M^r), \ldots, GOF_M(\boldsymbol{x}_M^r)$.

5. Repeat steps 1-4 for $r = 1, \ldots, R$, yielding R observations from the distributions of

$$\boldsymbol{GOF}_1 = [GOF_{1|1}, GOF_{2|1}, \ldots GOF_{M|1}],$$
$$\boldsymbol{GOF}_2 = [GOF_{1|2}, GOF_{2|2}, \ldots GOF_{M|2}],$$
$$\vdots$$
$$\boldsymbol{GOF}_M = [GOF_{1|M}, GOF_{2|M}, \ldots GOF_{M|M}],$$

   where $GOF_{i|j}$ is the goodness-of-fit to model i of data produced according to model j.

6. Meanwhile, fit all models to the observed data $\boldsymbol{x}$, yielding $\theta_1(\boldsymbol{x}), \theta_2(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})$. Calculate the goodness-of-fit of each model,

$$\boldsymbol{GOF}_{obs} = [GOF_1(\boldsymbol{x}), GOF_2(\boldsymbol{x}), \ldots, GOF_M(\boldsymbol{x})].$$

7. Compare the observed value $\boldsymbol{GOF}_{obs}$ to the distributions of $\boldsymbol{GOF}_1$, $\boldsymbol{GOF}_2$, $\ldots \boldsymbol{GOF}_M$:

   - The hypothesis to be preferred is that under which the observed value $\boldsymbol{GOF}_{obs}$ is most likely.

   - In other words, select the model satisfying

$$\underset{i=1,\ldots,M}{\operatorname{argmax}} f(\boldsymbol{GOF}_{obs} | \boldsymbol{GOF}_{obs} \sim \boldsymbol{GOF}_i).$$

A diagram outlining the multiple model comparisons method can be found in Figure 7.4.
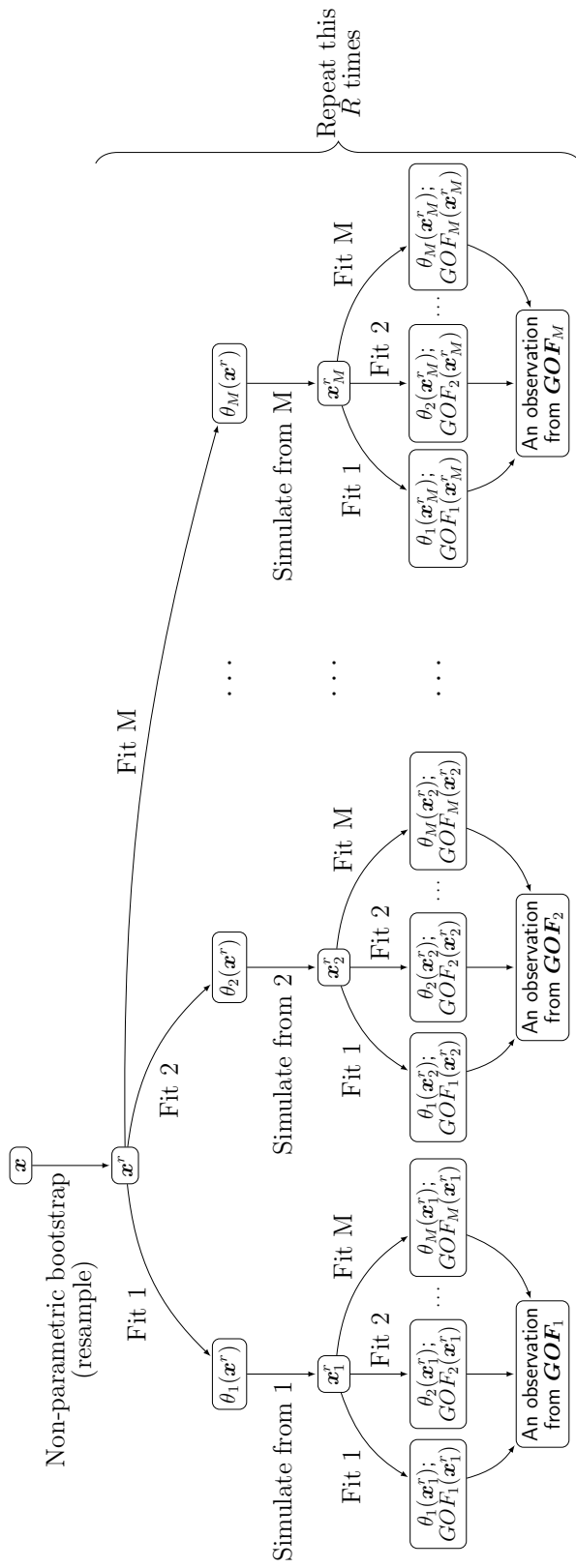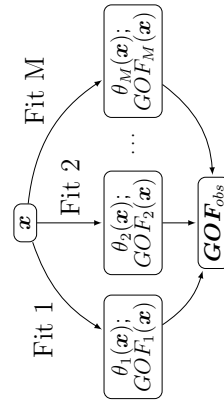
Figure 7.4: The multi-model variation of the parametric bootstrap cross-fitting method (PBCM) outlined by Schultheis and Naidu (2014). Simulated multivariate distributions of goodness-of-fit under competing models are compared to all models' goodness-of-fit to the observed data.

7.4.4   Classifying results from the multi-model PBCM

In Step 7 of the method adapted here from Schultheis and Naidu [156], a model is chosen which satisfies

$$\operatorname*{argmax}_{i=1,\ldots,M} f(\boldsymbol{GOF}_{obs}|\boldsymbol{GOF}_{obs} \sim \boldsymbol{GOF}_i).$$

Unfortunately, the distributions of $\boldsymbol{GOF}_i, i = 1, 2, \ldots, M$ are only known through the R simulated observations of these distributions. The task of choosing the model that maximises the density of $\boldsymbol{GOF}_{obs}$ under that model is, in other words, a supervised classification task, assigning a new point to one of M sets of observed points. Three popular methods for supervised classification are

- inspection,

- model-based classifiers, and

- non-parametric classifiers.

The first two such methods are discussed in the paragraphs below. For further discussion of non-parametric classifiers, including k-nearest neighbour methods, see Hastie, Tibshirani and Friedman [84]; these methods are not used here as the decision boundaries in these scenarios are rarely so irregular as to necessitate such an approach.

Inspection   With just a single data point to classify, it seems natural at first to determine the nearest distribution by inspection. In this context, this would amount to looking at a plot containing $\boldsymbol{GOF_{obs}}$ and all observations of $\boldsymbol{GOF}_i, i = 1, 2, \ldots, M$, and determining which set of observations appears closest to $\boldsymbol{GOF_{obs}}$. For simple analyses, there is nothing inherently wrong with this approach; indeed, in their paper introducing the PBCM, Wagenmakers *et al.* use inspection to classify models for their Example 1 [172]. However, there are two key flaws with this methodology; that inspection becomes more difficult in a higher-dimensional space, and that inspection may become inefficient for multiple data sets.

In the examples of Wagenmakers *et al.* [172], two models are differentiated using the original PBCM; that is, the version able to deal with a binary model selection. Wagenmakers *et al.* were thus able to make model selections by inspecting histograms. For the multi-model PBCM, multi-dimensional distributions of points are considered, making visualisation of $\boldsymbol{GOF_{obs}}$ within the goodness-of-fit space difficult. Pairwise scatterplots (by dimension) are possible, and are used in this thesis, though information is lost in showing just marginal goodness-of-fit distributions. An alternative used in this thesis is a two-dimensional principal components plot, which is able to represent a much larger proportion of the variation in the data than a two-dimensional marginal plot (see Section 3.1). This is one visualisation used by Schultheis and Naidu [156], though the principal components plot is also unlikely to fully convey the higher-dimensional system it represents. This thesis will thus propose the use of discriminant-based classifiers for PBCM output.

Another issue with model selection by inspection is that it becomes inefficient when a larger number of data sets are considered. For example, suppose a model is sought to describe multiple potential realisations of data sets. This situation is common in psychological modelling (see, e.g., Wagenmakers *et al.* [172]). In this instance, it may be necessary to classify many values of $\boldsymbol{GOF_{obs}}$, with separate sets of distributions of $\boldsymbol{GOF}_i, i = 1, 2, \ldots, M$ for each data set. By inspection, this would require considering a large number of visualisations in order to draw conclusions. Numerical classification of $\boldsymbol{GOF_{obs}}$ should then be considered.

Discriminant-based classifiers   Numerically, we seek the probability that $\boldsymbol{GOF_{obs}}$ belongs to some model $\mathcal{M}$. In other words, for each $i = 1, 2, \ldots, M$, we seek

$$P(\mathcal{M} = i \mid \boldsymbol{GOF_{obs}}).$$

Using Bayes' theorem, we can rearrange this to

$$\begin{aligned} P(\mathcal{M} = i \mid \boldsymbol{GOF_{obs}}) &= \frac{f(\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i)P(\mathcal{M} = i)}{f(\boldsymbol{GOF_{obs}})} \\ &= \frac{f(\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i)P(\mathcal{M} = i)}{\sum_{j=1}^{M} f(\boldsymbol{GOF_{obs}} \mid \mathcal{M} = j)P(\mathcal{M} = j)}. \end{aligned}$$

Since there are the same number (denoted earlier R) of observations for each model, and there is no assumed prior preference for any model, we can treat $P(\mathcal{M} = i) = P(\mathcal{M} = j) \, \forall \, i, j$:

$$P(\mathcal{M} = i \mid \boldsymbol{GOF_{obs}}) = \frac{f(\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i)}{\sum_{j=1}^{M} f(\boldsymbol{GOF_{obs}} \mid \mathcal{M} = j)}. \tag{7.4.1}$$

In order to classify $\boldsymbol{GOF_{obs}}$, we thus need to estimate $f(\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i)$. We briefly discuss here estimation of the distributions of $\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i$ in three manners of increasing complexity (all from Hastie, Tibshirani and Friedman (2009) [84]):

- Linear discriminant analysis (LDA), which assumes homoscedastic normal distributions;

- Quadratic discriminant analysis (QDA), which assumes heteroscedastic normal distributions; and

- Mixture discriminant analysis (MDA), which assumes mixtures of heteroscedastic normal distributions.

It should be noted that Hastie, Tibshirani and Friedman restrict their version of MDA to the assumption of mixtures of homoscedastic normal distributions ([84] at page 440), while this is generalised here to mixtures of heteroscedastic normal distributions for additional flexibility.

Linear discriminant analysis (LDA), the simplest of the three methods, is named because it produces linear decision boundaries; in other words, the boundary between the region whose points that would be classified to one model, and the region that would be classified to another, is always linear [84]. The LDA model supposes that each distribution can be expressed as

$$\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i \ \sim \ N(\boldsymbol{\mu_i}, \Sigma),$$

for $i = 1, 2, \ldots, M$. Note here that Σ is not dependent on i, meaning the variance is assumed to be the same for all models. Calculating the density functions can be done by maximum likelihood; each mean $\boldsymbol{\mu}_i$ is simply estimated to be the sample mean of each $\boldsymbol{GOF}_i$, while the variance Σ is estimated to be the weighted average of sample variance matrices $\hat{\Sigma}_i$ for each $\boldsymbol{GOF}_i$. Since in this case, there are the same number of observations of each goodness-of-fit distribution,

$$\hat{\Sigma} = \frac{1}{M} \sum_{i=1}^{M} \hat{\Sigma}_i,$$

for variance estimate $\hat{\Sigma}$, and sample variances $\hat{\Sigma}_i$ estimated by

$$\hat{\Sigma}_i = \frac{1}{R-1} \sum_{r=1}^{R} (\boldsymbol{GOF}_{i,r} - \overline{\boldsymbol{GOF}}_i)(\boldsymbol{GOF}_{i,r} - \overline{\boldsymbol{GOF}}_i)^{\mathrm{T}}, \qquad (7.4.2)$$

for $\boldsymbol{GOF}_{i,r}$ the ith observation of $\boldsymbol{GOF}_i$. Once these parameters have been estimated, Equation (7.4.1) can be used to select a model; the model that maximises the discriminant in (7.4.1) can be selected. The higher the value of the discriminant for the chosen model, the greater the confidence one can express in the selection of that model.

Quadratic discriminant analysis (QDA) relaxes the constant variance assumption from LDA. This means it is able to respond to differing covariance structures between goodness-of-fit distributions. The example later in this section demonstrates that this may be useful in the multi-model PBCM context; in that instance, the variance of distributions, as represented in Figure 7.5, was not the same for all candidate models.

The QDA model supposes that each distribution can be expressed as

$$\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i \ \sim \ N(\boldsymbol{\mu_i}, \Sigma_i),$$

for $i = 1, 2, \ldots, M$. Note here that Σ_i is dependent on i. Again, density functions are calculated by maximum likelihood, with means for each model estimated by sample mean $\overline{\boldsymbol{GOF}}_i$, and sample variances as in Equation (7.4.2) above.

Mixture discriminant analysis (MDA) presupposes that $\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i$ can be approximated by a Gaussian mixture, which is a weighted sum of normal random variables. Thus MDA allows for more flexible decision boundaries than LDA or QDA, by accounting for the fact that the distributions of goodness-of-fit

among candidate models may be multi-modal. The MDA model supposes that each
goodness-of-fit distribution can be expressed as

$$\boldsymbol{GOF_{obs}} \mid \mathcal{M} = i \ \sim\ \sum_{l=1}^{k_i} \pi_{i,l} N(\boldsymbol{\mu_{i,l}}, \Sigma_{i,l}),$$

for k_i components of the ith mixture model, $i = 1, 2, \ldots, M$. Note here that the
means, variances and even number of components may vary within and among
candidate distributions.

The allowance for the covariance matrices being not identical for all components
and for all $\boldsymbol{GOF}_i$ distributions is an extension of the discriminant suggested by
Hastie, Tibshirani and Friedman ([84] at page 440). The additional flexibility
provided by this extension is necessary in situations such as that in the example
in Figure 7.6, in which the distribution of goodness-of-fit under the log-normal
candidate model, unlike the distribution for other candidate models, has a highly
irregular covariance structure. Only a more flexible Gaussian mixture is able to
capture this irregular covariance structure, which requires multiple components with
different covariance matrices.

The parameters of the mixture model, within each group, can be estimated
using an Expectation Maximisation (EM) algorithm for Gaussian mixtures [183].
This EM algorithm is explained in Section 6.2.

Using LDA, QDA or MDA, models can be selected by maximising the dis-
criminant in Equation (7.4.1). If the LDA, QDA or MDA assumptions hold, the
discriminant in Equation (7.4.1) represents an estimate of the probability that each
model is true, given $\boldsymbol{GOF_{obs}}$, using an uninformative prior [84]. This provides for
a numerical representation of not just which model should be selected, but also the
uncertainty in this choice of model.

7.4.5   An example of multi-model mimicry

Using this thesis' expansion of the procedure of Schultheis and Naidu (2014) [156],
this section contains an example concerning data simulated from an exponential
distribution. Hypotheses that the data come from an exponential distribution,
a log-normal distribution and a chi-squared distribution are compared using the
multi-model variation of the parametric bootstrap cross-fitting method ('PBCM'),
using the energy statistic (Section 5.1.2) as a measure of goodness-of-fit to each
distribution. Since the three candidate distributions are similar, they should be
difficult to distinguish, making this example demonstrative of the power of the
multi-model variation of the PBCM.

First, 100 variates from the distribution Exp(1) were simulated. The multi-
model variation on the PBCM was then applied to this data, comparing the ability of
an exponential distribution, a log-normal distribution and a chi-squared distribution
to describe the data, with 500 replicates. This yielded 500 observations from the
distributions of $\boldsymbol{GOF_{exp}}$, $\boldsymbol{GOF_{logN}}$ and $\boldsymbol{GOF_{\chi^2}}$, the energy goodness-of-fit of all
models to the data simulated under the exponential, log-normal and chi-squared

models respectively. Each model was also fit to the observed data, yielding the observed goodness-of-fit of each model to the data, GOF_{obs}.

For a visual representation of this, pairwise plots of logarithms of two components of each GOF are given in Figure 7.5, alongside marginal distributions of logarithms of each component of each GOF. Logarithms are taken to make the plots more easy to visualise. Pairwise plots are used due to the difficulty of representing the three-dimensional clouds of variates from the vectors of goodnesses-of-fit. Each of GOF_{exp}, GOF_{logN} and GOF_{χ^2} are represented by a cluster of points in the pairwise scatterplots, and by a distribution function on the diagonals of this figure to represent the marginal distributions under each model. The observed value, GOF_{obs}, is represented by a black point or black line. It can be seen that the observed value appears to fit most closely to the exponential variates in all plots, indicating that the observed goodness-of-fit is more likely given that the true model is the exponential distribution, than if the true model is log-normal or chi-squared. Thus, the exponential model should be selected.

However, it is not sufficient to show that components of GOF_{obs} are, marginally, closest to components of the variates from GOF_{exp}. The components of GOF_{obs} are marginal distributions of the goodness-of-fit, and marginal distributions do not completely characterise a joint distribution. To confirm that the exponential model should be selected, linear, quadratic and mixture discriminant analyses were undertaken to classify the point GOF_{obs}. All discriminant analyses classified the observed goodness-of-fit into the exponential cluster, with probabilities 0.665, 0.998, and 0.997 for linear, quadratic and mixture discriminant analyses respectively.

Further visual representation can be provided by taking principal components of the collection of variates from each of GOF_{exp}, GOF_{logN} and GOF_{χ^2}, and projecting GOF_{obs} onto this space, providing two-dimensional projections of the goodness-of-fit space (see Section 3.1). A plot of this can be found in Figure 7.6. Figure 7.6 reinforces the selection of the exponential model. In this instance, the first two principal components capture 84.3% of the variance in the GOF distributions, meaning Figure 7.6 is sufficiently characteristic of the data to be of use in model selection.

In summary, the multi-model PBCM allows for the comparison of multiple models simultaneously, using any relevant goodness-of-fit measure. The approach is effective at distinguishing between models even at small sample sizes; here $n = 100$. While the final act of model classification is difficult to directly visualise, pairwise scatterplots and principal components plots can aid this. Classification can then be undertaken by inspection or using discriminant analyses, with discriminant analyses providing a clearer understanding of the level of uncertainty in model choice.

7.4.6   Application to the candidate models of political opinion

In the instance of the 2016 Australian Election Study model comparison, four candidate models have been chosen, with the goal of best explaining the error structure in estimating political opinion. This is because:

- error structure is of relevance to the political domain (see Section 7.2.2);
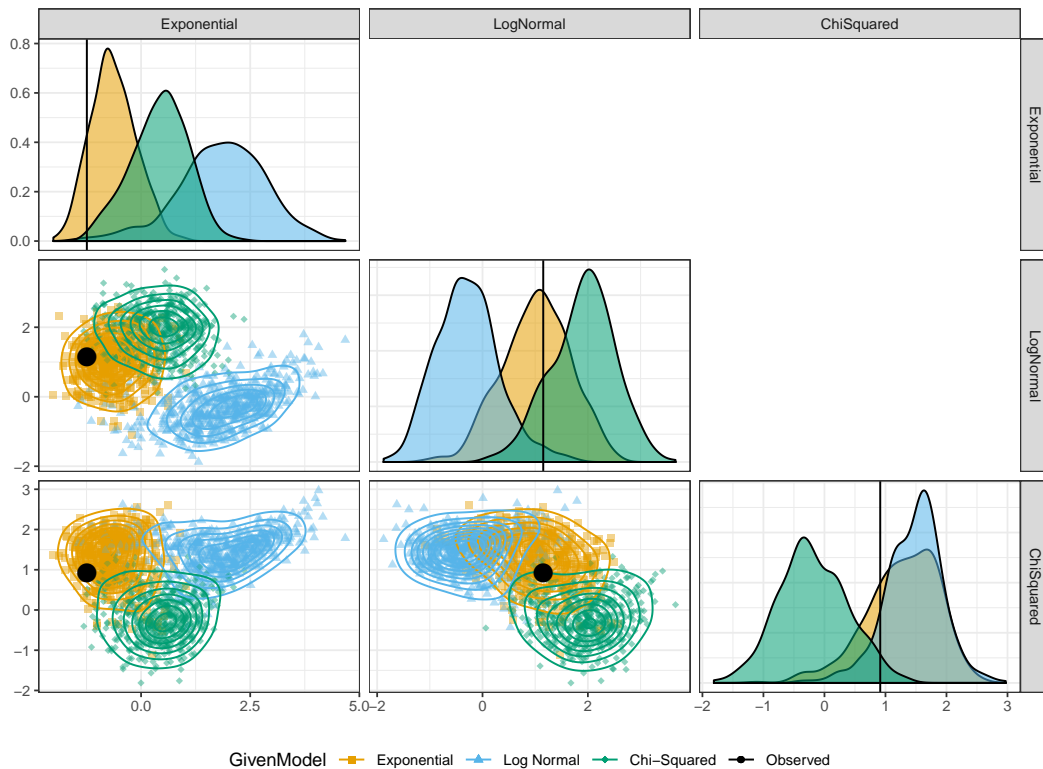
Figure 7.5: Plot comparing the fit of exponential, log-normal and chi-squared distributions to data simulated from the exponential distribution, using the multi-model variation of the PBCM. Each plot on the diagonal is of marginal, univariate densities of energy goodness-of-fit statistics to the labelled model, of data simulated from the three models as denoted in the legend. For example, the plot in the second row and second column is of how well data simulated under all three models fits to the log-normal distribution; the black line in this plot is the observed goodness-of-fit to the log-normal distribution. Scatterplots below the diagonal are of goodness-of-fit statistics to the two labelled models, of data simulated from the three models as denoted in the legend. For example, the plot in the second row and first column is of goodness-of-fit to the exponential (x-axis) and log-normal distributions (y-axis), of data simulated under all three distributions. A contour plot for each hypothesis is added to aid interpretation. The black dot in this scatterplot corresponds to the observed goodness-of-fit. The exponential distribution appears to be the best fit to the data.
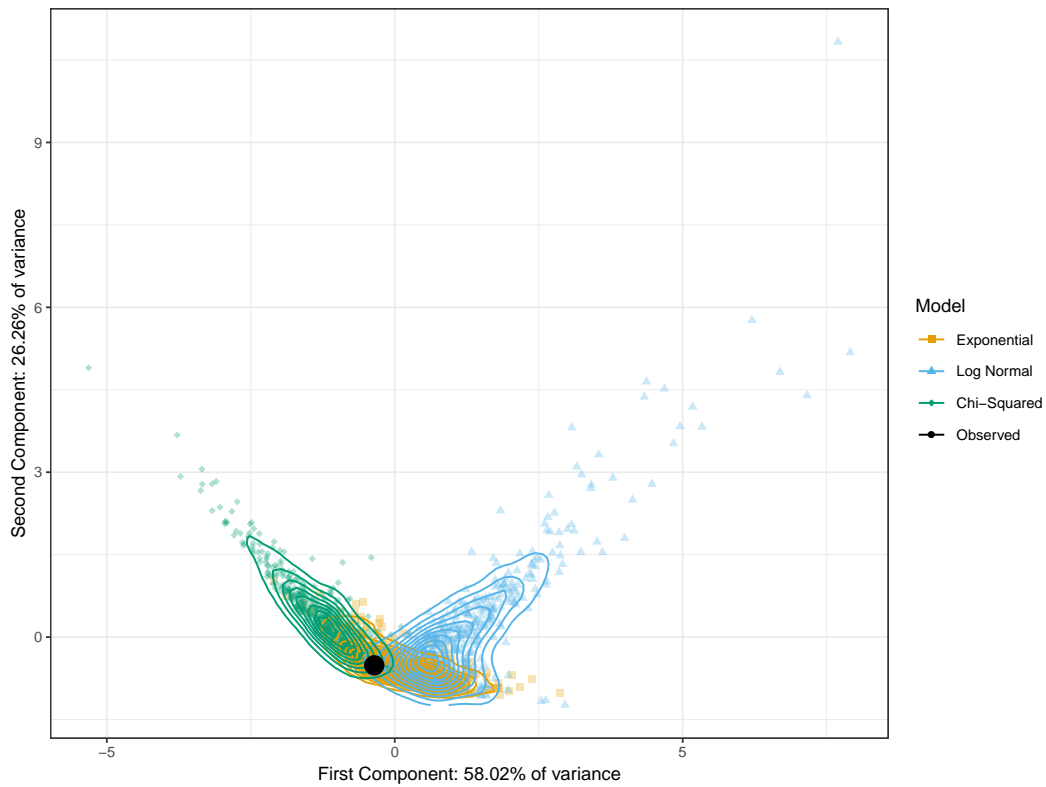
Figure 7.6: Plot comparing principal components of the fit of the exponential, log-normal and chi-squared distributions to data simulated from the exponential distribution, using the multi-model variation of the the PBCM. Each cluster of points represents data simulated under one of the hypotheses: that the data comes from an exponential (orange), a log-normal (blue) or a chi-squared distribution (green). A contour plot for each hypothesis is added to aid interpretation. Since the black point, the observed goodness-of-fit, fits in the centre of the exponential cluster, the exponential model is to be preferred.

- a model with an accurate error structure can be more confident in its error bounds; and

- a model should not be sought which does not abide by its error assumptions, as this affects the generalisability of the model.

As such, the four models need to be compared on the basis of how well the 2016 Australian Election Study data fits to the error distribution proposed by each model. The goodness-of-fit measure used to compare how the data fits to each proposed error distribution is the energy statistic [166], described in Section 5.1.2. The energy statistic uses the pairwise expected distances between the observed points and points from a proposed distribution to determine how dissimilar the observed values are to the proposed distribution. Unfortunately, the energy statistic is not measured on the same scale for all candidate distributions, so the raw energy statistic cannot directly be compared.

As the four models are not all nested, and are to be assessed on the basis of a goodness-of-fit measure whose raw values cannot directly be compared, this comparison is a prime candidate for this thesis' new probabilistic model selection approach, based on the multi-model variation of the parametric bootstrap cross-fitting method ('PBCM'), as described in Section 7.4.3.

The multi-model PBCM was applied in this instance with 1000 replicates. In other words, 1000 non-parametric bootstraps were applied to the 2016 AES data. All four models were fit to these bootstrapped data sets, producing 4000 sets of parameter estimates. For each of these models, a new synthetic data set (or 'parametric bootstrap') of the same size as the AES data set was produced, and the four candidate models were again fit to all 4000 simulated data sets. Compiling the energy goodness-of-fit statistics for all the models to each simulated data set, we have produced 4000 4×1 vectors of goodness-of-fit statistic—1000 for each of the four candidate models. These vectors are considered to be observations from the four distributions of goodness-of-fit statistics under each hypothesised model. The distributions are denoted $\boldsymbol{GOF_{multi}}$, $\boldsymbol{GOF_{rec}}$, $\boldsymbol{GOF_{mix}}$ and $\boldsymbol{GOF_{kde}}$ for the multivariate, recursive, mixture and kernel density models respectively.

Meanwhile, the goodness-of-fit statistic of each of the candidate models to the observed 2016 Australian Election Study data set was calculated, producing the vector of goodness-of-fit statistics $\boldsymbol{GOF_{obs}}$. Model comparison on the basis of goodness-of-fit to proposed error distributions is thus a problem of classifying this observed goodness-of-fit vector $\boldsymbol{GOF_{obs}}$ as belonging to one of the four distributions $\boldsymbol{GOF_{multi}}$, $\boldsymbol{GOF_{rec}}$, $\boldsymbol{GOF_{mix}}$ and $\boldsymbol{GOF_{kde}}$.

Figure 7.7 contains pairwise plots of the goodness-of-fits statistic distributions $\boldsymbol{GOF_{multi}}$, $\boldsymbol{GOF_{rec}}$, $\boldsymbol{GOF_{mix}}$ and $\boldsymbol{GOF_{kde}}$. Pairwise plots are required since each of the distributions contains four components. The figure contains four clusters of points, one for goodness-of-fit of data simulated under each hypothesised model, as well as a point representing the goodness-of-fit of the observed data. It appears from this plot that the observed goodness-of-fit matches best the distributions of goodness-of-fit under the mixture and kernel density models.
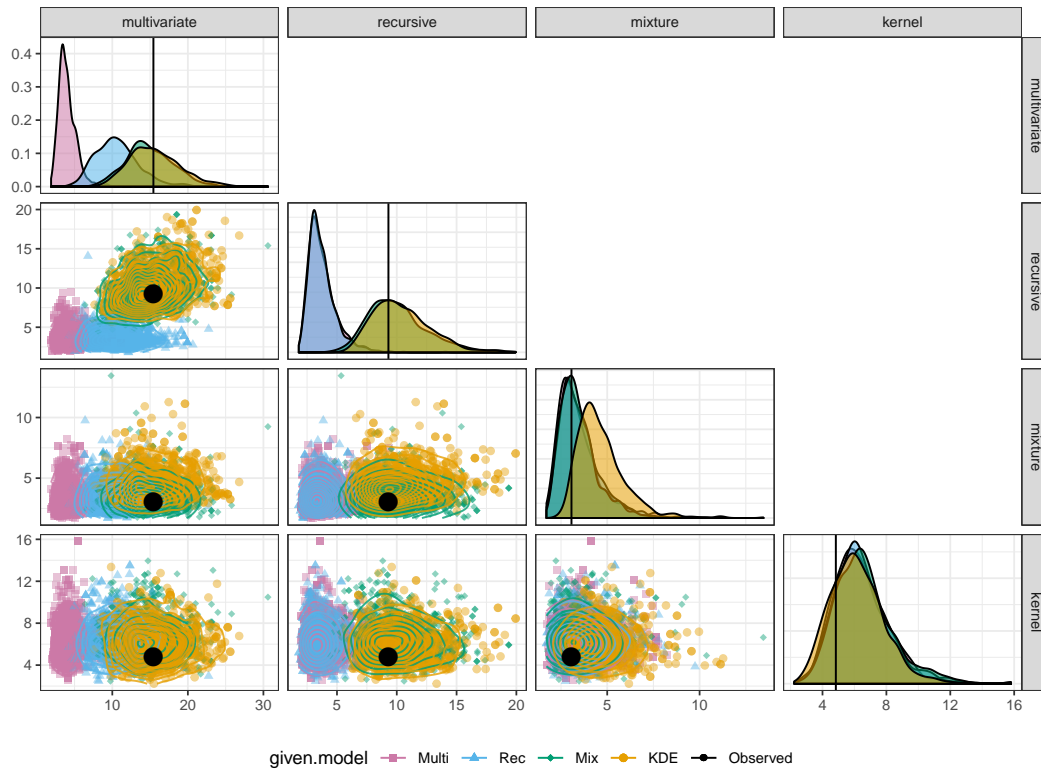
Figure 7.7: Plot comparing the fit of the multivariate, recursive, mixture and kernel density models to data from the 2016 Australian Election Study, using the multi-model variation of the PBCM. Each plot on the diagonal is of marginal, univariate densities of energy goodness-of-fit statistics to the labelled model, of data simulated from the four models as denoted in the legend. For example, the plot in the $(2, 2)$ position is of how well data simulated under all four models fits to the recursive model; the black line in this plot is the observed goodness-of-fit to the recursive model. Scatterplots below the diagonal are of goodness-of-fit statistics to the two labelled models, of data simulated from the four models as denoted in the legend. For example, the plot in the $(2, 1)$ position is of goodness-of-fit to the multivariate (x-axis) and recursive models (y-axis), of data simulated under all four models. A contour plot for each hypothesis is added to aid interpretation. The black dot in this scatterplot corresponds to the observed goodness-of-fit. The mixture and kernel density models appear to be the best fit to the data.
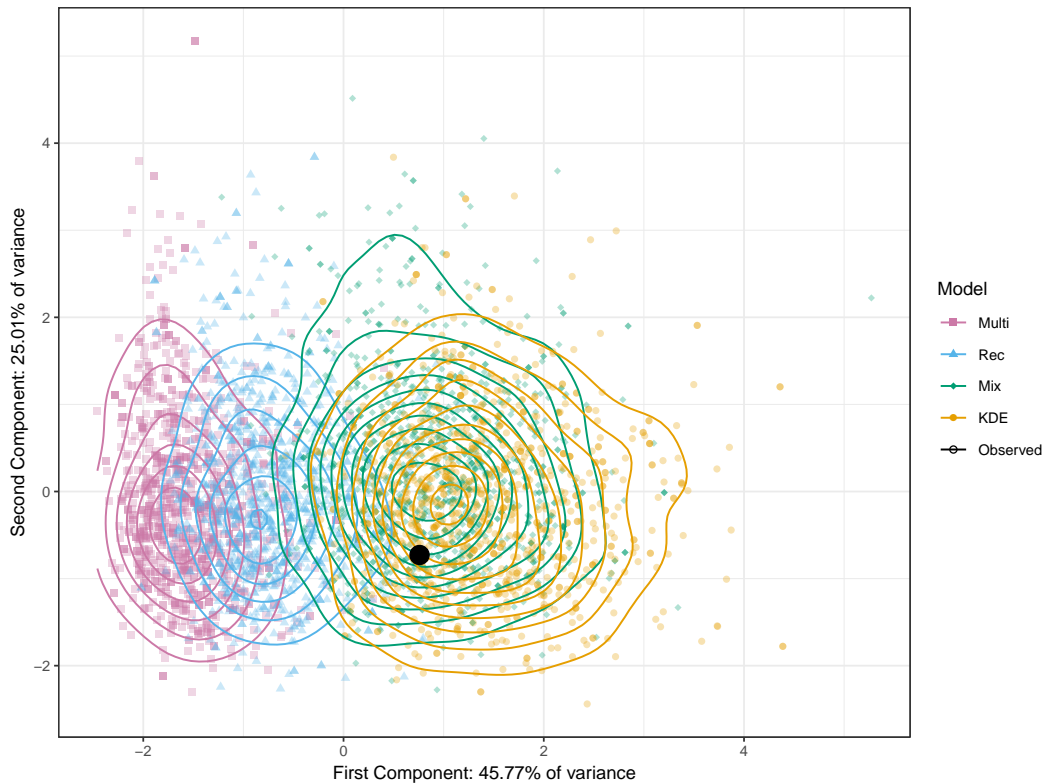
Figure 7.8: Plot comparing principal components of the fit of the multivariate, recursive, mixture and kernel density models to data from the 2016 Australian Election Study, using the multi-model variation of the PBCM. Each cluster of points represents data simulated under one of the hypotheses: that the data comes from the multivariate regression model (purple), the recursive model (blue), the mixture model (green) or the kernel density model (orange). A contour plot for each hypothesis is added to aid interpretation. Since the black point, the observed goodness-of-fit, fits close to the centre of the mixture and kernel density clusters, the mixture and kernel density models are to be preferred. A slight preference is evident for the mixture model, since the centre of the mixture cluster is slightly closer to the observed goodness-of-fit. The first principal component is best able to distinguish the four clusters.

	Multivariate	Recursive	Mixture	Kernel density
LDA	2.72×10^{-5}	6.49×10^{-3}	0.662	0.332
QDA	1.05×10^{-24}	6.92×10^{-6}	0.655	0.345
MDA	1.45×10^{-14}	2.11×10^{-4}	0.788	0.212

Table 7.2: Comparison of the four candidate models in their flexibility to produce data that behaves similarly to that from the 2016 Australian Election Study, as compared using the multi-model parametric bootstrap cross-fitting method. Simulated goodness-of-fit vectors under each of the four alternative hypotheses were compared to the observed goodness-of-fit. Using linear discriminant analysis ('LDA'), quadratic discriminant analysis ('QDA') and mixture discriminant analysis ('MDA'), the relative likelihood of the observed goodness-of-fit under each candidate model can be compared. The method concludes that the residuals were most likely to have come from an error distribution described by the mixture or kernel density models. This comparison provides strong evidence that the multivariate and recursive models cannot produce data that performs similarly to the 2016 AES, while the mixture and kernel density models perform much more strongly. The mixture model emerges as the most favoured on the basis of its fidelity to assumptions.

This is reinforced by considering the principal components plot of the goodness-of-fit distributions, as represented in Figure 7.8. In this instance, the first two principal components capture a combined 70.8% of the variance in the $\boldsymbol{GOF}$ distributions, meaning Figure 7.8 provides a good visual representation of the $\boldsymbol{GOF}$ distributions. The principal components plot also suggests that the observed goodness-of-fit is closest to the mixture and kernel density model, suggesting that these models, more than the other candidates, are able to produce data that behaves similarly to the 2016 Australian Election Study data.

Looking beyond comparison by inspection, the models can be compared using discriminant analyses, as described in Section 7.4.4. Using linear, quadratic and mixture discriminant analyses, we can produce estimates of the probability the observed goodness-of-fit belongs to each of the candidate models. Table 7.2 contains these estimated probabilities. This table reinforces the conclusion that the models best able to reproduce the desired error distributions are the mixture and kernel density models. All discriminant analyses express a strong preference for the kernel density and mixture models, over the recursive and multivariate regression models. Across the discriminant analyses, a preference for the mixture model is evident, with probabilities 0.662, 0.655 and 0.788 for linear, quadratic and mixture discriminant analyses respectively. This is substantially more than for alternative models, but we also cannot discount the kernel density model, whose probability lies between 0.212 and 0.345 depending on the discriminant chosen. As a result, no strong conclusion can be drawn as to the preferred model using the multi-model variation of the parametric bootstrap cross-fitting method alone.

Overall, as expressed by inspection of pairwise and principal components plots,

and as numerically expressed by this thesis' favoured approach of linear, quadratic and mixture discriminant analyses, the multi-model variation of the parametric bootstrap cross-fitting method favours the flexibility of the mixture and kernel density models over the recursive and multivariate regression models. This method demonstrates that the mixture and kernel density models are the candidates best able to reproduce data similar to that found in the 2016 Australian Election Study.

7.5 Model selection conclusions for the AES

This chapter has explored five bases for comparing the non-nested candidate models. These bases are:

- Cross-validated accuracy of prediction regions;

- Size of prediction regions;

- Parsimony;

- Relevance of the model to its domain; and

- Goodness-of-fit to model assumptions.

Since all of these bases are indicative, and not determinative, of the choice of model, the relative performance of each model with respect to each criterion needs to be weighed in order to determine the most appropriate model to describe the 2016 Australian Election Study data set.

With this in mind, Table 7.3 describes the performance of each model, with respect to each criterion, with a simple score based on the conclusions of the relevant sections of this chapter. Each model is rated as 'good', 'acceptable', or 'poor' in each area, meaning the models can be compared across the relevant model selection criteria.

Table 7.3 leads to two key conclusions. The first is a clear demonstration of the relationship between model complexity and size of prediction regions; as the candidate models become more complex, they are able to produce smaller cross-validated prediction regions, with no loss of accuracy. The lack of loss of accuracy in cross-validation reduces overfitting concerns, since the manifestation of overfitting is often in a lack of generalisability, leading to overconfidence in estimates. Cross-validation is able to assess, to some extent, a model's generalisability [84].

The second, more important conclusion is the clear preference exhibited for the mixture model. The mixture model performs best in all categories except prediction region size. Even with respect to prediction region size, the mixture model does not perform poorly. The preference for the mixture model over the kernel density model, the only model that outperforms the mixture model in any area, is particularly clear when considering that the model sought here is primarily an explanatory one. Since a model is sought with the aim of describing relationships between demographic and political identities, it is particularly important that the chosen model is relevant in this domain. Since the kernel density model performs poorly in this area, while the

	Multivariate	Recursive	Mixture	Kernel density
PR Accuracy	Good	Good	Good	Good
PR Size	Poor	Acceptable	Acceptable	Good
Parsimony	Good	Good	Good	Poor
Relevance	Acceptable	Good	Good	Poor
Assumption fit	Poor	Poor	Good	Good

Table 7.3: Comparison of the four candidate models from the 2016 Australian Election Study, in all measures described in Chapter 7. Models are rated 'poor', 'acceptable' or 'good' in each of the criteria: prediction region accuracy, prediction region size, parsimony, relevance to its domain and goodness-of-fit to assumptions. Cells are coloured for ease of interpretation. Overall, it appears that the mixture model should be preferred in all situations except when prediction region size is the dominant consideration.

mixture model performs well, it is to be preferred. The output from this mixture model will thus be used for the discussion chapter of this thesis, and more generally for reaching conclusions as to the demographic influences on political opinion.

This chapter has explored different ways of comparing the candidate models for describing the 2016 Australian Election Study dataset. The comparison methods have ranged from highly subjective to highly objective, and have focused on the desired applications of the model—to explain, simply and coherently, the relationships between demographic and political identity, without being overconfident in predictions. All predictions of continuous political ideology, based upon demographic identity, will be incorrect. It is thus important to understand the directions in which they are more likely to be incorrect.

This desire for an ability to explain error accurately has led to the development of techniques for choosing models based upon how well the desired models adhere to their assumptions. This has further led to discussion of the literature on non-nested model comparison, and the extension of current model comparison techniques to comparing multiple non-nested models on the basis of goodness-of-fit to model assumptions. This has also led to new numerical strategies for classifying data as belonging to a particular model on this goodness-of-fit basis, applying discriminant analyses to outputs from the multi-model variation of the parametric bootstrap cross-fitting method.

Using the techniques explored and developed in this chapter, a preference for the Gaussian mixture candidate model became evident. Chapter 8, which follows this, will describe the outputs of this model, and the key relationships between demographic and political identity that arise from the mixture model. Chapter 8 will also explain the shortcomings of models estimating political opinion, and how the mixture model can use the errors of estimates themselves to further describe relationships in political thought.

Chapter 8

Results in Context

Chapter 4 used dimension-reduction techniques to demonstrate that demographic and political identity, as expressed in the 2016 Australian Election Study ('AES'), can be reduced to a small number of key factors. For the purpose of building a model connecting demographic and political identity, the first three dimension-reduced political variables are of interest. We have labelled these:

1. 'Social inclusivity';

2. 'Attitudes to authority'; and

3. 'Spending priorities'.

Meanwhile, the first three dimension-reduced demographic variables are:

1. 'Socio-economic status and education';

2. 'Stage of life'; and

3. 'Cultural background'.

After the third axis, the dimension-reduced demographic variables have less clear-cut interpretations (see Chapter 4), but are nonetheless included in the models built in Chapter 6. These would still be useful in a predictive context, but do not impact other terms in each model; as all the dimension-reduced demographic variables are orthogonal, adding or removing one of the variables in the candidate linear models does not change the other variables' coefficients. Since we are describing this model in an explanatory, rather than a predictive, context, they are not discussed in this chapter.

In Chapters 5 and 6, four candidate models for describing the relationship between the dimension-reduced demographic variables and the dimension-reduced political variables were introduced. These are the multivariate regression model (Chapter 5), the recursive model (Section 6.1), the Gaussian mixture model (Section 6.2) and the kernel density model (Section 6.3).

In Chapter 7, it was decided that of the four introduced candidate models, the Gaussian mixture model was optimal. This was as it provides the best balance of

prediction region accuracy and size, parsimony, domain-relevance and fidelity to its assumptions.

In Section 8.1, the conditional expectations of political ideology given by the mixture model will be discussed in detail, response-by-response, in the context of previous literature on relationships between demographic and political identity. The error terms of the mixture model will be discussed in context in Section 8.2, to explain both the limitations of the model, and how these limitations in themselves provide insight into relationships in political thought. This will involve a discussion of different kinds of political identities that cannot be related to demography.

8.1   Conditional expectations of political ideology

As first stated in Section 6.2.4, the mixture model developed to connect demographic variables $\{X_1, X_2, \ldots, X_{50}\}$ to political variables $\{Y_1, Y_2, Y_3\}$ is

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 2.79 & -1.52 & 0.27 & 0.31 & 4.62 & 2.19 & -1.39 & -2.37 & -0.84 & -1.55 \\ -0.31 & -3.03 & -0.93 & -1.82 & -1.21 & 1.58 & -0.98 & -0.27 & -1.66 & -0.03 \\ -2.03 & -0.40 & -0.41 & -1.76 & -0.03 & -0.19 & 0.62 & 0.78 & -0.40 & 0.44 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_7 \\ X_9 \\ X_{10} \\ X_{11} \\ X_{20} \end{bmatrix} + \varepsilon_{mix},
$$

$$(8.1.1)$$

with the parameters of ε_{mix} to be identified and explained in Section 8.2. This section will explain what this model's point predictions, or conditional expectations, in terms of the first three dimension-reduced demographic variables, can tell us about each of the three dimension-reduced political variables, and how this relates to existing political science literature. While there are large coefficients relating to other dimension-reduced demographic variables, these variables do not have a clear social science interpretation, so will not be considered here.

8.1.1   Influences on social inclusivity

A high score on the social inclusivity axis, as explained in Section 4.4, is characterised by opposition to policies such as turning back boats containing asylum seekers, and responses in favour of increased support for Aboriginal and Torres Strait Islander peoples.

Using Equation (8.1.1), the marginal conditional expectations of individuals' social inclusivity can be written as

$$Y_1 = 2.79X_1 - 1.52X_2 + 0.27X_3 + 0.31X_4 + 4.62X_5 + 2.19X_7$$

$$- 1.39X_9 - 2.37X_{10} - 0.84X_{11} - 1.55X_{20}.$$

In other words, there is a substantial positive relationship between the demographic variable 'socio-economic status and education' (X_1) and social inclusivity, with a coefficient of 2.79. There is also a substantial negative relationship between the demographic variable 'stage of life' (X_2) and social inclusivity, with a coefficient of -1.52, and a small positive relationship between 'cultural background' (X_3) and social inclusivity, with a coefficient of 0.27.

Socio-economic status, education and social inclusivity With a coefficient of 2.79, the mixture model's conditional expectations indicate that there is a substantial positive relationship between socio-economic status, education and social inclusivity; those with a higher value of X_1, the socio-economic status and education axis, are more likely than the average Australian to hold socially inclusive views. Recalling from Section 4.3 that higher values of X_1 indicate an increased likelihood of having completed tertiary education, having a high income and being currently employed, we can conclude that on average, those with higher education and socio-economic status are more likely to be socially inclusive.

This conclusion also aligns with results in political science literature. Moore, writing in the context of the United States in 1986, notes that increased levels of education and income are associated with more inclusive views when it comes to immigration, noting that "the less threatened a person is economically, [...] the more favourably he or she will feel are immigration's consequences" [129]. In an Australian context, Bean in 2000 noted that support for anti-immigration party One Nation was lowest among "the affluent professional classes with tertiary education" ([20], see also [73]).

Further, using a subjective construction of an "attitudes to immigration" measure, and using data from previous Australian Election Studies, Bilodeau and Fadol found in 2011 that attitudes to immigration are more positive among people with university education [27]. This thesis builds on the work of Bilodeau and Fadol in two ways: we use a data-driven "social inclusivity" axis, rather than a subjective measure; and we associate not just education and immigration, but socio-economic status and education more broadly with a wider range of measures relating to social inclusivity, including attitudes to Aboriginal and Torres Strait Islander people and attitudes towards LGBT+ people.

In terms of explaining the association between socio-economic status, education and social inclusivity, Bilodeau and Fadol propose a twofold argument [27]. Firstly, more educated people are less likely to rely on stereotypes to make political assessments, because they have been taught "the value of cultural diversity". Secondly, people who are economically insecure might be afraid that assisting groups of which they are not a member might displace them from their already precarious social standing. This argument is reiterated by Moore [129].

Stage of life and social inclusivity The mixture model's conditional expectations also indicate that there is a substantial negative relationship between 'stage

of life' and social inclusivity, with a coefficient of -1.52. The stage of life variable, X_2, is associated with age, and markers relating to age, with a high score indicating a subject is more likely to be over the age of 65, be retired, own shares, and own their own home, with higher values of X_2 indicating an individual is older, or is otherwise demographically similar to older people. This means that people who are demographically similar to older people are more likely to be less socially inclusive than the average person.

This conclusion also aligns with results in political science literature, and in previous issue-by-issue polling. In polling for responses to univariate immigration questions, Chandler and Tsai found that older respondents to the United States General Social Survey were significantly more likely to harbour anti-immigration viewpoints [40]. A similar conclusion was drawn regarding Western European views by O'Rourke [140].

With regard to other social inclusivity issues, opinion polling leading up to Australia's Marriage Law Postal Survey consistently showed substantially lower support for reform among older Australians [48, 56, 130, 184]. On issues relating to Aboriginal and Torres Strait Islander peoples, polling by Essential Research has shown support for an indigenous 'voice to parliament' is highest among younger Australians [56] and polling by the Australian National University has shown support for constitutional recognition of indigenous people is highest among younger Australians [86].

This thesis statistically confirms the previous research, showing that the underlying relationship between stage of life and social inclusivity exists on a large scale and with statistical significance. While previous research has shown that particular realisations of this relationship exist, we go further by demonstrating the existence of a relationship between the underlying demographic dimension relating to one's 'stage of life' and the political dimension underlying a wide range of social views.

Cultural background and social inclusivity   The mixture model's conditional expectations also indicate that there is a small positive relationship between 'cultural background' and social inclusivity, with a coefficient of 0.27. High scores on the cultural background variable, X_3, are characterised by the subject having been born outside Australia, and having parents born outside Australia. This means that people who are born outside Australia are slightly more likely to be more socially inclusive than the average person.

It seems initially surprising that those who have migrated from overseas are not substantially more accepting of, say, asylum seekers arriving by boat, than the average person, but this is partially substantiated by previous literature in the area. One example of this is in a 1997 comparison of Mexican-Americans and Anglo-Americans living in Southern Texas, with respect to their views on immigration coming from Mexico [28]. While views relating to unlawful immigration differed between the two groups, the groups' views on levels of legal immigration were not substantially different, and general opposition to immigration was similar between more integrated Mexican-American people, and Anglo-Americans [28]. In the Australian Election Study data, we observe that there is similarly a slightly

higher level of socially inclusive views among people who were not born in Australia, but place of origin is not as strong a determinant of political perspective in this regard as is socio-economic status and education, or stage of life.

This can be further explained by the difference in views relating to other issues of social inclusivity between Australians born in Australia, and Australians born outside Australia. For example, an analysis of results from Australia's Marriage Law Postal Survey by the Guardian demonstrated that electorates with higher levels of migrants were slightly more likely to oppose marriage equality reform [114].

Overall, the relationship between place of origin and social inclusivity is complex. While a linear relationship exists between the variables, in order to shine a light on more complex determinants of political thought, additional intersectionalities in identity may need to be considered.

8.1.2 Influences on attitudes to authority

A high score on the attitudes to authority axis, as explained in Section 4.4, is characterised by low levels of confidence in the federal government in Canberra, in banks and corporations, and the belief that democracy is not working in Australia. Using Equation (8.1.1), the marginal conditional expectations for individuals' attitudes to authority can be written as

$$Y_2 = -0.31X_1 - 3.03X_2 - 0.93X_3 - 1.82X_4 - 1.21X_5 + 1.58X_7$$
$$- 0.98X_9 - 0.27X_{10} - 1.66X_{11} - 0.03X_{20}.$$

In other words, there is a small negative relationship between the demographic variable 'socio-economic status and education' (X_1) and attitudes to authority, with a coefficient of -0.31. There is also a substantial negative relationship between the demographic variable 'stage of life' (X_2) and attitudes to authority, with a coefficient of -3.03, and a negative relationship between 'cultural background' (X_3) and attitudes to authority, with a coefficient of -0.93.

Socio-economic status, education and attitudes to authority With a coefficient of -0.31, the mixture model's conditional expectations indicate that there is a small negative relationship between socio-economic status, education and attitudes to authority; those with a higher value of X_1, indicating higher socio-economic status and education, are slightly more likely than the average Australian to hold views trusting of government, law enforcement and corporations. This finding has also been observed in social science literature, with Gauld *et al.* collecting data showing Australians who trust government have slightly higher likelihood of being well-educated [72].

From the perspective of political science theory, Newton *et al.* note in a US context in 2018 that more educated people have no need to "distrust social arrangements that have served them well," while noting that at the other end of the spectrum, "anxiety and insecurity are the most powerful forces driving distrust among the poor" [106]. In other words, those who benefit from existing power structures are more likely to have faith in these structures, while those who are left out are more

likely to seek change. Newton *et al.* also cite empirical evidence from Paxton, who finds, using worldwide data, that additional years of education increase political trust [146].

Looking more directly at socio-economic status, rather than education, the picture of the relationship between the first dimension-reduced demographic axis, and trust, becomes less clear. While Newton *et al.* find that wealthier people are more likely to express trust in a general sense [106], Brehm and Rahn use data from the US General Social Survey to suggest that wealthier people have higher levels of social trust, but lower levels of political trust [33]. In other words, wealthier people are more likely to trust individuals—their neighbours, colleagues and strangers—while being less likely to trust government. Hibbing and Theiss-Moore even go so far as to say that "dissatisfaction is concentrated among people who are involved with and who have benefited from the system" [89].

This confusion in the literature is replicated in this thesis' results from the 2016 Australian Election Study. The literature's suggestion that education leads to higher trust in authority, while wealth leads to slightly lower trust in authority, confounds in the 2016 AES to lead to the general conclusion that there is a small negative relationship between socio-economic status and distrust in authority. In doing so, the AES dispels the notion that distrust is unique to people with lower levels of income and education, with no strong association in either direction.

Stage of life and attitudes to authority   The mixture model's conditional expectations demonstrate a large negative relationship between 'stage of life' and attitudes to authority, with a coefficient of -3.03. Thus those with higher values of X_2, people who are over 65, own shares, or are retired, are much more likely to have high levels of trust in institutions.

That people of a later stage of life are more likely to be trusting in general is strongly supported by political science literature. Paxton demonstrates this using early data from the World Values Survey, though with a measure of trust that was not empirically defined [146]. Paxton states that it is unclear whether this is a direct effect of age, or a cohort effect attached to generations; she notes that older generations are more trusting, but it is unknown whether younger generations will become more trusting as they get older.

This thesis, looking holistically at both age and factors relating to age, might shed some light on one possible cause of the relationship between age and trust in institutions. A high score on the dimension-reduced demographic variable 'stage of life' is associated with owning ones own home, and owning shares in corporations, among other factors. By forging positive relationships with corporations in purchasing shares, and by forging positive relationships with banks in having paid off their homes, a larger proportion of older people have reasons to trust banks and other large corporations. In the context of Newton *et al.*'s previously discussed arguments, older people might trust organisations that "have served them well" [106], while younger people might have less reason to build up trust.

A notable distinction emerges between those who score highly on the 'socio-economic status and education' variable and the 'stage of life' variable. While the

former group captures people who have higher current income, and are currently employed, the latter describes those who might have a lesser need for a high income, education or employment, since they are more likely to have lower costs and greater assets (see Chapter 4). The AES data demonstrates that those who have built strong relationships with corporations, but no longer strictly rely on these corporations for their economic stability, on average have much more social and political security than those who currently benefit from existing power structures but are more likely to be impacted by them. Through this lens, the results of the 2016 AES align quite closely with Newton *et al.*'s argument that distrust relates closely to insecurity [106].

Cultural background and attitudes to authority   The mixture model's conditional expectations also demonstrate a negative relationship between 'cultural background' and attitudes to authority, with a coefficient of -0.93. People who are born outside Australia or who have parents born outside Australia, those with higher values of X_3, are more likely to have high levels of trust in institutions like the federal government, big banks and corporations.

This conclusion aligns with previous research on the subject. André finds that migrants to the European Union have, in general, higher levels of political trust [6], while Maxwell finds that migrants to the UK on average have higher levels of satisfaction in government [119]. Both of these analyses relied on single measures of trust in government, while this thesis expands on this by developing a data-driven measure of political trust, and in doing so more rigorously validates the research of André and Maxwell in an Australian context.

The notion that first and second generation migrants to Australia on average have higher levels of trust in authority is also intuitively well-grounded. Most migrants to Australia migrate because they either are dissatisfied with the country in which they previously lived, or because they value opportunities brought about by their new community. Maxwell argues that this makes migrants predisposed to "have positive evaluations of host society institutions." Put simply, people are more likely to trust institutions in a country they actively elect to live in.

This thesis demonstrates that the predisposition among migrants to trust institutions is higher even when considering a response variable that includes not just trust in political institutions, but also corporate institutions.

8.1.3   Influences on spending priorities

A high score on the 'spending priorities' axis, as explained in Section 4.4, is characterised by favouring high levels of spending on health, the National Disability Insurance Scheme, pensions and law enforcement, while being less likely to prioritise taxation, or management of the economy, as a political issue. Using Equation (8.1.1), the marginal conditional expectations for individuals' spending priorities can be written as

$$Y_3 = -2.03X_1 - 0.40X_2 - 0.41X_3 - 1.76X_4 - 0.03X_5 - 0.19X_7$$
$$+ 0.62X_9 + 0.78X_{10} - 0.40X_{11} + 0.44X_{20}.$$

In other words, there is a substantial negative relationship between the demographic variable 'socio-economic status and education' (X_1) and spending priorities, with a coefficient of -2.03. There is also a negative relationship between the demographic variable 'stage of life' (X_2) and spending priorities, with a coefficient of -0.40, and a negative relationship between 'cultural background' (X_3) and spending priorities, with a coefficient of -0.41.

Socio-economic status, education and spending priorities   Our model's conditional expectations demonstrate a substantial negative relationship between 'socio-economic status and education' and spending priorities, with a coefficient of -2.03. Thus subjects with higher values of X_1, people with higher incomes and education, or otherwise have a higher socio-economic status, are more likely to prioritise lowering taxes over increasing government spending.

This association is not particularly new or noteworthy; people who pay the most tax naturally care more about their tax burden. On the other hand, those who rely on government services are more likely to favour increased social spending. These two theories find support in the work of Kaltenthaler and Ceccoli [105], Andersen and Curtis [4], and Schneider and Jacoby [155], in Western Europe, Canada and the United States respectively. In another example, Henderson *et al.* find that even adjusting for partisan orientation and gender, Americans of lower income are more likely to favour increased spending on child care [87].

The interesting development presented in this thesis is that the 'self-interest' motivation cited by Andersen [4] outweighs the more interventionist economic values seemingly common among people with more educational experience [124]. Here, the first dimension-reduced demographic axis, which is negatively correlated with views favouring government economic intervention, incorporates both socio-economic status, and levels of education. While socio-economic status is, in political science literature, associated with support for lower levels of taxation and social spending, there is broad agreement in the literature that higher levels of education are associated with support for social equality [124]. However, it appears that this support for social equality among more educated people does not extend to, say, favouring increased support for redistribution to marginalised people or increased spending on health and education. This conclusion is reiterated in the work of Hasenfeld and Rafferty [83] and Linos and West [113].

This thesis thus highlights the distinction between the general tendency among the professional classes to support measures promoting social inclusivity (the first dimension-reduced political axis), while opposing increased social spending (the third). This finding reiterates the importance of decomposing political opinion into social and economic axes, and has ramifications for the framing of policy proposals. For example, suppose one is attempting to promote a policy of increased spending in a particular area. Arguments that the policy might reduce social inequalities affecting marginalised groups might be more successful in convincing those of high socio-economic status and education, while arguments based around class might be more successful among people of a lower socio-economic status and education. This is because progressive views as to social inclusivity are more common among those

of high socio-economic status and education, while this same group of people is less likely to favour social spending in a general sense.

Stage of life and spending priorities   The mixture model's conditional expectations demonstrate a negative relationship between 'stage of life' and spending priorities, with a coefficient of -0.40. This means that people whose X_2 values are higher, that is, people who are older, or share demographic features common among older people, are slightly more likely to prioritise lowering taxes over increasing government spending.

As described in political science literature, this is a result of the confluence of two influences on political thought—an association between age and conservatism, and self-interest. In general, older people, or those with the characteristics of older people, are more likely to have economically conservative views. For example, Zagorski finds this effect in Australia in the mid-1980s [185], while Busemeyer *et al.* find that across the OECD, older people are more likely to favour less spending on education [36].

In this study, the effect of age on conservatism is mitigated by the fact that the 'taxes and spending' axis includes questions that specifically relate to spending that disproportionately benefits older Australians. The third dimension-reduced political axis includes questions relating to spending on old-age benefits, as well as health care. When it comes to public spending that aids older people, people of a more advanced stage of life are known to, on average, value this spending, even at the cost of possible tax increases. For example, Busemeyer *et al.* find that across the OECD, older people are more likely to favour higher spending on pensions, even when told this "might require a tax increase to pay for it" [36].

Thus, when benchmarking fiscal conservatism of older people in general, it is important to weigh up both the self-interest in favouring higher spending in some areas, with the desire for lower taxes. In producing a small, negative coefficient for the relationship between 'stage of life' and spending priorities, the 2016 AES data exhibits this balance.

Cultural background and spending priorities   Conditional expectations from the mixture model demonstrate a negative association between 'cultural background' and spending priorities, with a coefficient of -0.41. People who are born outside Australia, or whose parents are born outside Australia, are slightly more likely to be conservative when it comes to government spending, than the average Australian.

There is some support for this in social science literature. For example, Dancygier and Saunders, studying Germany and the United Kingdom, find that "immigrants in both countries appear less willing to spend more on social services" [50]. However, other research by Eger demonstrated no statistically significant relationship in Sweden between the country subjects were born, and whether they favoured increasing taxes and spending [55]. This research suggests only a small negative relationship, aligning somewhere in the middle of the work of Dancygier and Saunders, and Eger [50, 55].

8.2   Errors in estimating political ideology

No model estimating political opinion can be perfect; in fact, it is reassuring to the author of this thesis that political opinion is not socially determined to the extent that it can be explained wholly by a number of demographic influences. That the mixture model chosen in this thesis describes much less of the variance in the data than it leaves undescribed, is a consequence of the conscious decision to make an explanatory model, describing dynamics that exist in the population at large, rather than a predictive one. The existence of a substantial error term both leaves room for further research, and reassures us that political opinion is more complex than a function of a limited number of our experiences.

The statistician George Box wrote a number of times that "all models are wrong, but some are useful" [29, 30, 31, 32]. As shall demonstrate, the way in which our model is wrong is in itself useful, for it explains otherwise ignored relationships in political thought, and describes dynamics affecting the polarisation of our political system.

As briefly explained in Section 7.2.2, the error in estimating political opinion is itself informative, as it allows for the discussion of non-linear relationships between the orthogonal dimension-reduced political variables, which cannot be explained by demography alone. For example, in Chapter 5, it was seen that demography cannot explain the non-linear relationship between social inclusivity and trust in authority, in which more extreme social views are associated with distrust of authority. This section explains relationships such as this in more detail, with reference to their political science underpinnings.

8.2.1   The relationship between errors in estimating social inclusivity and trust in authority

In Section 6.2.1, the ordinary least squares procedure was applied to the dimension-reduced political variables, against the dimension-reduced demographic variables. Fitting a Gaussian mixture to the residuals from this procedure, three clusters of the error term emerged. There is substantial overlap between the three clusters, since respondents will not always clearly belong to one of three disjoint clusters, but as found in Section 6.2.1, all three are necessary to describe the shape of the error. The three clusters, with proportions 0.382, 0.209 and 0.409, have means

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -1.096 \\ -0.643 \\ -0.145 \end{bmatrix},$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} -1.908 \\ 0.561 \\ 0.468 \end{bmatrix}, \text{ and}$$

$$\boldsymbol{\mu}_3 = \begin{bmatrix} 2.000 \\ 0.315 \\ -0.104 \end{bmatrix},$$

and respectively have variance matrices

$$\Sigma_1 = \begin{bmatrix} 5.096 & -0.301 & -0.827 \\ -0.301 & 7.212 & -1.227 \\ -0.827 & -1.227 & 3.687 \end{bmatrix},$$

$$\Sigma_2 = \begin{bmatrix} 9.464 & -3.980 & -2.555 \\ -3.980 & 10.523 & 3.053 \\ -2.555 & 3.053 & 12.192 \end{bmatrix}, \text{ and}$$

$$\Sigma_3 = \begin{bmatrix} 8.008 & 0.907 & -0.550 \\ 0.907 & 6.196 & 0.828 \\ -0.550 & 0.828 & 2.670 \end{bmatrix}.$$

Recall that these three clusters represent the error in estimating the first three dimension-reduced political variables, using the dimension-reduced demographic variables. To remind the reader, the first dimension-reduced political variable is 'social inclusivity', with high scores representing high social inclusivity. The second dimension-reduced political variable is 'attitudes to authority', with high scores representing distrust in authority. The third dimension-reduced political variable is 'spending priorities', with high scores representing higher taxes and levels of government spending. The three clusters are, marginally, visually represented in Figures 8.1 and 8.2.

Looking then at the means of the three error clusters, we can note that the second and third clusters represent groups of people who are on average, more distrusting of authority than their conditional expectations as determined by their demographic characteristics, while the first represents a group of people who is on average more trusting of authority than expected. The two groups more distrusting of authority than expected have means located at errors in social inclusivity of -1.908 and 2.000, while the more trusting group has mean at a social inclusivity of -1.096. This leads to two key conclusions: that people who are more distrusting of authority than expected are more likely to also harbour more extreme views than expected when it comes to social inclusivity, and that the mean of the group most likely to trust authority is slightly less socially inclusive than the average person, relative to that expected. These conclusions are discussed in turn in the following paragraphs.

We note that the demographic associations described in Section 8.1 do not account for anywhere near the majority of the variation in the dimension-reduced political variables, meaning the political variables are dominated by the error terms. As such, our conclusions about associations within the error terms largely extend to associations within political opinion.

Extremity Since people who are less trusting of authority than the model expects are also more likely to have views either much more socially inclusive or much less socially inclusive than the model expects, it can be said that there is an association between extremity in social inclusivity and disenchantment with current power structures. This is visually represented by Figure 8.1. This is an intuitive result; it makes sense that people who seek substantial change with respect to social
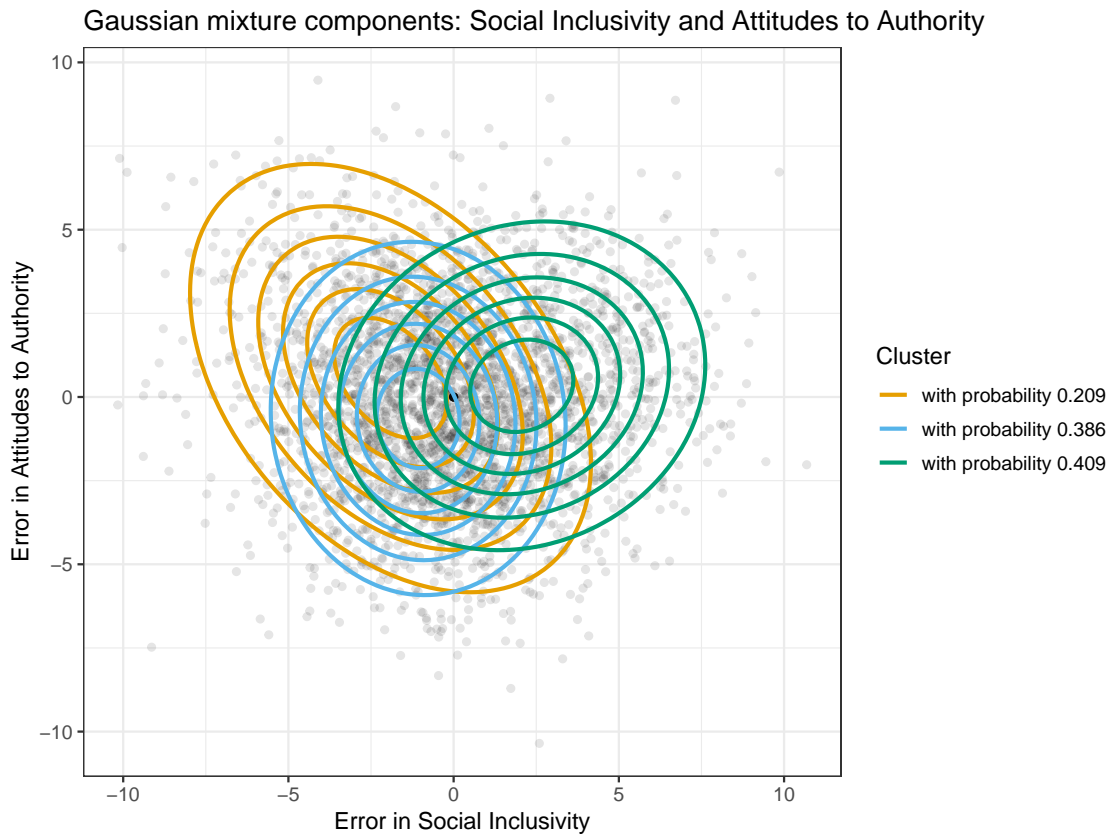
Figure 8.1: Plot of the residuals of the mixture model, with the three components of the Gaussian mixture overlaid. The first two dimension-reduced political axes, social inclusivity and attitudes to authority, are shown. The orange cluster captures respondents who are more distrustful of authority than expected, and less socially inclusive, while the green cluster captures respondents who are more distrustful of authority than expected, and more socially inclusive. The blue cluster captures respondents who are more trustful of authority, and this cluster has a mean slightly less socially inclusive than expected. This plot thus demonstrates the non-linear relationship between the error in social inclusivity, and error in trust in authority.

Figure 8.2: Plot of the residuals of the mixture model, with the three components of the Gaussian mixture overlaid. The third and second dimension-reduced political axes, spending priorities and attitudes to authority, are shown. No strong relationship between these error terms can be ascertained; while the greatest positive error in attitudes to authority, in the orange cluster, contains the subjects most economically conservative relative to expectations, this cluster also has the most economically progressive mean.

inclusivity policy are likely to have less faith in institutions which preserve the status quo. For example, an individual who strongly believes that asylum seekers should not be processed offshore would be more likely to have little faith in a parliament in which offshore processing has bipartisan support. Meanwhile, someone who takes a centrist view to issues relating to social inclusivity might have more confidence in institutions which align with their views.

This perspective is shared with previous social science research. For example, Miller suggests that in the US context, political distrust is not confined to one particular set of policy platforms; he writes that "no monolithic description of cynicism arises out of policy preferences" ([127], see also [44, 96]). Miller writes that instead, we can describe three groups of political thought: "cynicism of the left," "cynicism of the right," and "those in the center," who he describes as the "least cynical." This is a similar finding to that arising from the 2016 AES study, although we note that this relationship only occurs with relation to the social inclusivity spectrum, while Miller's study relates trust in government to a traditional left-right spectrum. Miller, in using the terms 'left' and 'right', does not distinguish between social and economic issues, as this thesis does.

This thesis moves beyond traditional notions of 'left' and 'right' (per Chapter 4.5), favouring instead the three axes 'social inclusivity', 'attitudes to authority' and 'spending priorities.' We find that extremity in the error in estimating social inclusivity is related to having lower than expected trust in authority, but this is not the case with respect to spending priorities. In fact, as seen in Figure 8.2, there is no strong relationship between error in trust in authority, and error in spending priorities.

This research demonstrates that there is a relationship between extremity of views with relation to social inclusivity, and distrust of authority. This is a more precise finding than previous research, which has demonstrated a relationship between ideological extremity more generally, and distrust of authority [127]. That distrust in authority is more common alongside more extreme social perspectives, as opposed to more extreme perspectives on an economic spectrum, might go some way to explaining current populist movements internationally, which overwhelmingly focus on social issues. Examples of populist movements centred around social issues include: the rise of Donald Trump in the United States, who was elected on a populist anti-immigration platform and whose economic platform is more uncertain [26, 95, 93], and the rise of Lega (formerly Lega Nord) in Italy, who are anti-immigration [2, 61, 187] and oppose abortion rights [186], devoting little campaign time to their economic views [2, 49].

A more complex case to describe in this paradigm is the movement for the United Kingdom to leave the European Union ('Brexit'), a populist movement based around distrust in European Union institutions [93]. Brexit is more difficult to classify as a social or economic issue, since it encompasses key debates about both freedom of movement [74, 170], which would fall under the first dimension-reduced political axis of 'social inclusivity', and free trade [52, 178], which is an economic position. The more extreme anti-immigration views espoused by pro-Brexit politicians [75, 154] would fall neatly within the distrust paradigm explained here. However, a

tendency toward economic conservatism is shown in this thesis to have less relation to attitudes to authority, than attitudes on social inclusivity. Findings in this thesis, if replicated in a UK context, would thus suggest that Brexiteers might be more likely to be motivated by their views with respect to social inclusivity, than their economic concerns. This perspective is echoed in an Ipsos MORI poll released on 16 June 2016, exactly one week before the Brexit referendum, which found that 52% of Leave voters listed the number of immigrants coming to the UK as a major influence on their vote, while only 18% listed Brexit's economic impact [131].

The relationship between social extremity and distrust in authority, and the different relationship between spending priorities and distrust in authority, can be thus used to gain a greater understanding of possible political motivations behind populist social movements.

Asymmetry Further inference can be drawn by looking at the cluster most trusting of authority. Recalling that the ordinary least squares error's grand mean is **0**, it can be seen that the most trusting cluster, whose mean is located at $(-1.096, -0.643, -0.145)$, has mean less socially inclusive than expected, at -1.096. In other words, the cluster mostly likely to be more trusting of authority than their demographics would suggest is also slightly less socially inclusive than their demographics would suggest. Of course, this is not necessarily a strong relationship; especially in light of the fact that the other two components have a larger variance in the 'trust in authority' direction, there will be large numbers of people in the more socially extreme clusters who are also trusting of authority. Still, despite the first two dimension-reduced political variables being uncorrelated, after the removal of the influence of demographics, the group most likely to be more trusting in authority is also likely to be slightly less socially inclusive than expected. This cluster is not insubstantial, either; in fact, the most trusting cluster is associated with a proportion of 0.382.

This asymmetry in the error clusters, from the perspective of the social inclusivity and attitudes to authority axes, is also manifested in the two more extreme clusters. One of these clusters, with proportion 0.209, is centred at $(-1.908, 0.561, 0.468)$, while the other, with proportion 0.409, is centred at $(2.000, 0.315, -0.104)$. We can observe from this that the less populous, less socially inclusive cluster is even less trusting of authority that the more populous, more socially inclusive cluster; 0.561 is almost twice as high as 0.315. While low trust in authority is more likely at both extremes, it is not a symmetrical relationship.

The asymmetry of the more socially extreme error clusters, when it comes to trust in authority, has two key consequences in social science research. Firstly, it goes some way to explain current trends in populist discourse. Secondly, it dispels any notion that our findings support a popular 'horseshoe theory' of ideological symmetry in attitudes to authority. These two consequences will now be discussed in turn.

Recent populist movements around the globe have had a decidedly asymmetrical structure in terms of their social inclusivity. For example, as previously discussed, Brexit [75, 131], the election of Donald Trump in the United States [26, 95, 93],

and the rise of Lega/Five Star Movement in Italy [187, 68] have all had an anti-immigration base of support. This has also been seen with the anti-immigration groups National Rally (formerly National Front) and New Zealand First gaining varying amounts of public support in France and New Zealand respectively [76, 136, 161, 164]. These groups are both populist (distrusting of authority) and not socially inclusive. Supporters of these parties would likely belong to the error cluster with proportion 0.209 and centred at $(-1.908, 0.561, 0.468)$. In Australia, Pauline Hanson's One Nation (formerly One Nation), who are not socially inclusive, and have high distrust in authority, are able to garner a small but influential amount of political support [73]. Like the other examples mentioned here, Pauline Hanson's One Nation focus on immigration, rather than strictly economic issues [73]. Pauline Hanson's One Nation do not come close to receiving 20.9% of the vote in Australia [19], so do not capture the entirety of this component of the Gaussian mixture. One factor for this may be the variance in this cluster, leading some members to support populist parties, and some to not, as well as the fact that this cluster just represents the error in conditional expectations of political opinion, so some individuals' lack of social inclusivity and low trust in authority may be offset by their demographic factors.

Notably, the error cluster centred at $(-1.908, 0.561, 0.468)$ is less trusting of authority than the other cluster of more extreme social views, whose mean is at $(2.000, 0.315, -0.104)$. This would suggest that it is easier for a core base of people of low socially inclusivity who are distrustful of authority to emerge, than for a core base of socially inclusive people who are distrustful of authority. In Australia, this might explain why there has been an appetite for anti-immigration populism, while there has not been a similar rise of pro-immigration populism. Further, if this is the case globally, it goes some way to justifying the prominence of populist, not socially inclusive political parties, while also explaining why there might be fewer populist parties with a socially inclusive *raison d'etre*. In other words, there are globally more successful populist movements based around low social inclusivity, than populist movements based around high social inclusivity. While some 'left populist' parties and movements have grown in Europe, South America and the United States, these have been focused largely on economic platforms [145, 148, 162].

'Horseshoe theory'   'Horseshoe theory' refers to the idea that people on what has traditionally been called the 'left' and the 'right' are very similar to each other, due to their shared distrust and disenchantment with authority [60]. Proponents of horseshoe theory argue that the 'left' and the 'right' are both further from the 'centre' than they are from each other [167]. This research provides an additional way of testing whether this is the case, in that through applying a Gaussian mixture model to the residuals of the ordinary least squares model of the dimension-reduced political variables against the dimension-reduced demographic variables from the 2016 AES, three main clusters of political thought have emerged; one of which is what traditionally might have been called a 'left cynicism', one of which is what traditionally might have been called a 'right cynicism', and one of which is a more 'centrist' perspective [127]. It is true that the centres of these three clusters are

positioned in such a way that the more socially inclusive cluster and the less socially inclusive cluster are both less trusting of authority than the 'centre' is. In that sense, the 'left' and the 'right' have in common this shared distrust. However, the manifestations of these three different clusters are such that the differences between them are asymmetrical; there is not a complete symmetry as proponents of horseshoe theory have suggested [167].

For example, the cluster of political thought that is the least socially inclusive, relative to the position their demography would suggest, represents 20.9% of the sample, and this group has a mean distrust of authority of 0.561 higher than their demography would suggest. Meanwhile, the cluster of political thought that is the most socially inclusive, relative to the position their demography would suggest, has a mean distrust 0.315 higher than their demography would suggest, and represents 40.8% of the sample. Thus, there is a difference between these clusters both in terms of their size and their distrust of authority. What we see in the group that might traditionally be called 'cynics of the right', the group least socially inclusive, is that this group represents a smaller group of people, and have less trust in authority, than do their more socially inclusive counterparts. This difference, noted by critics of horseshoe theory, means that the two groups cannot be equivocated; that they are both distrusting of authority does not make them the same as each other [42].

The finding that in an Australian context, there is an asymmetry between the 'cynicism of the left' and 'cynicism of the right' described by Miller [127] is a repudiation of horseshoe theory. The differences in the appetites of different groups for populist approaches is both of intellectual interest, and can be added to the suite of information from which policymakers understand the public whom they represent.

8.3 Summary

This chapter has described the social science implications of the Gaussian mixture model chosen to describe the results of the 2016 Australian Election Study. The chosen model relates the three most important dimension-reduced political variables to the three most important dimension-reduced demographic variables. The nine key findings with relation to the linear conditional expectations of the political variables are that, on average:

1. People of higher socio-economic status and education are substantially more likely to be socially inclusive;

2. People of a later stage of life are substantially less likely to be socially inclusive;

3. People of a diverse cultural background are slightly more likely to be socially inclusive;

4. People of higher socio-economic status and education are slightly more likely to trust authority;

5. People of a later stage of life are substantially more likely to trust authority;

6. People of a diverse cultural background are more likely to trust authority;

7. People of higher socio-economic status and education are substantially less likely to favour increased taxation and spending;

8. People of a later stage of life are slightly less likely to favour increased taxation and spending; and

9. People of a diverse cultural background are slightly less likely to favour increased taxation and spending.

These expectations are associated with a substantial error term. Understanding this error term can also bring insight into how political opinions relate to one another. The error was modelled as a Gaussian mixture, with three clusters. Our findings are that, on average:

1. People who are either substantially more, or substantially less, socially inclusive than their demographics would suggest, are likely to have greater than expected distrust of authority;

2. The group with largest positive error with relation to distrust of authority is the group least socially inclusive, accounting for their demographics;

3. The group most trusting of authority, relative to their conditional expectation based upon their demography is, on average, slightly less socially inclusive than expected; and

Section 8.2 describes the consequences of these results in a political science context. We note that political science research has in the past identified three clusters of political thought, namely "cynics of the left," "cynics of the right," and "those in the center" [127]. Accounting for demographic differences, we find that the Gaussian mixture model of the error suggests that the "cynics of the right" are more likely to be 'right-wing' in terms of their social views, while this is less the case for "cynics of the left". This has profound implications in light of the recent rise of both left- and right-populism. With populism seen throughout international political discourse, understanding its manifestations has never been more important.

Chapter 9

Conclusion

9.1 Contextual relevance of results

Suppose the raw results of an opinion poll suggest that those over the age of 65 have a lower than average approval of marriage equality. This is a valuable (though not particularly surprising) finding. In this thesis, we develop a more holistic, analogous finding—that with statistical significance, an increase in an Australian's underlying "stage of life" is associated with an expected decrease in that person's underlying "social inclusivity". Both results are important, but this thesis adds value by focusing on underlying relationships, over their specific realisations, and by undertaking analysis within a sound statistical framework. By undertaking this research in a mathematically-rigorous, holistic manner, we are able to clarify and extend existing political science literature.

Chapter 4 introduced new, entirely data-driven, demographic and political spectra for Australia, as based on the 2016 Australian Election Study. The political spectrum is comprised of three dimensions, which we labelled 'social inclusivity', 'attitudes to authority' and 'spending priorities'. In the context of other political spectra throughout history, in Section 4.5 we noted that the first political spectrum consisted of just the 'left' and the 'right' [71], which was later split into a 'social' and an 'economic' axis [57]. With our social inclusivity and spending priorities axes being remarkably similar to the traditional social and economic axes, our new spectrum confirms the existence of these political dimensions, while adding a new axis of 'attitudes to authority'. We discuss that given the recent rise of populist thought around the world—populism being a political approach characterised by a distrust of "corrupt elites" ([132], cited in [25])—the recognition of this axis of ideology is more relevant than ever.

We then sought to identify major influences on these political axes. A Gaussian mixture model was selected to describe the relationships between demographic and political identity in Chapter 7. While a full treatment of results can be found in Section 8.1, the strongest influences are as follows:

- People of higher socio-economic status and education are substantially more likely to be socially inclusive;

- People of a later stage of life are substantially less likely to be socially inclusive;

- People of a later stage of life are substantially more likely to trust authority;

- People of a diverse cultural background are more likely to trust authority; and

- People of higher socio-economic status and education are substantially less likely to favour increased taxation and spending.

These holistic conclusions are explained in Chapter 8 with reference to previous political science research.

Finally, using the Gaussian mixture model, we identify a key association within political opinion. People who are either substantially more, or substantially less, socially inclusive than their demographics would suggest, are likely to have greater than expected distrust of authority. Further, this is not a symmetrical relationship; the group least trusting of authority, relative to their demography, is the least socially inclusive group, while the group most trusting of authority, relative to their demography, is on average slightly less socially inclusive than expected. This is reflected in lived experience, since many populist movements around the world come from a position of extremely low social inclusivity, while fewer come from a position of high social inclusivity (see Section 8.2).

9.2   Statistical relevance of methods

The variety of statistical techniques used in this thesis are useful both because they provide a reproduceable framework with which to conduct further social analyses, and because they verify and extend model selection methods for non-nested models on the basis of generalised goodness-of-fit criteria.

This thesis relies on a process for connecting two sets of variables with noisy, high-dimensional realisations. The process consists of first, reducing the high-dimensional realisations to the underlying variables they represent, and secondly, connecting these variables using a multivariate model. This technique was especially useful in the context of political analysis because of the established theory that political opinions are noisy realisations of underlying ideologies (see, e.g., [57]), meaning it was known in advance that dimension reduction would be possible to reduce the 103 observed political survey questions to just a few underlying variables. In general, where a correlation structure exists between a set of observed variables, we can tease out a lower-dimensional representation of these variables to allow us to produce easier-to-construct models with clearer outputs. What might be lost in predictive power will be more than gained in explanatory power; high-dimensional predictors can easily confound and confuse interpretation. In the world of political science, the same techniques used here could be reproduced on a series of data sets of diverse geographical and temporal origin, with their results compared (see Section 9.3).

This thesis also implements a novel series of techniques for model selection. We sought to compare a number of non-nested candidate models for exploring the

relationship between demographic and political identity, based on how well these models adhered to their assumptions. First, in Chapter 7, a method for selecting between two non-nested models on the basis of some generalised goodness-of-fit criterion was introduced; this was the parametric bootstrap cross-fitting method ('PBCM', or 'model mimicry' method) of Wagenmakers *et al.* [172]. We placed this work in the context of a more complete history of non-nested model selection criteria, and demonstrated the effectiveness of the method using simulated data. This demonstration was performed using the 'energy' goodness-of-fit statistic introduced by Skèkely and Rizzo [166]. We then explored an extension to the PBCM introduced in a 2014 conference paper by Schultheis and Naidu [156], which allows for multiple models to be compared simultaneously, and discussed parametric classifiers for the results of the multi-model PBCM. These classifiers may be useful in situations where multiple data sets need to be assessed, or where outputs are too high-dimensional to visualise. Future work will consider the application of these classifiers in a range of contexts, as well as comparing their accuracy to methods like the non-parametric approaches of Schultheis and Naidu [156] (see Section 9.3).

9.3   Future work

The discussion in Chapter 8 described the outputs of the Gaussian mixture model as applied to the 2016 Australian Election Study data, and the potential manifestations of this model in the physical world. This has been done by reference to political movements around the world, and across the last two decades. As a result, the discussion has had to assume that our results are generalisable. A key extension to this thesis is thus determining how generalisable our results are. This can be done in a number of ways:

1. *Verifying* the political and demographic identity spectra, by repeating this analysis using other similar Australian social attitude studies, such as the Australian Survey of Social Attitudes [17];

2. *Temporally extending* the analysis, by comparing how these results from the 2016 Australian Election Study compare with those from previous editions of the AES; and

3. *Geographically extending* the analysis, by comparing the results from the 2016 Australian Election Study, to that from similar studies internationally, including ArabBarometer [8], AmericasBarometer [108], and the World Values Survey [94].

Difficulties in these comparative analyses will arise in the construction of the dimension-reduced political and demographic variables. For example, if two surveys consisting of the same questions were conducted at different times, and the most galvanising issues changed from survey to survey, then unsupervised dimension-reduction techniques would produce different political spectra from each survey.

This makes comparing the results related to each dimension-reduced political variable difficult, since the political variables will be defined according to different criteria. This issue could be resolved by performing dimension reduction for all studies simultaneously, so the same political spectrum is produced for all time periods, though this would mask the potential occurrence of substantive changes in political spectra themselves over time.

Another, more troubling issue is that not all surveys ask the same questions. For example, as discussed in Chapter 2, while editions of the Australian Election Study often include the same questions repeated word-for-word, this is not true of all variables in the study. One solution in comparing Australian Election Studies over time is to only keep variables that persist in all editions, but this may leave out variables of interest as public discourse changes. For example, while questions specifically relating to asylum seekers were not asked in the AES before 2001 [22], removing questions relating to asylum seekers would ignore a substantial part of Australian political discourse. In comparing results internationally, this difficulty would be accentuated by the fact that the questions vary to an even greater extent from study to study; the choice of questions is dependent on, *inter alia*, the political discourse of the region, the language the questions are to be asked in, the purpose of the study, and the identity of the researcher(s) writing the questions. All of these factors change between, for example, the ArabBarometer [8], AmericasBarometer [108], and the World Values Survey [94].

Despite this, producing similar analyses at other periods of time, and in other regions, is an extension to this thesis that would greatly add value in a social science context. The ability to produce data-driven political spectra from other studies in a similar manner makes the work in this thesis more reproducible, and allows for historical and geographical comparisons of political spectra. No single spectrum of political ideology is fit for all times and for all places, and understanding the difference between political spectra in different scenarios could provide a baseline to compare research across contexts. By understanding how political spectra change over time and place, we can learn from the experiences of other cultures.

Once spectra of political and demographic identity have been produced across different regions and different time periods, the results of the Gaussian mixture model fit can also be compared to determine what trends emerge. Key results in this thesis have been found relating to manifestations of populism, and demographic influences on political identity. By comparing results from different regions and at different time periods, we can determine if these results are unique to Australia in 2016, and if not, when and where these results also emerge.

Future statistical work

We also would like to expand upon the parametric bootstrap cross-fitting method used in Chapter 7. Currently in preparation is a paper outlining the model comparison technique described in Chapter 7. The paper places the technique we use in this thesis in the context of surrounding literature, and uses both synthetic and observed datasets to verify and demonstrate the technique's validity. Using a wide

range of data sets, we will also seek to compare methods for classifying results of the parametric bootstrap cross-fitting method, using both the parametric classifiers used in Chapter 7 and non-parametric classifiers discussed by Schultheis and Naidu [156]. This paper will provide an outline for other statisticians to select models on the basis of novel goodness-of-fit measures, including fidelity to model assumptions.

9.4 Final remarks

Our thoughts are more than just a linear function of our demography, and more than just a linear function of our underlying political beliefs. It would be a fool's errand to try to predict an individual's political beliefs merely on the basis of their demography; that we cannot do so is a testament to the diversity of political expression in Australia.

Broadly speaking, this thesis has sought to understand the lines along which Australians are most politically united, and most divided, and understand in which segments of our society divisions are most likely to occur. Our experiences bear strong influence over our ideology. But when we statistically assess and explain these associations, we must grapple with the manifestations of our diversity; we must recognise, and describe, the error in our models. This thesis achieves this in two ways. Firstly, we model underlying political ideologies and demographic realities, rather than their noisier realisations from inidividual questions, using dimension reduction in Chapter 4. Secondly, we build a model with the goal of best describing the error in predicting political belief, using the assumption-driven model selection methods of Chapter 7. We find that the error in our model is itself informative—there is an association between extremity in social views, and higher distrust of authority, relative to that suggested by ones demographics. By understanding the ways in which our expectations might be wrong, we have been able to highlight prominent clusters of political ideology, and better understand the political movements that arise from them.

An error doesn't become a mistake until you refuse to understand it.

Appendix A

The 2016 Australian Election Study Questionnaire

The following pages contain the paper version of the questionnaire from the 2016 Australian Election Study [121]. The questionnaire included is that mailed to the sample drawn from the Australian electoral roll, provided by the Australian Electoral Commission. Text in pink on the following pages was not printed on copies distributed to potential participants.

This is not the only format of the 2016 Australian Election Study. There was also:

- an online version; and

- a different paper version for subjects drawn from a random sample of Australian addresses (see Chapter 2).

All versions of the questionnaire contained the same questions, in the same order. Other questionnaires can be found at http://dx.doi.org/10.4225/87/7OZCZA. For more information on the 2016 Australian Election Study, see Chapter 2.

Australian
National
University

CRICOS No. 00213J   ABN 83 791 724 622

www.srcentre.com.au/aes

Username: <username>
Password: <password>

QR code

«Title» «paddrss»
<Add1>
<Add2>
 <SUBURB> <STATE> <PCODE>

<DATE>

Dear «Laddrss»,

AUSTRALIAN ELECTION STUDY 2016

We need to hear from Australian citizens to find out what people thought about the recent Federal election. Your views are important to us even if you were not eligible to vote, or you had little or no interest in the election.

This is a significant national study being carried out by the Australian National University and the Queensland University of Technology. It is the largest survey of its kind and has taken place after every federal election since 1987.

Return the questionnaire by 25 August 2016 for a chance to win 1 of 8 prizes of $500*.

For your convenience, we have provided more than one way for the questionnaire to be completed. You may:

1. Complete the attached form and post it back to us in the reply-paid envelope provided.

2. **OR** use the QR code provided on the top right hand corner of this page to open the questionnaire on your mobile or tablet device.

3. **OR** go to www.srcentre.com.au/aes and enter the username and password provided above.

The questionnaire will take around 30 minutes to complete, depending on your answers. All responses are strictly confidential and private and you will not be identified in any way.

For more information, please refer to the enclosed Participant Information sheet. If you have any questions about the survey, or if you would like to opt out, you can call the toll free hotline on 1800 023 040 or email aesurvey@srcentre.com.au.

Thank you in advance for your help! In participating you will be part of a unique research project.

Yours sincerely,

Professor Ian McAllister
Australian National University

Professor Clive Bean
Queensland University
of Technology

Dr Juliet Pietsch
Australian National University

*For full prize draw terms and conditions go to www.srcentre.com.au/aes.

AES 2016 - Page 1

Frequently Asked Questions

What if I have concerns about the ethical conduct of the survey?

If you have any concerns or complaints about the ethical conduct of the survey please contact the Ethics Manager, The ANU Human Research Ethics Committee on 02 6125 3427 or by email at human.ethics.officer@anu.edu.au. If you have concerns about your privacy in regards to this project you may contact the Office of Australian Information Commissioner at enquiries@oaic.gov.au or on 1300 363 992.

What are my options for completing the survey?

If you choose to participate, you can complete online using the QR code provided or by going to **www.srcentre.com.au/aes** and following the links. To access the survey you will need the username and password in the panel located in the top right hand corner of the front page. Alternatively, you can complete and return this questionnaire in the reply-paid envelope provided by the due date.

Participation in this survey is entirely voluntary and you can withdraw at any point. The information collected will be treated in strict confidence.

How To Fill Out This Questionnaire

To answer most of the questions you only need to **cross a box**. Please cross the box which is **closest** to your view - there are no right or wrong answers. Here is an example.

Do you think the government should spend more or less on education?

Spend more on education [X] 1
Doesn't matter [ ] 2
Spend less on education [ ] 3

If you think the government should spend more on education, you would put a cross in the box as shown.

Some boxes have 'Go to' instructions that look like this: [ ] ➡ **GO TO B3**

If you make a mistake, simply colour in the wrong box and mark the correct box like this [X] [▨]

Sometimes you are asked to write in an answer—in that case, simply write your answer in the space provided.

Please read each question carefully. Remember, there are no right or wrong answers—we just want to know **your own** personal opinions.

We hope you enjoy the questionnaire. Thank you very much for taking part in this study.

Professor Ian McAllister, Australian National University
Professor Clive Bean, Queensland University of Technology
Dr Juliet Pietsch, Australian National University

Section A: The Election Campaign

First, some questions about the Federal election campaign, the period between when the election was announced on Sunday 8 May and election day on Saturday 2 July.

A1 Generally speaking, how much interest do you *usually have* in what's going on in politics?

- ☐₁ A good deal
- ☐₂ Some
- ☐₃ Not much
- ☐₄ None

A2 How much attention did you pay to reports about the election campaign?

	A good deal	Some	Not much	None at all
In the newspapers	☐₁	☐₂	☐₃	☐₄
On television	☐₁	☐₂	☐₃	☐₄
On the radio	☐₁	☐₂	☐₃	☐₄
On the internet	☐₁	☐₂	☐₃	☐₄

A3 And how much interest would you say you took *in the election campaign* overall?

- ☐₁ A good deal
- ☐₂ Some
- ☐₃ Not much
- ☐₄ None at all

A4 Here is a list of things some people do during elections. How often did you do any of these things during the recent election?

	Frequently	Occasionally	Rarely	Not at all
Discuss politics with others in person (i.e. face to face or over the phone)	☐₁	☐₂	☐₃	☐₄
Discuss politics with others online (i.e. through email or on a social network site like Facebook or Twitter)	☐₁	☐₂	☐₃	☐₄
Talk to other people to persuade them to vote for a particular party or candidate	☐₁	☐₂	☐₃	☐₄
Show your support for a particular party or candidate by, for example, attending a meeting, putting up a poster, or in some other way	☐₁	☐₂	☐₃	☐₄
Go to any political meetings or rallies	☐₁	☐₂	☐₃	☐₄
Contribute money to a political party or election candidate	☐₁	☐₂	☐₃	☐₄

A5 Did you watch the televised debate between Bill Shorten and Malcolm Turnbull on Sunday 29 May?

- ☐₁ Yes
- ☐₂ No

A6 From what you saw or what you heard or read about it, who do you think performed better in the debate—Bill Shorten or Malcolm Turnbull?

- ☐₁ Shorten did much better
- ☐₂ Shorten did somewhat better
- ☐₃ About equal
- ☐₄ Turnbull did somewhat better
- ☐₅ Turnbull did much better

A7 During the election campaign, did a candidate or anyone from a political party contact you to persuade you to vote for them?
Please cross all that apply

- [ ] No
- [ ] Yes, by telephone
- [ ] Yes, by mail
- [ ] Yes, face-to-face
- [ ] Yes, by text message or SMS
- [ ] Yes, by email
- [ ] Yes, by social network site or other web-based method

A8 Which parties or candidates contacted you by any of these means?
Please cross all that apply

- [ ] None
- [ ] Liberal
- [ ] Labor
- [ ] National
- [ ] Greens
- [ ] Other party or candidate

A9 During the campaign, did a friend, family member, neighbour, work colleague or other acquaintance try to persuade you to vote for a particular party or candidate?
Please cross all that apply

- [ ] No
- [ ] Yes, by telephone
- [ ] Yes, by mail
- [ ] Yes, face-to-face
- [ ] Yes, by text message or SMS
- [ ] Yes, by email
- [ ] Yes, by social network site or other web-based method

A10 Do you think that voting at Federal elections should be compulsory, or do you think that people should only have to vote if they want to?

- [ ] 1 Strongly favour compulsory voting
- [ ] 2 Favour compulsory voting
- [ ] 3 Favour people voting only if they want to
- [ ] 4 Strongly favour people voting only if they want to

A11 Would you have voted in the election if voting had not been compulsory?

- [ ] 1 Definitely would have voted
- [ ] 2 Probably would have voted
- [ ] 3 Might/might not have voted
- [ ] 4 Probably not have voted
- [ ] 5 Definitely not have voted

AES 2016 - Page 4

A12 Do you think that the voting age in elections should be lowered to 16, or should it stay at 18?

- ☐ 1 Definitely lowered to 16
- ☐ 2 Probably lowered to 16
- ☐ 3 Probably stay at 18
- ☐ 4 Definitely stay at 18

A13 During the 2016 election campaign did you read or access any of the following?
Please cross all that apply

- ☐ Official party or candidate campaign sites
  (e.g. home pages, blogs, official Facebook profiles, official YouTube channels)
- ☐ Mainstream news media sites (e.g. ABC Online, SMH site etc.)
- ☐ Unofficial online content (i.e. non-party produced campaign material (e.g.YouTube))
- ☐ Federal Parliament site
- ☐ Australian Electoral Commission site
- ☐ None of the above

A14 During the 2016 election campaign did you do any of the following activities online?
Please cross all that apply

- ☐ Signed up to receive information from a party or candidate and/or registered as their follower/friend/
  supporter on Twitter, Facebook, or Instagram
- ☐ Shared official campaign/political content on a blog, Twitter feed or social network profile
- ☐ Shared unofficial campaign/political content (e.g. links to videos, news stories, jokes) on a blog,
  Twitter feed or social network profile
- ☐ Joined or started a political or election related group on a social networking site (e.g. Facebook etc.)
- ☐ None of the above

Section B: Party Preference and Voting

B1 Generally speaking, do you usually think of yourself as Liberal, Labor, National or what?

- ☐ 1 Liberal
- ☐ 2 Labor
- ☐ 3 National (Country) Party
- ☐ 4 Greens
- ☐ 5 Other party *(please specify)*

  ┌───┐
  │   │
  └───┘

- ☐ 6 No party ➜ **GO TO B3**

B2 Would you call yourself a very strong, fairly strong, or not very strong supporter of that party?

- ☐ 1 Very strong supporter
- ☐ 2 Fairly strong supporter
- ☐ 3 Not very strong supporter

B3 Would you say you cared a good deal which party won the Federal election or that you did not
care very much which party won?

- ☐ 1 Cared a good deal
- ☐ 2 Did not care very much
- ☐ 3 Did not care at all

AES 2016 - Page 5

B4 When did you decide how you would *definitely* vote in this election?

- ☐₁ A long time ago
- ☐₂ A few months ago, *before* election day was announced
- ☐₃ About the time the election was announced, on 8 May
- ☐₄ In the first few weeks of the campaign
- ☐₅ A few days before election day
- ☐₆ On election day
- ☐₇ Did not vote in this election → **GO TO B6**

B5 In deciding how you would vote in the election, which was *most* important to you?
Please cross one box only

- ☐₁ The party leaders
- ☐₂ The policy issues
- ☐₃ The candidates in your electorate
- ☐₄ The parties taken as a whole

B6 Which do you think is better—when the Federal Government has a majority in *both* the House of Representatives *and* the Senate, OR when the Federal Government in the House of Representatives *does not* control the Senate?

- ☐₁ Much better when Government controls *both*
- ☐₂ Better when Government controls *both*
- ☐₃ Neither/doesn't matter
- ☐₄ Better when Government *does not* control the Senate
- ☐₅ Much better when Government *does not* control the Senate

B7 Considering everything the Labor Party and the Liberal Party stand for, would you say there is....

- ☐₁ A good deal of difference between the parties
- ☐₂ Some difference between the parties
- ☐₃ Not much difference between the parties
- ☐₄ No difference between the parties

B8 In politics, people sometimes talk about the 'left' and the 'right'. Where would you place yourself on a scale from 0 to 10, where 0 means the left and 10 means the right?

Left										Right
☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10

Using the same scale, where would you place each of the Federal political parties?

	Left										Right
Liberal Party	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
Labor Party (ALP)	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
National Party	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
Greens	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10

B9 In the Federal election for the House of Representatives on Saturday 2 July, which party did you vote for *first* in the House of Representatives? And in the Senate election?
If you did not vote in the election on 2 July, go to B14

	House of Representatives	Senate
Liberal Party	□ 1	□ 1
Labor Party (ALP)	□ 2	□ 2
National (Country) Party	□ 3	□ 3
Greens	□ 4	□ 4
Other *(please specify party below)*	□ 5	□ 5
Voted informal	□ 6	□ 6

B10 Did you vote in person on election day or did you cast your vote before election day?
If you did not vote in the election on 2 July, go to B14

- □ 1 Voted in person on election day
- □ 2 Voted in person before election day (at a pre-polling voting centre)
- □ 3 Cast postal vote before election day

B11 *If your first preference was for the Greens or other minor party:* In the end, which of the two major parties, the Liberal-National Coalition or the Labor Party, did you give your preference to in the House of Representatives? And in the Senate election?

	House of Representatives	Senate
Liberal-National Coalition	□ 1	□ 1
Labor Party (ALP)	□ 2	□ 2
Not sure/Don't know	□ 3	□ 3

B12 In voting for the House of Representatives, did you follow a party 'How to Vote' card or did you decide your own preferences?

- □ 1 Followed a 'How to Vote' card
- □ 2 Decided my own preferences

B13 Was there any time during the election campaign when you *seriously* thought you might give your *first* preference to another party in the House of Representatives?
Please cross one box only

- □ 1 No
- □ 2 Yes, Liberal Party
- □ 3 Yes, Labor Party (ALP)
- □ 4 Yes, National (Country) Party
- □ 5 Yes, Greens
- □ 6 Yes, another party/independent

B14 In the last Federal election in September 2013, when Labor was led by Kevin Rudd and the Liberals by Tony Abbott, which party got your first preference then in the House of Representatives election?

- □ 1 Liberal Party
- □ 2 Labor Party (ALP)
- □ 3 National (Country) Party
- □ 4 Greens
- □ 5 Another party/independent
- □ 6 Did not vote
- □ 7 Don't know/can't recall

Need new scan marks

AES 2016 - Page 7

B15 Before this current Federal election for the House of Representatives, had you always voted for the same party, or had you sometimes voted for different parties?
Please cross all that apply

☐ Always voted for the same party

☐ Sometimes voted Liberal

☐ Sometimes voted Labor (ALP)

☐ Sometimes voted National (Country)

☐ Sometimes voted Greens

☐ Sometimes voted for other party/independent

☐ Did not vote before this election

B16 Some people say that political parties in Australia care what ordinary people think. Others say that political parties in Australia don't care what ordinary people think. Where would you place your view on this scale from 1 to 5?

Political parties in Australia care what ordinary people think				Political parties in Australia don't care what ordinary people think
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

B17 Where would you place your view on this scale from 1 to 5, where 1 means that political parties are necessary to make our political system work, and 5 means that political parties are not needed in Australia?

Political parties are necessary to make our political system work				Political parties are not needed in Australia
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

B18 Finally in this section, we would like to know what you think about each of our political parties. Please rate each party on a scale from 0 to 10, where 0 means you strongly *dislike* that party and 10 means that you strongly *like* that party. If you are neutral about a particular party or don't know much about them, you should give them a rating of 5.

How do you feel about the:	Strongly dislike					Neutral					Strongly like
Liberal Party	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
Labor Party (ALP)	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
National Party	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
Greens	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10

Section C: Politicians and Government

C1 Again using a scale from 0 to 10, please show how much you *like* or *dislike* the party leaders. If you don't know much about them, you should give them a rating of 5.

How do you feel about:	Strongly dislike					Neutral					Strongly like
Bill Shorten	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
Malcolm Turnbull	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
Barnaby Joyce	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
Tony Abbott	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10
Richard Di Natale	☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10

C2 Here is a list of words and phrases people use to describe party leaders. Thinking first about **Bill Shorten**, in your opinion how well does each of these describe him—extremely well, quite well, not too well or not well at all?

	Extremely well	Quite well	Not too well	Not well at all
Intelligent	□ 1	□ 2	□ 3	□ 4
Compassionate	□ 1	□ 2	□ 3	□ 4
Competent	□ 1	□ 2	□ 3	□ 4
Sensible	□ 1	□ 2	□ 3	□ 4
Provides strong leadership	□ 1	□ 2	□ 3	□ 4
Honest	□ 1	□ 2	□ 3	□ 4
Knowledgeable	□ 1	□ 2	□ 3	□ 4
Inspiring	□ 1	□ 2	□ 3	□ 4
Trustworthy	□ 1	□ 2	□ 3	□ 4

C3 Now thinking about **Malcolm Turnbull**, in your opinion how well does each of these describe him—extremely well, quite well, not too well or not well at all?

	Extremely well	Quite well	Not too well	Not well at all
Intelligent	□ 1	□ 2	□ 3	□ 4
Compassionate	□ 1	□ 2	□ 3	□ 4
Competent	□ 1	□ 2	□ 3	□ 4
Sensible	□ 1	□ 2	□ 3	□ 4
Provides strong leadership	□ 1	□ 2	□ 3	□ 4
Honest	□ 1	□ 2	□ 3	□ 4
Knowledgeable	□ 1	□ 2	□ 3	□ 4
Inspiring	□ 1	□ 2	□ 3	□ 4
Trustworthy	□ 1	□ 2	□ 3	□ 4

C4 Do you approve or disapprove of the way the Liberal Party handled the leadership change in September of last year, when Malcolm Turnbull replaced Tony Abbott?

□ 1 Strongly approve

□ 2 Approve

□ 3 Disapprove

□ 4 Strongly disapprove

C5 On the whole, are you very satisfied, fairly satisfied, not very satisfied or not at all satisfied with the way democracy works in Australia?

□ 1 Very satisfied

□ 2 Fairly satisfied

□ 3 Not very satisfied

□ 4 Not at all satisfied

C6 In general, do you feel that the people in government are too often interested in looking after themselves, or do you feel that they can be trusted to do the right thing nearly all the time?

□ 1 Usually look after themselves

□ 2 Sometimes look after themselves

□ 3 Sometimes can be trusted to do the right thing

□ 4 Usually can be trusted to do the right thing

C7 Would you say the government is run by a few big interests looking out for themselves, or that it is run for the benefit of all the people?

- ☐ 1 Entirely run for the big interests
- ☐ 2 Mostly run for the big interests
- ☐ 3 About half and half
- ☐ 4 Mostly run for the benefit of all
- ☐ 5 Entirely run for the benefit of all

C8 Over the past five years or so, have you done any of the following things to express your views about something the government should or should not be doing?

	Yes	No
Contacted a politician or government official by email	☐ 1	☐ 2
Contacted a politician or government official either in person, or in writing	☐ 1	☐ 2
Taken part in a protest, march or demonstration	☐ 1	☐ 2
Worked together with people who shared the same concern	☐ 1	☐ 2
Signed a written petition	☐ 1	☐ 2
Signed an online or e-petition	☐ 1	☐ 2

C9 Some people say that Federal politicians know what ordinary people think. Others say that Federal politicians don't know much about what ordinary people think. Where would you place your view on this scale from 1 to 5?

Federal politicians know what ordinary people think				Federal politicians don't know what ordinary people think
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

C10 Some people say it makes a big difference who is in power. Others say it doesn't make any difference who is in power. Using the scale below, where would you place yourself?

It makes a big difference who is in power				It doesn't make any difference who is in power
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

C11 Some people say that no matter who people vote for, it won't make any difference to what happens. Others say that who people vote for can make a big difference to what happens. Using the scale below, where would you place yourself?

Who people vote for can make a big difference				Who people vote for won't make any difference
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

C12 Should there be more efforts to increase the number of women MPs? If so, what means would you prefer?
Please cross one box below only

- ☐ 1 No, there is no need to increase the number of women MPs
- ☐ 2 No, nothing needs to be done, it will happen naturally
- ☐ 3 Yes, by legally requiring all political parties to select more women candidates by means of a 'quota'
- ☐ 4 Yes, the political parties should make their own voluntary commitments to increase the number of women MPs
- ☐ 5 Yes, by encouraging more women to participate in politics

Section D: Election Issues

D1 Here is a list of important issues that were discussed during the election campaign. **When you were deciding about how to vote**, how important was each of these issues to you personally?

Issues	Extremely important	Quite important	Not very important
1. Taxation	☐₁	☐₂	☐₃
2. Immigration	☐₁	☐₂	☐₃
3. Education	☐₁	☐₂	☐₃
4. The environment	☐₁	☐₂	☐₃
5. Government debt	☐₁	☐₂	☐₃
6. Health and Medicare	☐₁	☐₂	☐₃
7. Refugees and asylum seekers	☐₁	☐₂	☐₃
8. Global warming	☐₁	☐₂	☐₃
9. Superannuation	☐₁	☐₂	☐₃
10. Management of the economy	☐₁	☐₂	☐₃

D2 Still thinking about these same issues, whose policies—the Labor Party's or the Liberal–National Coalition's—would you say **come closer** to your own views on each of these issues?

Issues	Labor (ALP)	Liberal–National Coalition	No difference	Don't know
1. Taxation	☐₁	☐₂	☐₃	☐₄
2. Immigration	☐₁	☐₂	☐₃	☐₄
3. Education	☐₁	☐₂	☐₃	☐₄
4. The environment	☐₁	☐₂	☐₃	☐₄
5. Government debt	☐₁	☐₂	☐₃	☐₄
6. Health and Medicare	☐₁	☐₂	☐₃	☐₄
7. Refugees and asylum seekers	☐₁	☐₂	☐₃	☐₄
8. Global warming	☐₁	☐₂	☐₃	☐₄
9. Superannuation	☐₁	☐₂	☐₃	☐₄
10. Management of the economy	☐₁	☐₂	☐₃	☐₄

D3 Still thinking about the same 10 issues, which of these issues was **most** important to you and your family during the election campaign? And which next?
Please put the number of the issue in the appropriate box below

Issue of most concern [ ]   **Second issue of concern** [ ]

D4 How does the **financial situation of your household** now compare with what it was 12 months ago? And how do you think the **general economic situation in Australia** now compares with what it was 12 months ago?

	A lot better	A little better	About the same	A little worse	A lot worse
Financial situation of your household	☐₁	☐₂	☐₃	☐₄	☐₅
Economic situation in Australia	☐₁	☐₂	☐₃	☐₄	☐₅

D5 Compared with 12 months ago, would you say that the Federal government's policies have had a good effect, a bad effect, or that they really have not made much difference to the financial situation of your household? And what effect do you think they have had on the general economic situation in Australia as a whole?

	A good effect	Not much difference	A bad effect
Financial situation of your household	☐ 1	☐ 2	☐ 3
Economic situation in Australia	☐ 1	☐ 2	☐ 3

D6 Compared to now, what do you think the financial situation of your household will be *in 12 months time*? And what do you think the general economic situation in Australia as a whole will be in 12 months time?

	A lot better	A little better	About the same	A little worse	A lot worse
Financial situation of your household	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Economic situation in Australia	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

D7 Do you think that, *12 months from now*, the Federal government's policies will have had a good effect, a bad effect, or that they really will have not made much difference to the financial situation of your household? And what effect do you think they will have had on general economic situation in Australia as a whole?

	A good effect	Not much difference	A bad effect
Financial situation of your household	☐ 1	☐ 2	☐ 3
Economic situation in Australia	☐ 1	☐ 2	☐ 3

D8 Please say whether there should be more or less public expenditure in each of the following areas. Remember if you say 'more' it could require a tax increase, and if you say 'less' it could require a reduction in those services.

	Much more than now	Somewhat more than now	The same as now	Somewhat less than now	Much less than now
Health	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Education	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Unemployment benefits	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Defence	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Old-age pensions	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Business and industry	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Police and law enforcement	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
The National Disability Insurance Scheme	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Public transport infrastructure	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Child care	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

D9 Over the next ten years or so, how likely or unlikely is it that you will improve your standard of living?

☐ 1 Very likely
☐ 2 Somewhat likely
☐ 3 Somewhat unlikely
☐ 4 Very unlikely

D10 On the whole, how satisfied are you with the present state of the economy in Australia?

Extremely dissatisfied										Extremely satisfied
☐ 0	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6	☐ 7	☐ 8	☐ 9	☐ 10

AES 2016 - Page 12

D11 How likely or unlikely do you think it is that your household's income could be severely reduced *in the next 12 months*?

- ☐ 1 Very likely
- ☐ 2 Somewhat likely
- ☐ 3 Somewhat unlikely
- ☐ 4 Very unlikely

D12 If you lost your job, how easy or difficult would it be to find another job **in the next 12 months**?

If your spouse/partner lost their job, how easy or difficult would it be for them to find another job *in the next 12 months*?

	Very easy	Somewhat easy	Somewhat difficult	Very difficult	Do not have spouse/partner
You	☐ 1	☐ 2	☐ 3	☐ 4	
Your spouse/partner	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

D13 Please say whether you strongly agree, agree, disagree or strongly disagree with each of these statements.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
High income tax makes people less willing to work hard	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
The trade unions in this country have too much power	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Big business in this country has too much power	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Income and wealth should be redistributed towards ordinary working people	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
There should be stricter laws to regulate the activities of trade unions	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
The government should take measures to reduce differences in income levels	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Section E: Social Policy

E1 If the government had a choice between reducing taxes or spending more on social services, which do you think it should do?

- ☐ 1 Strongly favour reducing taxes
- ☐ 2 Mildly favour reducing taxes
- ☐ 3 Depends
- ☐ 4 Mildly favour spending more on social services
- ☐ 5 Strongly favour spending more on social services

E2 The statements below indicate some of the changes that have been happening in Australia over the years. For each one, please say whether you think the change has gone too far, not gone far enough, or is it about right?

	Gone much too far	Gone too far	About right	Not gone far enough	Not gone nearly far enough
Equal opportunities for migrants	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
The right to show nudity and sex in films and magazines	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Aboriginal land rights	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Equal opportunities for women	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
The number of migrants allowed into Australia at the present time	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Government help for Aborigines	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Building closer relations with Asia	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

E3 Which one of these statements comes closest to how you feel about abortion in Australia?

☐ 1 Women should be able to obtain an abortion readily when they want one

☐ 2 Abortion should be allowed only in special circumstances

☐ 3 Abortion should not be allowed under any circumstances

☐ 4 Don't know

E4 Do you personally favour or oppose same sex couples being given the same rights to marry as couples consisting of a man and a woman?

☐ 1 Strongly favour

☐ 2 Favour

☐ 3 Oppose

☐ 4 Strongly oppose

E5 What do you think is the best way to handle the processing and resettlement of asylum seekers who come by boat and manage to reach Australian waters?

☐ 1 Process and resettle offshore

☐ 2 Process offshore but resettle in Australia

☐ 3 Process and resettle onshore in Australia

☐ 4 None of these options

E6 Here are some statements about general social concerns. Please say whether you strongly agree, agree, disagree or strongly disagree with each of these statements.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
The death penalty should be reintroduced for murder	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
The smoking of marijuana should NOT be a criminal offence	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
People who break the law should be given stiffer sentences	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Women should be given preferential treatment when applying for jobs and promotions	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
All boats carrying asylum seekers should be turned back	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Terminally ill patients should be able to end their own lives with medical assistance	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
The government should increase opportunities for women in business and industry	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

E7 If a referendum were held to recognise Indigenous Australians in the Constitution would you support or oppose such a change to the Constitution?

☐ 1 Strongly support

☐ 2 Support

☐ 3 Oppose

☐ 4 Strongly oppose

E8 Next, a question about what you think the aims of Australia should be for the next ten years. Here is a list of four aims that different people would give priority.   If you had to choose among these four aims, which would be your first choice?   And which would be your second choice?
Put the number of the statement in the appropriate box

1. Maintain order in the nation

2. Give people more say in important government decisions

3. Fight rising prices

4. Protect freedom of speech

First choice ☐   **Second Choice** ☐

E9 How much confidence do you have in the following organisations?

	A great deal of confidence	Quite a lot of confidence	Not very much confidence	None at all
The armed forces	☐ 1	☐ 2	☐ 3	☐ 4
The legal system	☐ 1	☐ 2	☐ 3	☐ 4
The press	☐ 1	☐ 2	☐ 3	☐ 4
Television	☐ 1	☐ 2	☐ 3	☐ 4
Trade unions	☐ 1	☐ 2	☐ 3	☐ 4
The police	☐ 1	☐ 2	☐ 3	☐ 4
The Federal government in Canberra	☐ 1	☐ 2	☐ 3	☐ 4
Australian political parties	☐ 1	☐ 2	☐ 3	☐ 4
The Federal parliament	☐ 1	☐ 2	☐ 3	☐ 4
The public service	☐ 1	☐ 2	☐ 3	☐ 4
Major Australian companies	☐ 1	☐ 2	☐ 3	☐ 4
Banks and financial institutions	☐ 1	☐ 2	☐ 3	☐ 4
Universities	☐ 1	☐ 2	☐ 3	☐ 4
The Australian political system	☐ 1	☐ 2	☐ 3	☐ 4

E10 Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?

☐ 1 Most people can be trusted

☐ 2 Can't be too careful

Section F: General Political Views

F1 How important do you feel the Queen and the Royal Family are to Australia?

☐ 1 Very important

☐ 2 Fairly important

☐ 3 Not very important

F2 Do you think that Australia should become a republic with an Australian head of state, or should the Queen be retained as head of state?

☐ 1 Strongly favour becoming republic

☐ 2 Favour becoming republic

☐ 3 Favour retaining the Queen as head of state

☐ 4 Strongly favour retaining the Queen as head of state

F3 How important do you think the Australian alliance with the United States under the ANZUS treaty is for protecting Australia's security?

☐ 1 Very important

☐ 2 Fairly important

☐ 3 Not very important

☐ 4 Not at all important

F4 If Australia's security were threatened by some other country, how much trust do you feel Australia can have in the United States to come to Australia's defence?

- ☐ 1 A great deal
- ☐ 2 A fair amount
- ☐ 3 Not very much
- ☐ 4 None at all

F5 In your opinion, are any of the following countries likely to pose a threat to Australia's security?

Countries	Very likely	Fairly likely	Not very likely
1. Japan	☐ 1	☐ 2	☐ 3
2. United States	☐ 1	☐ 2	☐ 3
3. China	☐ 1	☐ 2	☐ 3
4. Vietnam	☐ 1	☐ 2	☐ 3
5. Malaysia	☐ 1	☐ 2	☐ 3
6. Indonesia	☐ 1	☐ 2	☐ 3

F6 Please say whether you strongly agree, agree, disagree or strongly disagree with each of the following statements.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
Australia's defence is stronger now than it was 10 years ago	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Australia would be able to defend itself successfully if it were ever attacked	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Australia's trading future lies in Asia	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Australia should provide military assistance for the war on terrorism	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

F7 Do you think the number of immigrants allowed into Australia nowadays should be reduced or increased?

- ☐ 1 Increased a lot
- ☐ 2 Increased a little
- ☐ 3 Remain about the same as it is
- ☐ 4 Reduced a little
- ☐ 5 Reduced a lot

F8 There are different opinions about the effects that immigrants have on Australia. How much do you agree or disagree with each of the following statements?

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
Immigrants increase the crime rate	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Immigrants are generally good for Australia's economy	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Immigrants take jobs away from people who are born in Australia	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Immigrants make Australia more open to new ideas and cultures	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

F9 How serious a threat do you think global warming will pose to you or your way of life in your lifetime?

- ☐ 1 Very serious
- ☐ 2 Fairly serious
- ☐ 3 Not very serious
- ☐ 4 Not at all serious

AES 2016 - Page 16

F10 And a final question for this section, a quick quiz on Australian government. For each of the following statements, please say whether it is true or false. If you don't know the answer, cross the "don't know" box and try the next one.

	True	False	Don't know
1. Australia became a federation in 1901	☐ 1	☐ 2	☐ 3
2. There are 75 members of the House of Representatives	☐ 1	☐ 2	☐ 3
3. The Constitution can only be changed by the High Court	☐ 1	☐ 2	☐ 3
4. The Senate election is based on proportional representation	☐ 1	☐ 2	☐ 3
5. No-one may stand for Federal parliament unless they pay a deposit	☐ 1	☐ 2	☐ 3
6. The longest time allowed between Federal elections for the House of Representatives is four years	☐ 1	☐ 2	☐ 3

Section G: Education and Work

G1 How old were you when you left secondary school?

Age when left secondary school ☐☐

☐ 1	No formal schooling	➜	**GO TO G3**
☐ 2	Went to primary school only	➜	**GO TO G3**
☐ 3	Still at secondary school	➜	**GO TO G3**

G2 In all, how many years of **tertiary** study have you completed since you left secondary school?
If your tertiary study was part-time, give the number of years of equivalent full-time study.

Years of full-time tertiary study ☐☐

G3 Have you obtained a trade qualification, a degree or a diploma, or any other qualification since leaving school? What is your highest qualification?

☐ 1 **No qualification** since leaving school

Qualified—*If multiple apply, please indicate your highest qualification only*

☐ 2 Postgraduate Degree or Postgraduate Diploma
☐ 3 Bachelor Degree (including Honours)
☐ 4 Undergraduate Diploma
☐ 5 Associate Diploma
☐ 6 Trade qualification
☐ 7 Non-trade qualification

G4 Now some questions about the work you are doing now. Last week, what were you mainly doing?

☐ 1 Working full-time for pay
☐ 2 Working part-time for pay
☐ 3 Unemployed—looking for full-time work
☐ 4 Unemployed—looking for part-time work
☐ 5 Retired from paid work
☐ 6 A full-time school or university student
☐ 7 Keeping house
☐ 8 Other *(please specify)* [_____]

G5 What kind of work do you do? Please give your full job title, and as much detail as you can. For public servants, state official designation and occupation. For armed services personnel, state rank and occupation.
If you are unemployed or retired, please describe your last regular paid job
If you have never worked for pay, please skip this section and go to Section H

a. Full job title

b. What are (or were) the ***main tasks*** that you usually perform? Please give full details.

Main tasks

c. What kind of business or industry is (or was) that in? What do they do or make at the place where you work (or used to work)?

Kind of industry

d. Which of the following best describes the position that you hold (or held most recently)?
- [ ] 1 Upper managerial
- [ ] 2 Middle managerial
- [ ] 3 Lower managerial
- [ ] 4 Supervisory
- [ ] 5 Non-supervisory

e. Whom do (or did) you work for?
- [ ] 1 Self-employed
- [ ] 2 Employee in private company or business
- [ ] 3 Employee of Federal/State/Local Government
- [ ] 4 Employee in family business or farm

f. As an official part of your job, do (or did) you directly supervise the work of other employees or tell other employees what work to do?
- [ ] 1 Yes
- [ ] 2 No

G6 Do you belong to a trade union?
- [ ] 1 Yes
- [ ] 2 No

Section H: Personal Background

This final section includes questions about yourself and your family background. These characteristics are very important to our research on how people in different circumstances feel about the issues covered earlier.
The information you provide is totally confidential and will be used only for this research.

H1 Are you male or female?
- [ ] 1 Male
- [ ] 2 Female

H2 In what year were you born?   Year **19** [ ][ ]

AES 2016 - Page 18

H3 In which country were you born? And your mother? And your father?

Country of birth	Yourself	Mother	Father
Australia	1	1	1
New Zealand	2	2	2
United Kingdom	3	3	3
Republic of Ireland	4	4	4
Italy	5	5	5
Germany	6	6	6
Greece	7	7	7
China	8	8	8
India	9	9	9
Vietnam	10	10	10
Philippines	11	11	11
Other (please specify below)	12	12	12

Yourself [   ] Mother [   ] Father [   ]

H4 Do you identify yourself as Aboriginal or Torres Strait Islander?

- 1 Yes
- 2 No

H5 Did your father have any particular preference for one of the political parties when you were young, say about 14 years old? And how about your mother?

	Father	Mother
Liberal (includes UAP)	1	1
Labor	2	2
National/Country (includes Progressive Party)	3	3
Democratic Labor (DLP)	4	4
Australian Democrats	5	5
Other parties	6	6
Don't know	7	7

H6 What is your religion or faith?

- 1 Roman Catholic
- 2 Anglican/Church of England
- 3 Uniting Church/Methodist
- 4 Orthodox Church
- 5 Presbyterian
- 6 Other (please specify) [   ]
- 7 No religion

AES 2016 - Page 19

H7 Apart from weddings, funerals and baptisms, about how often do you attend religious services?

- ☐ 1 At least once a week
- ☐ 2 At least once a month
- ☐ 3 Several times a year
- ☐ 4 At least once a year
- ☐ 5 Less than once a year
- ☐ 6 Never

H8 What is your current marital status?

- ☐ 1 Never married
- ☐ 2 Now married (including de facto relationships)
- ☐ 3 Widowed
- ☐ 4 Divorced or separated

H9 Do you own outright, are you buying or renting the dwelling in which you now live?

- ☐ 1 Own outright
- ☐ 2 Own, paying off mortgage
- ☐ 3 Rent from private landlord or real estate agent
- ☐ 4 Rent from public housing authority
- ☐ 5 Other (boarding, living at home, etc.)

H10 Do you own any investment properties?

- ☐ 1 Yes
- ☐ 2 No

H11 Do you have a self-managed superannuation fund?

- ☐ 1 Yes
- ☐ 2 No

H12 In general, how often do you use the internet (on any device)?

- ☐ 1 Several times a day
- ☐ 2 About once a day
- ☐ 3 Three to five days a week
- ☐ 4 One to two days a week
- ☐ 5 Every few weeks
- ☐ 6 Less often
- ☐ 7 Do not use the internet ➜ **GO TO H15**

AES 2016 - Page 20

H13 When did you first start using the internet?

- 1 Within the last six months
- 2 About a year ago
- 3 Two or three years ago
- 4 More than three years ago

H14 Have you done any of the following tasks on the Internet?
Please cross all that apply

- Sent an attachment with an email
- Posted audio, video or image files
- Personally designed a webpage or blog
- Downloaded a software program to your computer

H15 Which social class would you say you belong to?

- 1 Upper class
- 2 Middle class
- 3 Working class
- 4 None

H16 Would you say you now live in …

- 1 A rural area or village
- 2 A small country town (under 10,000 people)
- 3 A larger country town (over 10,000 people)
- 4 A large town (over 25,000 people)
- 5 A major city (over 100,000 people)

H17 What is the gross annual income, before tax or other deductions, for you and your family or others living with you from all sources? Please include any pensions and allowances, and income from interest or dividends.

1 Less than $10,000 per year	12 $70,001 to $80,000 per year
2 $10,001 to $15,000 per year	13 $80,001 to $90,000 per year
3 $15,001 to $20,000 per year	14 $90,001 to $100,000 per year
4 $20,001 to $25,000 per year	15 $100,001 to $110,000 per year
5 $25,001 to $30,000 per year	16 $110,001 to $120,000 per year
6 $30,001 to $35,000 per year	17 $120,001 to $130,000 per year
7 $35,001 to $40,000 per year	18 $130,001 to $140,000 per year
8 $40,001 to $45,000 per year	19 $140,001 to $150,000 per year
9 $45,001 to $50,000 per year	20 $150,001 to $160,000 per year
10 $50,001 to $60,000 per year	21 $160,001 to $180,000 per year
11 $60,001 to $70,000 per year	22 More than $180,000 per year

AES 2016 - Page 21

H18 Do you own shares in any company listed on the Australian Stock Exchange (shares registered in your name or that of your family company)?

- [ ] 1   No, don't own any shares
- [ ] 2   Yes, one company
- [ ] 3   Yes, in two to five companies
- [ ] 4   Yes, in six to 10 companies
- [ ] 5   Yes, in more than 10 companies

H19 Would you be prepared to take part in a similar survey at the next Federal election?

If you choose to stay in the study we will retain your name and address and from time to time send you updates about the project, key findings and media activity. If you decide at a later stage you want to withdraw you will just need to contact us and we will delete your details.

- [ ] 1   Yes   a. Name: _____
-           b. Postal address: _____
-       OR   c. Email: _____ @ _____
- [ ] 2   No

> **The next few questions (H20 to H25)**
> **are ONLY for those who are MARRIED or living with a PARTNER.**
> **If you are NOT currently married or living in a de facto relationship,**
> **please skip this section and GO TO page 24.**

YOUR PARTNER

H20 What was your partner's main activity last week?

- [ ] 1   Working full-time for pay
- [ ] 2   Working part-time for pay
- [ ] 3   Unemployed—looking for full-time work
- [ ] 4   Unemployed—looking for part-time work
- [ ] 5   Retired from paid work
- [ ] 6   A full-time school or university student
- [ ] 7   Keeping house
- [ ] 8   Other *(please specify)* _____

AES 2016 - Page 22

H21 What kind of work does (or did) your partner do? Please give their full job title, and as much detail as you can. For public servants, state official designation and occupation. For armed services personnel, state rank and occupation.
If not working for pay now, please describe your partner's last regular paid job
If your partner has never worked for pay, please go to H25
a. Job title

b. What are (or were) the ***main tasks*** that your partner usually performs? Please give full details.
Main tasks

H22 Whom does (or did) your partner work for?
☐ 1 Self-employed
☐ 2 Employee in private company or business
☐ 3 Employee of Federal/State/Local Government
☐ 4 Employee in family business or farm

H23 As an official part of their job, does (or did) your partner directly supervise the work of other employees or tell other employees what work to do?
☐ 1 Yes
☐ 2 No

H24 Does your partner belong to a trade union?
☐ 1 Yes
☐ 2 No

H25 Generally speaking, does your partner usually think of himself or herself as Liberal, Labor, National or what?
☐ 1 Liberal
☐ 2 Labor
☐ 3 National (Country) Party
☐ 4 Greens
☐ 5 Other party *(please specify)*

☐ 6 No party

AES 2016 - Page 23

As a thank you for taking part in this important study, you have the chance to win in the prize draw competition. To enter, please provide your name and details for the Social Research Centre to contact you if your name is announced as the winner (the details included here will be used only for this purpose and will not be linked to survey results):

1. Name:

2. Phone: (   ) -

3. Email: @

*For full prize draw terms and conditions go to www.srcentre.com.au/aes.

Your views will remain strictly confidential and will be reported only as part of the general findings from the survey.

If you have any questions about the survey, you can call the toll free hotline on 1800 023 040, or email aesurvey@srcentre.com.au.

That is the end of the questionnaire.
Please put the questionnaire in the pre-paid reply envelope
and post it back to us.
Thank you again for your co-operation.

Australian Election Study
Reply Paid 90428
LAW COURTS VIC 8010

Barcode

AES 2016 - Page 24

Bibliography

[1] Nataraj Akkiraju, Herbet Edelsbrunner, Michael Facello, Ping Fu, Ernst P. Mücke, and Carlos Varela. Alpha shapes: definition and software. *Proceedings of the 1st International Computational Geometry Software Workshop*, 63:66–69, 1995.

[2] Daniele Albertazzi, Arianna Giovannini, and Antonella Seddone. 'no regionalism please, we are Leghisti!' the transformation of the Italian Lega Nord under the leadership of Matteo Salvini. *Regional & Federal Studies*, 28(5):645–671, 2018.

[3] David J. Allcroft and Chris A. Glasbey. A simulation-based method for model evaluation. *Statistical Modelling*, 3(1):1–13, 2003.

[4] Robert Andersen and Josh Curtis. Public opinion on social spending, 1980-2005. In Keith Banting and John Myles, editors, *Inequality and the Fading of Redistributive Politics*, chapter 6, pages 141–64. UBC Press Vancouver, 2013.

[5] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Publications in Statistics. Wiley, 1958.

[6] Stéfanie André. Does trust mean the same for migrants and natives? testing measurement models of political trust with multi-group confirmatory factor analysis. *Social Indicators Research*, 115(3):963–982, 2014.

[7] Santiago Anria. Social movements, party organization, and populism: insights from the Bolivian MAS. *Latin American Politics and Society*, 55(3):19–46, 2013.

[8] Arab Reform Initiative. *ArabBarometer*. University of Michigan and Princeton University, 2013. http://www.arabbarometer.org/.

[9] Benjamin Arditi. Populism as a spectre of democracy: a response to Canovan. *Political Studies*, 52(1):135–143, 2004.

[10] Aristotle (trans by T. A. Sinclair). *Politics*. Penguin, 1981 (first published 350 B.C.E.).

[11] Steven F. Arnold. *The Theory of Linear Models and Multivariate Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1981.

[12] Attorney-General's Department of the Commonwealth of Australia. *Australian Government Guidelines on the Recognition of Sex and Gender*, 2013.

[13] Vincent Audigier, François Husson, and Julie Josse. Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of statistical computation and simulation*, 86(11):2140–2156, 2016.

[14] Vincent Audigier, François Husson, and Julie Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10(1):5–26, 2016.

[15] Vincent Audigier, François Husson, and Julie Josse. MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and computing*, 27(2):501–518, 2017.

[16] Australian Bureau of Statistics. Australian Census, 2016.

[17] Australian Consortium for Social and Political Research Incorporated. *Australian Survey of Social Attitudes*. Australian National University, 2014. http://aussa.anu.edu.au/.

[18] Australian Electoral Commission. *Elector Count by Division, Age Groups and Gender for all States/Territories*. Commonwealth of Australia, 2016. https://www.aec.gov.au/Enrolling_to_vote/Enrolment_stats/elector_count/index.htm.

[19] Australian Electoral Commission. *First preferences by party*. AEC, 2019. https://tallyroom.aec.gov.au/HouseStateFirstPrefsByParty-24310-NAT.htm.

[20] Clive Bean. Nationwide electoral support for One Nation in the 1998 federal election. In Geoffrey Stokes Michael Leach and Ian Ward, editors, *The Rise and Fall of One Nation*, chapter 7, pages 136–152. University of Queensland Press, 2000.

[21] Clive Bean, David Gow, and Ian McAllister. *Australian Election Study 1998*. Australian National University, via the Australian Data Archive, 1998. http://www.australianelectionstudy.org/.

[22] Clive Bean, David Gow, and Ian McAllister. *Australian Election Study 2001*. Australian National University, via the Australian Data Archive, 2001. http://www.australianelectionstudy.org/.

[23] Clive Bean, Ian McAllister, Rachel Gibson, and David Gow. *Australian Election Study 2004*. Australian National University, via the Australian Data Archive, 2004. http://www.australianelectionstudy.org/.

[24] Clive Bean, Ian McAllister, Juliet Pietsch, Rachel Kay Gibson, and The Social Research Centre. *Australian Election Study 2013*. Australian National University, via the Australian Data Archive, 2013. http://www.australianelectionstudy.org/.

[25] Mabel Berezin. *Illiberal politics in neoliberal times: culture, security and populism in the new Europe*. Cambridge University Press, 2009.

[26] Gurminder K Bhambra. Brexit, Trump, and 'methodological whiteness': on the misrecognition of race and class. *The British Journal of Sociology*, 68:S214–S232, 2017.

[27] Antoine Bilodeau and Nada Fadol. The roots of contemporary attitudes toward immigration in Australia: contextual and individual-level influences. *Ethnic and Racial Studies*, 34(6):1088–1109, 2011.

[28] Norman E Binder, Jerry L Polinard, and Robert D Wrinkle. Mexican American and Anglo attitudes toward immigration reform: A view from the border. *Social Science Quarterly*, pages 324–337, 1997.

[29] George Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[30] George Box and Norman Draper. *Response surfaces, mixtures, and ridge analyses*. John Wiley & Sons, 2007.

[31] George Box, J Stuart Hunter, and William G Hunter. Statistics for experimenters. In *Wiley Series in Probability and Statistics*. Wiley, second edition, 2005.

[32] George Box, Alberto Luceno, and Maria del Carmen Paniagua-Quinones. *Statistical control by monitoring and adjustment*. John Wiley & Sons, 2011.

[33] John Brehm and Wendy Rahn. Individual-level evidence for the causes and consequences of social capital. *American Journal of Political Science*, pages 999–1023, 1997.

[34] Wayne Brittenden. *Political Compass*. Pace News Ltd, 2012. http://politicalcompass.org.

[35] Roderick D Buchanan. *Playing with Fire: The Controversial Career of Hans J. Eysenck*. Oxford University Press, 2010.

[36] Marius R Busemeyer, Achim Goerres, and Simon Weschle. Attitudes towards redistributive spending in an era of demographic ageing: the rival pressures from age and income in 14 OECD countries. *Journal of European Social Policy*, 19(3):195–212, 2009.

[37] Margaret Canovan. Trust the people! Populism and the two faces of democracy. *Political Studies*, 47(1):2–16, 1999.

[38] Margaret Canovan. Populism for political theorists? *Journal of Political Ideologies*, 9(3):241–252, 2004.

[39] Hadley Cantril. *Gauging public opinion.* Princeton University Press, 1944.

[40] Charles R Chandler and Yung-mei Tsai. Social factors influencing immigration attitudes: an analysis of data from the General Social Survey. *The Social Science Journal*, 38(2):177–188, 2001.

[41] Christopher Chatfield and Alexander J. Collins. *Introduction to multivariate analysis.* Springer, 2013.

[42] Simon Choat. *'Horseshoe theory' is nonsense – the far right and far left have little in common.* The Conversation, 12 May 2017. https://theconversation.com/horseshoe-theory-is-nonsense-the-far-right-and-far-left-have-little-in-common-77588.

[43] Becky L. Choma, Michael C. Ashton, and Carolyn L. Hafer. Conceptualizing political orientation in Canadian political candidates: A tale of two (correlated) dimensions. *Canadian Journal of Behavioural Science*, 42(1):24, 2010.

[44] Jack Citrin. Comment: The political relevance of trust in government. *The American Political Science Review*, 68(3):951–972, 1974.

[45] *Commonwealth Electoral Act 1918* (Cth).

[46] D. R. Cox. Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 105–123, Berkeley, Calif., 1961. University of California Press.

[47] D. R. Cox. Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24:406–424, 1962.

[48] David Crowe. *It's a Yes as SSM support gathers pace in the post: Newspoll.* The Australian, 13 November 2017. https://www.theaustralian.com.au/nation/politics/its-a-yes-as-ssm-support-gathers-pace-in-the-post-newspoll/news-story/d0f80b04859f38a7ee54fcd5c09d8b55.

[49] Roberto D'Alimonte. How the populists won in Italy. *Journal of Democracy*, 30(1):114–127, 2019.

[50] Rafaela Dancygier and Elizabeth N Saunders. A new electorate? comparing preferences and partisanship between immigrants and natives. *American Journal of Political Science*, 50(4):962–981, 2006.

[51] *Déclaration des droits de l'homme et du citoyen de 1789* [Declaration on the Rights of Man and the Citizen of 1789] (France), 26 August 1789.

[52] Swati Dhingra, Gianmarco IP Ottaviano, Thomas Sampson, and John Van Reenen. The consequences of Brexit for UK trade and living standards. *London School of Economics and Political Science Centre for Economic Performance*, 2016.

[53] Tarn Duong. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*, 21(7):1–16, 2007.

[54] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983.

[55] Maureen A Eger. Even in Sweden: the effect of immigration on support for welfare state spending. *European Sociological Review*, 26(2):203–217, 2009.

[56] Essential Research. *The Essential Report*. Essential Research, 7 November 2017.
https://www.essentialvision.com.au/wp-content/uploads/2017/11/Essential-Report_071117.pdf.

[57] Hans J Eysenck. *The Psychology of Politics*. Routledge and Kegan Paul, 1954.

[58] Hans J Eysenck. *Sense and Nonsense in Psychology*. Penguin Books, 1957.

[59] Hans J Eysenck. *Race, intelligence and education*. Temple Smith, 1971.

[60] Jean-Pierre Faye. *Le Siècle des idéologies*. Armand Colin, 1996.

[61] Stefano Fella and Carlo Ruzza. *Re-inventing the Italian Right: Territorial politics, populism and 'post-fascism'*. Routledge, 2009.

[62] Leonard Ferguson. Primary social attitudes. *Journal of Psychology*, 8:217–223, 1939.

[63] Leonard Ferguson. The measurement of primary social attitudes. *Journal of Psychology*, 10:199–205, 1940.

[64] Leonard Ferguson. The stability of the primary social attitudes: religionism and humanitarianism. *Journal of Psychology*, 12:283–288, 1941.

[65] Leonard Ferguson. An appraisal of the validity of the factor loadings employed in the construction of the primary social attitude scales. *Psychometrika*, 7(2):135–138, 1942.

[66] Leonard Ferguson. The isolation and measurement of nationalism. *Journal of Social Psychology*, 16:215–228, 1942.

[67] Leonard Ferguson. *Personality Measurement*. McGraw-Hill, 1952.

[68] Paolo Franzosi, Francesco Marone, and Eugenio Salvati. Populism and Euroscepticism in the Italian Five Star Movement. *The International Spectator*, 50(2):109–124, 2015.

[69] Daniel R. Fusfeld. Papers and proceedings of the eighty-fifth annual meeting of the American Economic Association: Types of radicalism in American economics. *The American Economic Review*, 63(2):145–151, 1973.

[70] Peter Gale. The refugee crisis and fear: Populist politics and media discourse. *Journal of Sociology*, 40(4):321–340, 2004.

[71] Marcel Gauchet. La droite et la gauche. In Pierre Nora, editor, *Les Lieux de mémoire*, volume 2, chapter 29, pages 2533–2601. Gallimard, 1997.

[72] Robin Gauld, Shaun Goldfinch, and Peter Herbison. The participation divide? Political participation, trust in government, and e-government in Australia and New Zealand. *Australian Journal of Public Administration*, 68(3):333–350, 2009.

[73] Rachel Gibson, Ian McAllister, and Tami Swenson. The politics of race and immigration in Australia: One Nation voting in the 1998 election. *Ethnic and Racial Studies*, 25(5):823–844, 2002.

[74] Stuart Gietel-Basten. Why Brexit? The toxic mix of immigration and austerity. *Population and development review*, pages 673–680, 2016.

[75] Matthew Goodwin and Caitlin Milazzo. Taking back control? Investigating the role of immigration in the 2016 vote for Brexit. *The British Journal of Politics and International Relations*, 19(3):450–464, 2017.

[76] Gouvernement de France. *Final results of the first round of the 2017 presidential election, drawn up by the Constitutional Council.* Gouvernement.fr, 26 April 2017. https://www.gouvernement.fr/en/final-results-of-the-first-round-of-the-2017-presidential-election-drawn-up-by-the-constitutional.

[77] John C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

[78] Michael Greenacre. *Correspondence Analysis in Practice.* CRC Press, 2007.

[79] Edward Guadagnoli and Wayne F. Velicer. Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2):265, 1988.

[80] Peter Hall, J.S. Marron, and Byeong U. Park. Smoothed cross-validation. *Probability theory and related fields*, 92(1):1–20, 1992.

[81] Lachlan Harris and Andrew Charlton. *The fundamental operating model of Australian politics is breaking down.* Sydney Morning Herald, 2 April 2018. https://www.smh.com.au/politics/federal/the-fundamental-operating-model-of-australian-politics-is-breaking-down-20180322-p4z5o9.html.

[82] Sharon Harris. Online quiz taken 20 million times! *The Liberator*, 18(21), November 1971.

[83] Yeheskel Hasenfeld and Jane A Rafferty. The determinants of public attitudes toward the welfare state. *Social Forces*, 67(4):1027–1048, 1989.

[84] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition, 2009.

[85] Patrick C. L. Heaven and John Connors. Politics and toughmindedness. *Journal of Social Psychology*, 128(2):217, 1988.

[86] Anna Henderson. *Indigenous recognition in the constitution overwhelmingly supported in new national poll*. Australian Broadcasting Corporation, 2015. https://www.abc.net.au/news/2015-03-27/poll-surveys-australians-on-indigenous-disadvantage/6350098.

[87] Tammy L Henderson, Pamela A Monroe, James C Garand, and Diane C Burts. Explaining public opinion toward government spending on child care. *Family Relations*, pages 37–45, 1995.

[88] John R. Hibbing and James T. Smith. Is it the middle that is frustrated? *American Politics Research*, 32(6):652–678, 2003.

[89] John R Hibbing and Elizabeth Theiss-Morse. *Congress as public enemy: Public attitudes toward American political institutions*. Cambridge University Press, 1995.

[90] M. Hills. Review of an introduction to multivariate techniques for social & behavioral sciences (book). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(3):339, 1977.

[91] John W. Hooper and Arnold Zellner. The error of forecast for multivariate regression models. *Econometrica*, 29(4):544–555, 1961.

[92] Rob Hyndman, Xibin Zhang, and Maxwell King. Bandwidth selection for multivariate kernel density estimation using MCMC. In *Econometric Society*, number 120 in Australasian Meetings. Econometric Society, 2004.

[93] Ronald Inglehart and Pippa Norris. Trump and the populist authoritarian parties: The Silent Revolution in reverse. *Perspectives on Politics*, 15(2):443–454, 2017.

[94] Ronald Inglehart, Bi Puranen, Thorleif Pettersson, Juan D Nicolas, and Yilmaz Esmer. *The World Values Survey*. 2015. http://www.worldvaluessurvey.org/.

[95] Ronald F Inglehart and Pippa Norris. Trump, Brexit, and the rise of populism: economic have-nots and cultural backlash. *Harvard Kennedy School Working Papers*, No. RWP16-026, 2016.

[96] Gilles Ivaldi, Maria Elisabetta Lanzone, and Dwayne Woods. Varieties of populism across a left-right spectrum: The case of the Front National, the Northern League, Podemos and Five Star Movement. *Swiss Political Science Review*, 23(4):354–376, 2017.

[97] O. A. Y. Jackson. Some results on tests of separate families of hypotheses. *Biometrika*, 55(2):355–363, 1968.

[98] Joanne Simon-Davies. Population and migration statistics in Australia. *Australian Parliamentary Library Research Paper Series, 2018-19*, 2018.

[99] Dieter W. Joenssen and Jürgen Vogel. A power study of goodness-of-fit tests for multivariate normality implemented in R. *Journal of Statistical Computation and Simulation*, 84(5):1055–1078, Nov 2012.

[100] Ian T Jolliffe. A note on the use of principal components in regression. *Applied Statistics*, pages 300–303, 1982.

[101] Roger Jones, Ian McAllister, David Denemark, and David Gow. *Australian Election Study 1993*. Australian National University, via the Australian Data Archive, 1993. http://www.australianelectionstudy.org/.

[102] Roger Jones, Ian McAllister, and David Gow. *Australian Election Study 1996*. Australian National University, via the Australian Data Archive, 1996. http://www.australianelectionstudy.org/.

[103] Karl G. Jöreskog. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443–482, 1967.

[104] Julie Josse and François Husson. missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[105] Karl C Kaltenthaler and Stephen J Ceccoli. Explaining patterns of support for the provision of citizen welfare. *Journal of European Public Policy*, 15(7):1041–1068, 2008.

[106] Dietlind Stolle Kenneth Newton and Sonja Zmerli. Social and political trust. In Eric Uslaner, editor, *The Oxford Handbook of Social and Political Trust*, chapter 3, pages 37–56. Oxford University Press, 2018.

[107] J. B. Kruskal. Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[108] Latin American Public Opinion Project. *AmericasBarometer*. Vanderbilt University, 2014. http://www.americasbarometer.org/.

[109] Sébastien Lê, Julie Josse, and François Husson. FactoMineR: an R package for multivariate analysis. *Journal of statistical software*, 25(1):1–18, 2008.

[110] Brian G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.

[111] J. C. Lester. The evolution of the political compass (and why libertarianism is not right-wing). *Journal of Social and Evolutionary Systems*, 17(3):231–241, 1994.

[112] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22 (140):5–55, 1932.

[113] Katerina Linos and Martin West. Self-interest, social beliefs, and attitudes to redistribution: re-addressing the issue of cross-national variation. *European Sociological Review*, 19(4):393–409, 2003.

[114] Anna Livsey and Andy Ball. *Same-sex marriage survey: religious belief matched no vote most closely*. The Guardian, 17 November 2017. https://www.theguardian.com/australia-news/datablog/2017/nov/17/same-sex-marriage-survey-religion-drove-the-no-vote.

[115] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, 2:49–55, 1936.

[116] Luke Mansillo and Nick Evershed. *Australian politics becoming more polarised*. The Guardian, 7 August 2014. https://www.theguardian.com/news/datablog/2014/aug/07/australian-politics-becoming-more-polarised.

[117] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Wiley, 1979.

[118] Stephanie Marken. *Still Listening: The State of Telephone Surveys*. Gallup, 2018. https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx.

[119] Rahsaan Maxwell. Trust in government among British Muslims: the importance of migration status. *Political Behavior*, 32(1):89–109, 2010.

[120] Ian McAllister. Border protection, the 2001 Australian election and the coalition victory. *Australian Journal of Political Science*, 38(3):445–463, 2003.

[121] Ian McAllister, Clive Bean, Rachel Kay Gibson, and Toni Makkai. *Australian Election Study 2016*. Australian National University, via the Australian Data Archive Dataverse, 2016. http://www.australianelectionstudy.org/ http://dx.doi.org/10.4225/87/7OZCZA.

[122] Ian McAllister, Roger Jones, Elim Papadakis, and David Gow. *Australian Election Study 1990*. Australian National University, via the Australian Data Archive, 1990. http://www.australianelectionstudy.org/.

[123] Ian McAllister and Anthony Mughan. *Australian Election Study 1987*. Australian National University, via the Australian Data Archive, 1987. http://www.australianelectionstudy.org/.

[124] Herbert McClosky and Alida Brill. *The dimensions of tolerance: What Americans believe about civil liberties*. Russell Sage Foundation, 1983.

[125] Jude McCulloch. National (in)security politics in Australia: Fear and the federal election. *Alternative Law Journal*, 29:87–91, 2004.

[126] Christopher J. Mecklin and Daniel J. Mundfrom. A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, 75(2):93–107, Feb 2005.

[127] Arthur H. Miller. Political issues and trust in government: 1964-1970. *The American Political Science Review*, 68(3):951–972, 1974.

[128] B. Mönch. *Der politische Wortschatz der französischen Restauration in Parlament und Presse*. Rheinische Friedrich-Wilhelms-Universität, 1960.

[129] Stephen Moore. Social scientists' views on immigrants and us immigration policy: A postscript. *Annals of the American Academy of Political and Social Science*, 487(1):213–217, 1986.

[130] Gary Morgan and Michele Levine. *Australians are voting in favour of same sex marriage*. Roy Morgan Research, 2017. http://www.roymorgan.com/findings/7368-same-sex-marriage-postal-plebiscite-survey-october-2017-201710130354.

[131] Ipsos MORI. *Immigration is now the top issue for voters in the EU referendum*. Ipsos MORI, 16th June 2016. https://www.ipsos.com/ipsos-mori/en-uk/immigration-now-top-issue-voters-eu-referendum.

[132] Cas Mudde. The populist zeitgeist. *Government and opposition*, 39(4):541–563, 2004.

[133] Denis Muller. *Outrage, polls and bias: 2019 federal election showed Australian media need better regulation*. The Conversation, 22 May 2019. https://theconversation.com/outrage-polls-and-bias-2019-federal-election-showed-australian-media-need-better-regulation-117401.

[134] National Health and Medical Research Council. *National Statement on Ethical Conduct in Human Research*, 2007 (updated 2018).

[135] National Health and Medical Research Council. *Australian Code for the Responsible Conduct of Research*, 2018.

[136] New Zealand Electoral Commission. *2017 General Election - Official Result*. Govt.nz, 2017.
https://www.electionresults.govt.nz/electionresults_2017/.

[137] David Nolan. Classifying and analyzing politico-economic systems. *The Individualist*, pages 5–11, January 1971.

[138] *NSW Registrar of Births, Deaths and Marriages v Norrie* (2014) 250 CLR 490.

[139] J Eric Oliver and Wendy M Rahn. Rise of the Trumpenvolk: Populism in the 2016 Election. *The Annals of the American Academy of Political and Social Science*, 667(1):189–206, 2016.

[140] Kevin H O'Rourke and Richard Sinnott. The determinants of individual attitudes towards immigration. *European Journal of Political Economy*, 22(4):838–861, 2006.

[141] Jason W. Osborne and Anna B. Costello. Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation*, 9(11):8, 2004.

[142] Jérôme Pagès. *Multiple factor analysis by example using R*. CRC Press, 2016.

[143] Jennifer Pan and Yiqing Xu. China's ideological spectrum. *The Journal of Politics*, 80(1):254–273, 2017.

[144] Fransisco Panizza. *Populism and the Mirror of Democracy*. Verso, 2005.

[145] Maryalice Parks. *Bernie Sanders Defends His Unwavering Focus on Economic Issues*. American Broadcasting Company, 2015.
https://abcnews.go.com/Politics/bernie-sanders-defends-unwavering-focus-economic-issues/story?id=35635871.

[146] Pamela Paxton. Association memberships and generalized trust: A multilevel model across 31 countries. *Social Forces*, 86(1):47–76, 2007.

[147] Pierre Bréchon. *The evolution of France's left and right politics, from the 1789 French Revolution to this year's election*. The Conversation, 2017.
https://theconversation.com/the-evolution-of-frances-left-and-right-politics-from-the-1789-french-revolution-to-this-years-election-76226.

[148] Nancy Postero. Morales's MAS government: building indigenous popular hegemony in Bolivia. *Latin American Perspectives*, 37(3):18–34, 2010.

[149] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[150] William Ramp and Trevor W Harrison. Libertarian populism, neoliberal rationality, and the mandatory long-form census: implications for sociology. *The Canadian Journal of Sociology*, 37(3):273–294, 2012.

[151] Rainer Riemann, Claudia Grubich, Susanne Hempel, Susanne Mergl, and Manfred Richter. Personality and attitudes towards current political topics. *Personality and Individual Differences*, 15(3):313 – 321, 1993.

[152] Steven Rose. Racism. *Nature*, 275:86, 1978.

[153] Steven Rose. Racism refuted. *Nature*, 274:738, 1978.

[154] Anealla Safdar. *Brexit: UKIP's 'unethical' anti-immigration poster*. Al Jazeera, 28 June 2016.
https://www.aljazeera.com/indepth/features/2016/06/brexit-anti-immigration-ukip-poster-raises-questions-160621112722799.html.

[155] Saundra K Schneider and William G Jacoby. Reconsidering the linkage between public assistance and public opinion in the American welfare state. *British Journal of Political Science*, 37(3):555–566, 2007.

[156] Holger Schultheis and Praneeth Naidu. Multi-model comparison using the cross-fitting method. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, pages 1389–1394, 2014.

[157] Howard Schuman and Stanley Presser. Question wording as an independent variable in survey analysis. *Sociological Methods & Research*, 6(2):151–170, 1977.

[158] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289 – 317, 2016.

[159] Roger N. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27(2):125–140, 1962.

[160] Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[161] Paul Spoonley and Andrew Butcher. Reporting superdiversity. the mass media and immigration in new zealand. *Journal of Intercultural Studies*, 30(4):355–372, 2009.

[162] Yannis Stavrakakis and Giorgos Katsambekis. Left-wing populism in the European periphery: the case of SYRIZA. *Journal of political ideologies*, 19(2):119–142, 2014.

[163] Robert Sternberg, Elena Grigorenko, and Kenneth Kidd. Intelligence, race, and genetics. *The American psychologist*, 60:46–59, 2005.

[164] Daniel Stockemet. *The Front National in France: Continuity and Change Under Jean-Marie Le Pen and Marine Le Pen*. Springer, 2017.

[165] Jon M Sutter, John H Kalivas, and Patrick M Lang. Which principal components to utilize for principal component regression. *Journal of chemometrics*, 6(4):217–225, 1992.

[166] Gábor J. Székely and Maria L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58 – 80, 2005.

[167] Jeff Taylor. *Where Did the Party Go?* University of Missouri Press, 2006.

[168] Paul Thureau-Dangin. *Le parti libéral sous la Restauration*. E. Plon, Nourrit, 1888.

[169] Warren S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[170] Sofia Vasilopoulou. Uk Euroscepticism and the Brexit referendum. *The Political Quarterly*, 87(2):219–227, 2016.

[171] Vox Pop Labs. *Vote Compass*. Vox Pop Labs, 2015. http://votecompass.com.

[172] Eric-Jan Wagenmakers, Roger Ratcliff, Pablo Gomez, and Geoffrey J. Iverson. Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48(1):28 – 50, 2004.

[173] A. M. Walker. Some tests of separate families of hypotheses in time series analysis. *Biometrika*, 54(1-2):39–68, 1967.

[174] Matt P. Wand and M. Chris Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116, 1994.

[175] Matt P. Wand and M. Chris Jones. Kernel smoothing. *Chapman & Hall*, 1995.

[176] Paul V. Warwick. Toward a common dimensionality in West European policy spaces. *Party Politics*, 8(1):101–122, 2002.

[177] Eric W Weisstein. *Matrix inverse*. Wolfram Research, 2014. http://mathworld.wolfram.com/MatrixInverse.html.

[178] Brian Wheeler, Paul Seddon, and Richard Morris. *Brexit: All you need to know about the UK leaving the EU*. British Broadcasting Corporation, 10 May 2019. https://www.bbc.com/news/uk-politics-32810887.

[179] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[180] D. A. Williams. Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, 26(1):23–32, 1970.

[181] D. A. Williams. In discussion of "A method for discriminating between models" by A.C. Atkinson. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(3):323–353, 1970.

[182] Shaun Wilson and Trevor Breusch. Taxes and social spending: The shifting demands of the australian public. *Australian Journal of Social Issues*, 38(1):39–56, 2003.

[183] Lei Xu and Michael I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1):129–151, 1996.

[184] YouGov. *YouGov / Fifty Acres Survey Results*. YouGov, 12th-16th October 2017.
https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/6dku1w0cx2/50AcresResults_171016_Wave13_W.pdf.

[185] Krzysztof Zagorski. Public opinion on taxes and government spending. *Australian Journal of Social Issues*, 23(4):300–308, 1988.

[186] Giada Zampano. *Italy's politics gives new life to anti-abortion campaign*. Politico, 18 November 2018.
https://www.politico.eu/article/italy-abortion-divide-politics-gives-new-life-to-anti-abortion-campaign/.

[187] Andrej Zaslove. *The re-invention of the European radical right: populism, regionalism, and the Italian Lega Nord*. McGill-Queen's Press, 2011.