Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics

Non-parametric Bayesian Forecasts Of Election Outcomes

Name of student: Banu Atav
Student ID number: 355108
Name of Supervisor: Prof. dr. R. Paap
Name of Second assessor: A. Castelein
Date final version: August 22, 2018

Abstract

Although there is much literature that covers election outcome forecasting, few to no methods of prediction have been able to consistently deliver accurate results. This problem essentially stems from the fact that election results are greatly influenced by idiosyncratic factors. This makes model selection difficult as, at the time of election, it is not clear which (type of) model will perform best. In this research, the problem of election forecasting is approached with a non-parametric Bayesian individual-level model using voting intentions and sociodemographic variables on Dutch elections in 2010 and 2012. Making use of a Dirichlet Process mixture (DPM) model, a flexible model specification is proposed. This specification is useful as the model's flexible nature allows it to be able to adapt to the characteristics of a new election. Furthermore, results of previous elections can be incorporated by adapting the prior specification. The results show that the DPM model improves on the forecast of the benchmark election models. Using the outcome of the DPM model applied to earlier years, forecasts of present elections can be further improved.

Keywords: non-parametric, Bayesian, Dirichlet Process, mixture modeling, election forecasting.

"I knew you were trouble when you walked in."

— T.A. Swift

1 Introduction

With pioneers dating back to the mid-20th century, predicting voting results of elections and other forms of ballot have been a significant scientific concern motivated by the desire to determine the drivers of an election outcome. The interest in and the application of forecasting election outcomes, however, is not merely limited to scientific literature. Driven by the abundant supply and widespread availability of public opinion polls and many sources of election outcome forecasts, forecasting election outcomes has become a large business. The evolution of voting preferences throughout the campaign period is closely tracked by the media, which influences the manner in which candidates and their campaigns are perceived. Hundreds of millions of dollars are invested in advertising and manpower which are allocated based on the knowledge about the topic (Linzer, 2013). Furthermore, this interest has translated into the emergence of prediction markets, where large amounts of money are invested, and various day-to-day forecasting websites that gained an enormous amount of popularity.

In the vast amount of literature, various types of election forecasting models are used to predict future election outcomes. Yet, there have been few to no methods of prediction, whether it be scientific or not, that have been able to consistently deliver accurate results under different circumstances. This leads to the main problem in forecasting election outcomes; model selection is difficult as it is unknown which model performs best under the new circumstances.

This research aims to alleviate part of this problem by taking a non-parametric Bayesian approach. Under a Dirichlet Process mixture model specification, individual-level voting intentions are modeled. The main rationale behind this choice is the model's flexible nature. This allows the model to be able to adapt to the characteristics of a new election, while it is still possible to include results of previous elections. Furthermore, the general model specification allows the model to be applied on various election systems, which expands and diversifies its track record.

Using the DPM model specification, where the prior distribution is a Dirichlet Process (DP), the joint distribution of the response variable y and covariates x are modeled by specifying appropriate distributions for $y|x$ (multinomial logit model) and x (multivariate normal and categorical distribution). Forecasts for the election outcome are made based on draws of the resulting posterior distribution. These draws are obtained using a hybrid MCMC sampler, where draws of the parameters of the multinomial logit model are generated using a Random Walk sampler, while a Gibbs sampler is used for all other parameters. Data pertaining to Dutch parliamentary elections for the years 2010 and 2012 are used. Using a sample of voting intentions y of eligible voters and their sociodemographic characteristics x, a forecast for the general election outcome is made with the DPM model and several standard benchmark models.

The results show that the DPM model has great potential to improve upon election forecasts of the benchmark models. This is especially true when the DPM model results of past elections are used to improve the model specification of the present election forecast. Unfortunately, the usage of the DPM model is no panacea. Challenges, such as finding a representative sample, make this a difficult topic.

This paper is organized in two parts. The first part outlines the existing literature and explains the problems in election forecasting into more detail. These insights are used to show in which way the current model specification can be useful. The second part is dedicated to the Dirichlet Process mixture model specification and the empirical application on the aforementioned data.

Part I

Modeling Issues In Election Forecasting

Most election forecasting models can be grouped under three categories. The political science literature is dominated by fundamental models. Driven by theoretical foundations, these models rely on macro-level information to predict voter share of the incumbent to estimate their long-term relationship to election results. The empirical counterpart to these models are the aggregators. These models predict election outcomes mainly using national polls and exploiting the benefits of quantitative methods. Lastly, synthesizers combine the previous two models to benefit from both sources of information.

Table 1: Absolute forecast errors for the US national popular vote under different models for 1992-2016 obtained from (PollyVote, nd).

	1992-2012	2016	1992-2016
Polls	2.8	1.6	2.6
Econometric Models	2.5	1.5	2.4
Expert Judgment*	1.5	2.1	1.8
Index Models**	1.4	2.5	1.8
Prediction Markets	1.3	4.8	1.6
Citizen forecast	1.2	1.2	1.2
PollyVote	1	1.9	1.1

* Forecast only available from 2004-2016.
** Forecast only available from 2008-2016

Table 1 presents an overview of the average absolute forecast errors of several types of models. These figures are determined based on a selection of models made for the PollyVote forecasts that predicts the US presidential elections' national popular vote. The figures for 1992-2012 seem to indicate that most models fall in the same range of average forecast error where two types of models seem to underperform. The data for 2016 shows, however, that the accuracy of the several types of models do not follow the same order as the average between 1992-2012.

The same argument can be made for the models that fall within a certain type of category. Table 2 is an example of this and shows the forecast errors of fundamental models predicting the national popular vote for US presidential elections. The figures show that, although the overall accuracy seems to have improved a lot since 2008, the model that predicts the outcome well in one year does not necessarily predict the outcome well in the next election year. This inconsistency is not limited to models within the fundamental model category and also applies to models that fall within the aggregate and synthesizer categories, the most popular example being the failure of Nate Silver's model (Silver, 2016b) to predict the 2016 US presidential elections after its great success in 2012.

Table 2: Forecast error of the US presidential elections' national popular vote for several fundamental models.

	Forecast error				
Forecaster	2016	2012	2008	2004	2000
Abramowitz	-2.5	-0.5	-1.4	5.4	2.9
Campbell	0.1	0.9	0.8	5.5	2.5
	(-0.4)	(0.2)	(-0.9)		
Cuzan (& Bundrick)		-4.2	0.9		
		(-5.5)			
Erikson & Wlezien	0.9	1.5	0.7	4.6 - 3.4	4.9
Holbrook	1.4	-3.2	-2.8	6.2	10
Lewis-Beck & Tien	-0.1	-2.9	-3.7	1.6	5.1
		(1.6)	(2.8)		
Lockerbrie	-0.7	2.7	-5.3	9.3	10
Norpoth (& Bednarczuk)	-3.6	2.1	2.8	6.4	4.7
Jerome & Jerome-Speziari	-1	0.5			

[1] The forecast errors are computed as $y - \hat{y}$, where y is the actual national popular vote of the incumbent and $\hat{y}$ the forecast of this popular vote.

[2] When there are multiple errors reported for one model at a given year, the items enclosed by parentheses are the forecast error of the researchers' least preferred forecast.

[3] The models are taken from the list of election forecasting models that appeared in the special US election issue of PS: Political Science & Politics. The selection has been made based on the number of forecasts published between 2000-2016 to make comparison possible.

[4] It should be noted that some of the listed forecasters altered their forecasting model in the reported time period, making the figures not completely comparable over time.

The lack of consistently accurate results is the main difficulty that lingers with forecasting election outcomes and essentially stems from the fact that election results are greatly influenced by idiosyncratic factors (Graefe, 2015). Evidently, an appropriate model is to be chosen prior to the event taking place. Therefore, the issue of idiosyncrasy enhances the difficulty that comes with the task of the selecting a forecasting model as one cannot know which type of model will perform well under the circumstances of the election at hand. This issue is further magnified as a result of the limitations in data. Elections are often held several years apart, due to which, depending on the type of the used data, the sample can be limited. Furthermore, elections that were held a significant amount of years ago may not carry information that is useful in predicting elections in the present. Yet, at the same time, it is desirable to evaluate models based on their overall track record for many elections to alleviate the model selection problem (Campbell et al., 2017).

An overlooked form of election forecasting is the individual-level model, which makes use of voting intentions and (mostly) sociodemographic variables to model election outcomes. These models have elements of both fundamental models as well as empirically driven models. Furthermore, due to the characteristics of the used data, they present the opportunity to opt for models that are more general, flexible and complex.

The remainder of part I is outlined as follows. In the next section, an overview of the existing literature is presented. This section discusses the aforementioned models and explains their approach into more detail. Furthermore, it sheds light on their advantages, identifies their limitations and explains in which way choosing an individual-level model aims to mitigate the issues that exist with forecasting elections. Section 3 explains how the choice of non-parametric Bayesian model can be beneficial in

this setting. Together, the information laid out in these sections provide the arguments that have driven the choices behind the model specification, which is discussed in Part II.

2 Literature and Modeling implications

The difference between existing forecast models is mainly driven by the type of data that is used. They are further diversified by different modeling assumptions that come with the data of choice. The political economy literature mostly deals with models that make use of macroeconomic and political data, mainly categorized as fundamental models, while aggregators and synthesizers are much less popular in the scientific literature. This section gives an overview of the numerous election forecasting models. In addition, it provides a discussion of the performance and limitations of these models to establish an overview of the model characteristics that are desired in election forecasting.

2.1 Fundamental Models

Fundamental models dominate the current literature and empirical forecasts (Lewis-Beck and Dassonneville, 2015a). These models assume a theoretical based approach where a political economy explanation of voting is used to build a model and make use of data that is independent from the election in question, such as economic and political data, that are in line with the theoretical framework (Hummel and Rothschild, 2013). The predictor variables are assumed to determine the voting outcome by some unknown function

$$\text{Vote}_\text{t} = f(\text{Economic Variables}, \text{Political Variables})$$

The model is estimated based on historical data which captures the relationship between the election outcome and the predictor variables of choice (Campbell, 2016a). The advantage of this type of models is the possibility to gain insight into the drivers behind voting behavior and thus election outcomes. Additionally, forecasts of the election outcome can be made far before any election related data is available and much before opinion polls have meaning (Hummel and Rothschild, 2014).

Lewis-Beck and Rice (1984) created what is now considered the classical political economy model by predicting voting results based on economic growth and political popularity. Many fundamental models that are currently applied have some similarities to this model. Much of the discussion and research pertaining to these models relates to selecting the correct predictor variables. It is save to say that all fundamental models recognize that voting behaviour is in some way related to the economic conditions in a country (Hibbs, 2008) and that voting behaviour is related to change in, not level of, economic conditions (Wlezien, 2015). The models do, however, disagree about whether voting is retrospective or prospective. Examples of economic covariates in accordance with retrospective voting are real GDP growth (Abramowitz, 2016; Lewis-Beck and Tien, 2016; Campbell, 2016b), real disposable personal income growth (Hibbs, 2008), unemployment rate (Lewis-Beck and Tien, 2008), national conditions relative to foreign countries (Kayser and Leininger, 2016) and index variables that are meant to capture economic conditions (Erikson and Wlezien, 2016; Holbrook, 2016). Prospective voting is less popular and is, for example, included by Lockerbie (2016) in the form of a questionnaire response variable in which voters are asked to predict the economic conditions in the future.

The economic conditions in the country reflect how voters perceive the incumbent has performed

to improve the countries' conditions and are one way to account for the voters' judgment of the incumbent. Essentially, the fundamental models work based on this principle; they perceive each election as a referendum on whether the incumbent should remain elected based on several factors that determine how the incumbent is evaluated (Campbell et al., 2017). Economic conditions are a significant factor in this evaluation, but research also suggests that there is a tendency to punish the incumbent regardless of the state of the economy or the popularity of the current president (Abramowitz, 2000). In general, current fundamental models account for this using a dummy variable for incumbency (Abramowitz, 2016), (the logarithm of) the time the incumbent has been in the White House (Lockerbie, 2016) or an interaction term for economic conditions and rerunning of the incumbent president (Holbrook, 2012).

Economic conditions and the overall tendency of punishing incumbents cannot completely explain the extent to which the incumbent is affected in terms of popularity among voters as there is a myriad of other factors at play. Hibbs (2012), for example, includes the sum of US military fatalities during term as his research shows that voters punish the president that initiated the unprovoked commitment of US forces as an additional variable. The main tendency to solve this issue, however, is to add political indicators to fundamental models. The indicators are measures that capture the overall voters' opinion of the candidates given current affairs and thus summarize the effects of all factors. Adding a variable of this sort seems to improve the model by mitigating the omitted variable problem. For fundamental models, adding more variables to obtain the same effect is not a viable option as these models typically have very few data points usually ranging from 15 to 25.

Frequently used political indicators are presidential approval ratings, which measures the overall popularity of the incumbents (Lewis-Beck and Tien, 2016; Abramowitz, 2016). An alternative to using popularity measures is to make use of a trial-heat variable that measures the candidates' performance in pre-election polls at a particular point in time, usually months from the election date (Erikson and Wlezien, 2016; Campbell, 2016b; Holbrook, 2016). Furthermore, the findings of Norpoth (2016) suggest that, historically speaking, presidential primary votes seem to be indicators of the winner of the presidential election.

From the outline of the various examples of fundamental models it is clear that the main issue this part of the field deals with is selecting the right variable that explains the share of votes the incumbent receives. There are countless approaches to this problem. The vast array of examples mentioned here is only a small extract of all fundamental model literature. The focus of this selection is mainly on most recent elections and models that have been tenaciously used over time.

2.2 Aggregators and Synthesizers

Thus far, this section gave an overview of main approaches among the copious election forecasting models that are based on theoretical foundations. The rival approaches, aggegrators and synthesizers, while being less popular, are much more empirically driven and almost non-existent in European election forecasting. Aggregators solely use empirical data to forecast the election outcome. Synthesizers, on the other hand, use both sources of information: empirical data of any sort and theoretical models. Among other advantages, as opposed to fundamental models, both types of models are dynamic in the sense that their forecast can be updated with recent information as the election nears, thereby foregoing the time advantage that the fundamental models posses.

With the explosive growth of the Internet, making collecting and communicating data easier, the supply of polls has become abundant (Blumenthal, 2014). As a result, the aggregator and synthesizer approaches, although almost exclusively used for the US national elections, have emerged in election

forecasting. Polling data, which measures voting intentions at a particular point in time, is a very rich information source at the same time as being very noisy (Linzer, 2013). Two main methodological points of consideration are faced in dealing with polling data. The first concerns generating the data which, next to collecting individual voting intentions, also consists of merging the data into an aggregate. Even though collecting polling data has become easier, there still are many challenges pollsters face in measuring polls. A few examples of these difficulties are identifying which voters to contact, how to survey, who is likely to vote and eventually how to appropriately weight the incoming information in order to reach a representative sample (Pasek, 2015). These difficulties make room for many sources of bias including sampling error, non-response and errors due to wording of the interview or other interview characteristics.

The second methodological concern is the manner in which polling data are processed to obtain a forecast. Over time, polls converge to an average and, on average, polls held close to election are very accurate. However, the observed data has a large variance, partially due to sampling error, which complicates identifying the actual movement of preferences from the noisy data at hand. Moreover, a poll is the result of voting intentions at a particular point in time and not the election day. This means that it cannot be directly interpreted as a forecast (Pasek, 2015; Blumenthal, 2014). Given that the voting intentions are measured before election day, the polling data is subject to a time-dependent error as voting intentions before the election day are likely to be different from voting elections on election day (Silver, 2014). The model that translates this type of data should, therefore, take these two factors into account.

The methodology of aggregators typically consists of combining multiple sources of polling data to arrive at a forecast for an election outcome. Due to the aforementioned factors, pollsters' methodological choices cause polls to have idiosyncratic errors. If these errors persist in a certain direction they lead to house effects. Consequently, these errors can be mitigated by combining many sources of data. By aggregating individual polls, the magnitude of systematic bias diminishes, as this bias is non-universal and has a smaller contribution when aggregated, but also because random errors are lower due to more information that is accounted for in the final forecast (Pasek, 2015). The techniques for combining polling data range from taking simple averages to much more complicated quantitative methods. Although taking a simple average seems naive, as it neglects the sample size of a poll, which is inversely related to its uncertainty, and the differences in quality between polls (Panagopoulos, 2009), to this date it is unclear which strategy is the best (Pasek, 2015). What is clear, however, is that aggregating can have benefits as the outcome is less variable and less susceptible to idiosyncrasies. This is beneficial to forecasting as a greater precision implies that smaller changes in preferences can be detected as most of the variability in polling data is not due to changes in preferences, but rather due to sampling error (Erikson and Wlezien, 1999).

While the methodology used by fundamental models is rather uniform, the opposite is true for aggregators and synthesizers. Rigdon et al. (2009) forecast US presidential election outcomes using a Bayesian approach that seems innate to election data. Starting out with previous election results, used to specify an informative prior, the model forecasts the winning probability for the candidates in each state by updating the prior believes with current polling data. Wang (2015), on other other hand, estimates the probability of a given state winning from the polling margin by assuming that the the median polling margin has a t-distribution. These probabilities are then combined to determine the probability distribution of all possible outcomes which is used to predict the electoral vote for each candidate.

There are a couple of models that received a lot of attention in the past decade. Due to his great success in forecasting the 2008 and 2012 US presidential elections, Nate Silver is probably the most well known one. Initially starting out with developing pollster ratings, Silver eventually converts this information

into a election forecasting model. The ratings are based on the observation that the performance of polls is predictable and that pollsters vary in accuracy. For example, the best polls in Silver's database are 1% more accurate than average, while worst perform 2.3% worse than average (Silver, 2014). Silver exploits this observation by constructing forecasts based on a weighted combination of the polling data with weights that are based on the pollster rating, while also accounting for sample size, recency[1] and different versions of the same poll held at different times. The latter is used to calculate a trend of the poll using loess regression to adjust the most recent data point used in the weighted combination. Next to a couple of other adjustments that are mainly US election specific, a house effect adjustment is included (Silver, 2016a). For recent estimations, a proxy that aims to account for polling firms' methodological standards is added (Silver, 2014). The polling aggregate is eventually used to create an estimate of how voting intentions are moving by regressing the aggregate[2] at state level on the partisan voter index. In effect, by including an attempt at accounting for all the difficulties that are faced with polling data outlined in the previous paragraph and more, three forecasts are created by blending the polling aggregate and the regressions where the polling aggregate receives more weight the closer we get to election day. Eventually, Silver computes a winning probability[3] (Silver, 2016a).

Alongside Nate Silver, Jackman and Linzer have created models that almost matched Silver's success in 2012, while taking completely different approaches. Jackman (2014) uses model-based poll averaging to estimate the outcome of the US presidential elections. His approach consists of using a national-level forecast, for instance one of the fundamental models' forecasts or an average thereof, to predict the national popular vote on election day. This is then used to determine the difference between the previous election outcome and the forecasted popular vote. This figure, called the uniform swing, reflects the change in voting preferences on national level. Jackman assumes that the state-level swing in voter preferences is constant over states and, thus, equal to the uniform swing, which is supported with the observation that the correlation between state-level outcomes is very high (0.98). This assumption is used to arrive at the forecasts, which is the sum of the previous state-level election result and the uniform swing (Jackman, 2014).

Linzer (2013), on the other hand, proposes a dynamic Bayesian approach where he combines the regression-based historical forecasting method and state-level polling data. To generate a forecast, he updates state-level fundamental forecasts using most recent polling data. Similar to Silver's strategy, older polls contribute less as these are only used to compute a trend in opinion polls. With this approach, both the fundamental as aggregate field are combined. The model gains its strength from a hierarchical definition and is build on the assumption that current state-level preferences are a function of state-level and national-level factors, both of which have been assigned a Bayesian reverse random-walk prior. The scales of the state-level and national-level effects are anchored on election day, by assuming the national-effects to be zero and the state-level effects to follow a prior distribution that is specified by means of the fundamental forecasts at this day.

Synthesizer models are not strictly dominated by complex quantitative models. The PollyVote forecast created by Graefe et al. (2016) and the work of Lewis-Beck and Dassonneville (2015a), for example, consist of weighted combinations of individual forecasts generated by different models. Lewis-Beck and Dassoneville combine their fundamental model forecast and poll data with fixed weights that favor poll data as the election approaches in Lewis-Beck and Dassonneville (2015a), while in Lewis-Beck and Dassonneville (2015b) they regress the forecasts of the two models on voter share in order to determine their weights. For the PollyVote forecast, the sources of data are not limited to fundamental model

[1] The rationale for this is that more recent polls carry more accurate information about voting intentions and therefore should receive more weight.

[2] Several other variables predictors are used where Silver takes three different strategies. These are blended on a later stage.

[3] After introducing a national, demographic and regional, and state-specific error drawn from a fat tailed distribution, this probability is determined by means of simulations.

forecasts and polling data. Data is obtained from trial-heat polls, prediction markets, fundamental models, expert judgment[4], index models[5] and citizen forecasts[6]. Forecast are created by taking the average within and across all components. Previous research (see for example (Graefe et al., 2015)) shows that equal weights tend to outperform more complex combining methods, possibly due to the inconsistency of the accuracy of various methods across time and different elections (Graefe et al., 2016). Therefore, equal weights are applied first within each component to determine a component and are then averaged again across components. This implies that each individual forecast does not obtain an equal weight in the final forecast. Hence, components with a small number of individual forecasts are assigned relatively more weight, while models with many forecast, such as fundamental models, do not dominate the forecast.

2.3 Limitations

The limitations of fundamental models are numerous. The small number of data points, typically around 20, is one of the major drawbacks of this approach and causes the models to only include very few explanatory variables. This implies that the predictions of these models are subject to high uncertainty and only in cases where the fundamental data clearly favors a candidate, predictions can be made with confidence (Linzer, 2013; Jackman, 2014). Furthermore, the focus of these models on finding a long-term relationship between fundamentals and election outcomes causes the forecasts to be based on the historic relationship between these variables. As elections generally occur about every 4 years, this means that this estimation is based on a relationship that is assumed to hold for more or less 80 years. Even assuming the historic relationship does hold for every new election to come, the approach completely fails at accounting for any idiosyncrasies that the election at hand may have. Models that aim to include some of the idiosyncrasy by using pre-election polls are not powerful enough. This is because the estimated regression weights are subject to high uncertainty and early polling results are not accurate (Linzer, 2013).

The econometric approach of most, if not all, fundamental models is a linear regression where a voter share is typically regressed on two or three fundamental variables. From a technical perspective, this method completely fails to recognize that voter share is not an unrestricted continuous variable. In addition, the model is limited to cases with only two running parties, where the voter share of one party is estimated, while the second is implied by the forecast. For the US case, this limitation causes models to disregard any third parties and to use the national popular vote as dependent variable. The latter implies that this model does not account for the, on state-level, winner-takes-all feature of the US electoral system, diminishing the relevance of the estimate of the national popular vote to generate forecasts even further[7]. This critique is supported by the fact that differences between popular vote and Electoral College are observed frequently (Rigdon et al., 2009). A handful of fundamental models recognize this issue and propose a state-level model (see for example Jerme and Jerme-Speziari (2016)), but this approach is far less common than the abundant national-level models (Jackman, 2014). Although it alleviates some of the issues national-level fundamental models face, such as the limited number of data points and restricted dependent variable, other issues arise due to the panel feature of the data and the limited availability of the required data on state-level.

While aggregators and synthesizers seem to have the answer to some of the problems faced with fundamental models, their methodology is subject to criticism too. Much of the discussion within this

[4] A panel of 15 experts is polled.

[5] Index models obtain forecasts based on regressions of vote share on various index variables regarding, for instance, characteristics of presidential candidate or economic conditions.

[6] These are constructed by means of surveys where citizens are asked to predict the winner.

[7] This argument also applies to models from the aggregate and synthesizer category that fail to account for this.

category relates to methodological differences among models which are rather model specific. One of the main points, however, is that poll data are technically not forecasts. These are rather snapshots of voting intentions at a particular time and only nearing election date, they come close to being forecasts (Kayser and Leininger, 2016). This means that, for forecasts produced based on polling data to have any meaning, they have to be generated close to election date, which is undesirable. In addition, even without sampling error, polling data has a large variance as it is highly influenced by sentimental moments such as party conventions (Pasek, 2015). Another point of criticism is related to pollster 'herding'. Even though polls converge to an average over time, it is not clear whether this is because voter preferences stabilize over time or because of tinkering in the polling results (Blumenthal, 2014). Given the strong popularity of polling results and the expectations of accuracy, there seems to be the incentive for pollsters to make ad hoc adjustments based on other pollster's results or discard results in order to not be an outlier. This is a significant problem as it decreases the amount of information captured in the data, while, perhaps more importantly, making all information dependent on each other (Silver, 2014).

It is widely accepted that elections are subject to more factors than just fundamental variables alone and that sometimes the effect of these factors may overshadow the effect of fundamental variables. However, fundamental model advocates argue that these factors are often temporary and different for each election and that, a priori, it is not possible to incorporate these in the model. The justification for fundamental models lies in that, overall, there seems to be a persistent historical relationship between election outcomes and the various fundamental variables. Additionally, forecasting based on other type of data does not yield any profound insight into drivers behind voting (Hibbs, 2013). Political scientists conclude that the focus should be on the general pattern as many of the idiosyncrasies are, despite media attention, not useful for predicting elections (Pasek, 2015).

The overwhelming amount of literature concerning election forecasting models presents a wide range of results that are contradicting to some extent. Empirical evidence suggest that the relative accuracy of different models cannot be put into a clear order; methods that perform well in one election often perform badly in others (Graefe et al., 2014). Past track records of these models support this ambiguity in relative performance both within and between types of models. Even though fundamental models have become more accurate over time and sometimes have remarkably accurate results (Linzer, 2013), there are periods where they can be extremely off or contradicting. In 2008 and 2012, the opinions regarding the victory of Obama were very divided. While fundamental models gave Obama 60% chance of winning in 2012, aggregate models thought this chance was about 90% (Wang, 2015). In 2008, predictions of well-known fundamental models ranged from clear wins to high probability wins for Obama according to Abramowitz (2008) and Lewis-Beck and Tien (2008), to a toss-up by Erikson and Wlezien (2008) and a clear loss by Campbell (2008). All the while, historically speaking, the national popular vote of 53.7% for Obama in 2008 is considered to be a significant margin.

Among fundamental models, a mean absolute forecast error of less than about 2% is classified as 'quite accurate' (Campbell and Lewis-Beck, 2008). This established benchmark is reflected in the satisfaction Campbell et al. (2017) voice in the accuracy of the reported models in 2016, where errors ranged from 0.1% to 3.6%. An error of this magnitude may be reasonable when the goal at hand is to determine the general drivers behind voting. When the aim is to forecast election outcomes, however, slight differences in the national popular vote can tip the election outcome (Rigdon et al., 2009). In fact, this was the case for the 2016 US election where Donald Trump won the election with only 48.9% of the popular vote; an occurrence only forecasted by Norpoth (2016) with 0.87 probability, which is based on a forecast that overpredicted Trump's share in the popular vote by 3.6 per cent.

Similar observations apply to comparison of results within the category off aggregators and synthesizers. After the "triumph of the quants" where several approaches, among which Nate Silver's, Drew

Linzer's and Simon Jackman's approaches, successfully predicted the 2012 US presidential election results (Jackman, 2014), it seemed as if consensus regarding the most preferred model for US election forecasting was reached. However, these models are now criticized for failing to predict the winner of the 2016 US presidential elections (Kennedy et al., 2017). In light of the unusual circumstances of this particular election, the failure to predict this event would have been expected of models that rely on a historic pattern, but came as a surprise given the nature of empirically driven models that, due to their use of polling data, should be able to account for idiosyncrasies.

Reflected in the overall interest in the media, scientific literature regarding election forecasting is mainly devoted to US (presidential) elections with some attempts for Western European countries mostly concerning Germany, France, and UK (Kennedy et al., 2017). Even though forecasting of election outcomes in European countries is becoming more popular, its literature is greatly under-represented and dominated by fundamental models. Aggregators are almost non-existent, while in the synthesizers category Lewis-Beck and Dassonneville (2015a) and Lewis-Beck and Dassonneville (2015b) are the only examples. Hence, little is known about whether existing methods are useful for many other countries and their electoral systems. Even though a general application of election forecasting models is desired and can be useful in gaining knowledge in this topic, the majority of the models specified for US election are too specific to be applied to European countries. Multi-party electoral systems where parliamentary coalitions are indirectly chosen, which are common in Europe, are one of the challenging features that current models often do not account for. Most models circumvent this issue by predicting the governing coalitions' voter share in the election outcome instead, but lose out on being able to draw a full picture of the election outcome. Kayser and Leininger (2016) justify this choice by arguing that estimating the percentage of votes for all parties is not a forecast of the outcome due to coalition bargaining that takes place after the election. However, when taking this approach one has to make the assumption that the current governing parties will continue to stay in power together if they are big enough after the elections, which in case of coalition forming may not be true.

2.4 Modeling Implications

Clearly, aggregating many sources of data and forecasts can improve accuracy. Given the ambiguity regarding the relative accuracy of different types of models and the fact that an appropriate model is to be chosen a priori, combining different models seems to be the best option. However, the main rationale behind Silver's aggregating methodology applies here: when constructing aggregates it is vital that one carefully selects models and weights them appropriately. Ideally, these decisions would be based on the track record of the various models. However, choosing appropriate models and weights is complicated due to the fact that elections only take place once in a couple of years. To validate a model in respect of accuracy one would have to track a model for many election cycles. Although the infrequency of elections cannot be resolved, additional knowledge on the track record and more can be gained from applying a forecast model to various countries and elections. This requires specifying a forecast model that can be generally applied.

Provided that the quality of the polls are up to par and the data is treated the proper way, polling data provide an excellent source of information that can capture any idiosyncratic factors and allow the forecasts to be updated as election nears. However, polling data only gains meaning close to the election date and does not provide any insight into the drivers behind voting behaviour. This is yet another reason why hybrid modeling may be fruitful, but a good initial starting point is needed. Most synthesizers make use of a structural forecasts as a starting point. Models that make use of long-term historic relationships, however, seem inadequate for forecasting.

An underrated form of election forecasting uses individual-level data, which provides an alternative to the structural counterpart of hybrid models. It provides advantages that are similar to the benefits of fundamental models. For example, forecasts can be made much in advance and various sources of data, such as economic conditions, can be included. In addition, it is a very natural departing point as voting behaviour is based on the voters' opinions, which are likely to be related to the voters' characteristics. Valuable empirical relationships, such as the fact that most people vote for the same party year after year (Jackman, 2014) and that previous voting behaviour is an important predictor for current voting preferences (Paap et al., 2005), can be exploited. As long as the sample is large enough and representative of the electorate, conclusions drawn from these voting choices apply to the underlying population.

Voting choice models based on individual-level data offer a solution to many of the difficulties that are faced in election forecasting. The data can be cross-sectional, which is relatively easy to gather, or panel data, which captures much more information. In both cases, compared to fundamental models, the number of data points and, as a result of this limitation, the uncertainty regarding parameter estimates are no longer of concern. Additionally, the data allows for more complex model specifications. Using multinomial discrete choice models, the model becomes suitable for many types of electoral systems, while including more predictor variables can improve the forecasts. Furthermore, even though some historical information may be of value, information regarding elections held decades ago not likely to carry information that is useful in predicting elections in the present. Individual-level data offers the possibility to include some historic information without limiting the model to long-term historic patters. Voting choice models are also useful in Bayesian approaches such as the approach described in Linzer (2013) where initial starting point is needed for prior specification.

In sum, the lack of preferred method as a results of inconsistent relative accuracy and the lack of a general application that can account for multi-party electoral systems indicates the need for additional research on the topic. As argued above, individual-level voting choice models have a great potential to be a solution for this. To meet the requirement of general application, however, a flexible model with the little model specific assumptions that can account multi-party response is needed.

3 Methodological Choices

In the previous section, the desired characteristics for an election forecasting model are established. As argued in this section, the Dirichlet Process Mixture model offers the possibility to account for these features.

When using data on an individual level, the simplest approach is to specify a model that is assumed to hold for every individual: a pooled model. However, this restriction may not always be true and differences between individuals may be such that one model cannot describe the relationship between input and outcome variables. For instance, it could be that parameters vary across individuals, i.e. there is unobserved heterogeneity in the parameters. Then, the heterogeneity in the parameters can be included in the econometric model by assuming that the parameters vary according to a distribution, which leads to the random parameters model or mixture modeling approach (Cameron and Trivedi, 2005).

$$p(d) = \int f(d|\theta, \lambda)\pi(\theta|\lambda)d\theta \tag{1}$$

By averaging the distribution of the data d conditional on the parameter(s) θ over a mixing distribution, the marginal distribution is obtained. By means of this approach flexibility is introduced in the assumed model. In fact, this approach enables us to pick a wide range of marginal distributions for d by simply varying the choice of $\pi(\theta|\lambda)$ (Rossi, 2014).

Equation (1) displays a model that assumes that the unobserved heterogeneity can be accounted for by a continuous distribution. In practical applications, one often benefits from limiting the heterogeneity assumption to simplify the model by using a finite mixture approach. In that case, the mixing distribution π is discrete and it is assumed that the data $d = (d_1, \cdots, d_n)$ are heterogeneous in the sense that there are K clusters. Each data point belongs to a certain cluster and each cluster has its own parametric model, i.e. cluster j has a distribution parametrized by θ_j. Prior to the analysis of the data, the unconditional probability of a data point belonging to a cluster j is described by π_j. This type of generative model, described in (2), has an intuitive interpretation: there may be K types or subpopulations and individuals within a cluster respond to changes in covariates in the same way (Cameron and Trivedi, 2005).

$$\begin{aligned} d_i|z_i, \theta_{z_i} \\ z_i|\pi \\ \pi = (\pi_1, \cdots, \pi_K) \end{aligned} \tag{2}$$

In this representation, π contains the mixing proportions and the variable z_i determines cluster membership such that data point i belongs to cluster z_i. Constructing the model in this way implies that the unconditional distribution of d is a weighted, linear combination of distributions as represented in equation (3).

$$p(d) = \sum_{j=1}^{K} \pi_j f(d|\theta_j) \tag{3}$$

This representation gives rise to an alternative interpretation: the weighted average of the distributions constitutes a good approximation of the empirical distribution of the variable of interest (Cameron and Trivedi, 2005). Depending on the choice of K and the values of the remaining parameters, this establishes a great amount of flexibility in the model. In fact, by choosing the right values of the parameters, one can approximate a continuous distribution of any shape. For example, the distribution can be multimodal, skewed, uniform near the mode or have fat tails as a consequence of the choice of K, π and θ. Counterintuitively, this feature is even more so present in finite mixtures than in the random parameter model. Partially, this is due to the fact that scaling is not enough to model asymmetric characteristics in the distribution. More importantly, when the model choices in a continuous random parameter model are not in line with the correct densities generating the data, the model estimates (and, most likely, its forecasts) will be biased. On the other hand, as long as enough clusters are chosen, the finite mixture approach will produce consistent results (Rossi, 2014). This benefit, along with the flexibility that this model specification provides, makes it a good candidate for election forecasting with individual-level data. Dividing the data into subpopulations can promote more accurate inference and forecasting. Furthermore, identifying clumps of individuals that behave in a similar way is appealing for, for example, formulation of election campaigns.

The most widely applied specification of this model is the mixture of normals, where it is assumed that the distribution of the data f is normal for all clusters with cluster specific parameter $\theta_j = (\mu_j, \Sigma_j)$. This choice for f is particularly appealing due to the fact that linear combinations of

normal distributions have tractable representations. Furthermore, the choice is often justified due to the flexibility argument given above; depending on the choice of mixing distribution and K, we can approximate any marginal distribution.

The common ways in which finite mixture models are estimated are maximum likelihood (ML) estimation and Bayesian analysis. Estimation with ML implies that the value of the parameter of interest is chosen such that the highest likelihood of the data being generated by the assumed model is obtained. Bayesian analysis, on the other hand, does not deal with optimization. It rather treats the parameter of interest as a random variable that, prior to the analysis, are postulated to have a certain distribution. This accommodates a systematic way in which any information can be included in the model estimation (Greenberg, 2012). This is particularly appealing for election forecasting models as it provides a natural way to influence the forecasting problem at hand with information obtained from previous elections. This may improve the forecasts as the dependence on previous election results is inherent to election data.

Besides this, Bayesian analysis offers numerous other advantages over ML. Firstly, it avoids any technical issues that accompany optimization problems. Possibly the most attractive feature of Bayesian analysis is the fact that it does not suffer from overfitting in its strict sense as it does not attempt to find the optimal fit of the data into the model. In addition, certain models that are computationally demanding, such as multinomial probit models with many categories, can be estimated relatively easily with Bayesian analysis.

The Bayesian representation of the model is similar to the above presented finite mixture model. Again, in this model K represents the number of components, π contains the mixing proportions and $\theta = (\theta_1, \cdots, \theta_K)$ describes the parameter(s) the researcher is interested in. The difference here is the introduction of prior distributions on unknown parameters. Often, the prior distribution for the mixing proportions is defined as the Dirichlet distribution (with hyperparameter α) due to the Multinomial-Dirichlet conjugacy. Any distribution suitable to the model and parameter characteristics can be picked as the prior of the component parameters θ, here denoted with H.

$$\begin{aligned} \pi|\alpha &\sim Dir(\frac{\alpha}{K}, \cdots, \frac{\alpha}{K}) \\ \theta_k^*|H &\sim H \\ z_i|\pi &\sim Mult(\pi) \\ d_i|z_i, \theta_k^* &\sim F(\theta_{z_i}^*) \end{aligned} \tag{4}$$

Despite the many advantages the (Bayesian normal) mixture approach has to offer, the approach is criticized for its parametric nature (Rossi, 2014). Given a space with all possible functions that attempt to describe the true model, a parametric method restricts the functional form to a set of possibilities by means of formulated assumptions about the true model. A mixture approach is essentially preferred when one would like to eliminate the bias caused by misspecification by increasing the flexibility of the model. Then, the modeling choice that leads to the ultimate flexible model is one with a highly parametrized model specification. A non-parametric model meets this criterion. Contrary to the parametric approach, non-parametric models put prior mass on all possible functional forms. This does not imply that no assumptions are made. As a matter of fact, the chosen model reflects a preference for simpler functions; these functions are assigned more mass prior to the analysis (Teh, 2013). A (finite) mixture model is not non-parametric as it cannot measure any distribution (Rossi, 2014). As the number of components cannot grow with n, no matter how large K is to be chosen, eventually the model will hit its limit as n approaches infinity (Broderick, 2015).

Several issues arise when opting for a parametric model. As a result of the desire for parsimony, the model can suffer from underfitting. Moreover, methods for model selection are somewhat arbitrary and different methods may lead to different conclusions. This makes choosing the right complexity for a model very difficult. A non-parametric method offers a solutions to these issues. A naive solution to the underfitting problem can be to always choose a large number of parameters in a parametric model, especially in case parsimony is not important. This leads to another issue in model selection. Models with an arbitrarily chosen large number of parameters can be subject to overfitting. Even though these models may perform well in explaining the data at hand, the forecasting performance can suffer from the consequences of overfitting. Forecasts may possibly be biased if the available data point to the wrong model. On top of that, the large number of parameters results in a loss of accuracy. One can alleviate the overfitting problem by taking a Bayesian approach. In a Bayesian approach (with proper priors) the overfitting problem is not significant as, due to the formulated priors, it mimics a shrinkage method; the priors impose a 'penalty' on the parameters (Rossi, 2014).

Taking a non-parametric approach comes very natural in Bayesian analysis. The main principle in Bayesian ideology is that when one encounters an unknown parameter, the parameter is treated as random. In that case, a prior distribution is introduced that reflects the existing beliefs about the parameter and the uncertainty around these beliefs. Hence, if the distribution of a parameter is unknown, a Bayesian puts a distribution on this parameter too. In the non-parametric approach, this prior is chosen such that the support of this distribution accounts for all possible distributions (Teh, 2010). Hence, a non-parametric Bayesian approach can be taken to counteract both issues. Here, the Bayesian approach ensures that the overfitting is mitigated, while the unbounded complexity that the non-parametric characteristic brings with, mitigates underfitting (Rossi, 2014; Teh, 2010).

One of the most popular and, in fact, the basis for many extensions of non-parametric Bayesian models is the Dirichlet Process Mixture (DPM) model, which is similar to the finite mixture model. The difference between the models is the fact that in the DPM model, the number of components and, therefore, the number of parameters is unlimited. Although one can try to choose a large number of components in a finite mixture model as an attempt to mimic this effect, the mere fact that the number of components is still fixed implies that this approach cannot make non-parametric claims (Rossi, 2014). The DPM approach facilitates a model with an infinite amount of parameters by allowing the number of parameters to grow with the number of data points n (Broderick, 2015). Hence, due to its infinite parameter space and Bayesian approach a DPM model provides us with a model that can account for the desired flexibility (Teh, 2013).

Part II

Dirichlet Process Mixture Model And Empirical Application

The previous sections have outlined the issues faced with election forecasting and argued that the main problems are a) the difficulty of model selection and b) the idiosyncratic factors elections are subject to. Section 2 of Part I concluded that an individual-level model can provide a solution to this problem to some extent. Among many other advantages, this source of data allows for a generally applicable model specification, due to which more knowledge can be gained on the track record of the model from applying it on various countries and elections. This alleviates the problem of model

election to some extent. Section 3 argued that a mixture approach may bring the desired flexibility. A non-parametric Bayesian approach is preferred to mitigate under- and overfitting, while maintaining this flexibility. More importantly, it allows the data to greatly influence the model due to which the model can be adapted to the needs of each individual election, while allowing past information to be included using the prior specification. In light of these considerations, the previous part concluded that a DPM model provides the desired model.

The focus of this part is twofold. First, the technical details of the DPM model are outlined. Second, the DPM model is put to a test with an empirical application on the dutch parliamentary elections in 2010 and 2012. The remainder of this part is organized as follows. First, Section 4 clarifies which data is used and discusses the characteristics of this data. These characteristics are necessary for proper model definition. Section 5 discusses the relevant definitions and representations of the general DPM model, while 6 discusses the adjustments and additional specifications needed in the DPM model to make it suitable for this research and the characteristics of the data. Section 7 presents the complete MCMC sampler for all parameters using the Gibbs sampler method discussed in earlier sections. In Section 8, the approaches to detecting convergence and generating forecasts are discussed. Section 9 completes the model specification by specifying the values of hyperparameters. Multiple sets of prior specification are used to detect the influence of the prior on the final outcome. Finally, the last two sections discuss the results and conclusions of this research.

4 Data

The data for voting choices are obtained from the LISS Panel collected by CentERdata. The panel consists of roughly 7000 subjects situated in the Netherlands who complete monthly online surveys related to various topics in social sciences since October 2007. The sample of subjects is composed such that it is in line with the true probability sample obtained from Statistics Netherlands. The data of interest is taken from three different questionnaires: Background Variables (BV), Religion and Ethnicity (RE), and Politics and Values (PV).

Nine waves of the PV data set are available. These waves correspond to the years 2007-2013, 2015 and 2016. During each wave two types of questions regarding to voter preferences are posed. The individuals are asked for their party choice during a particular election as well as their hypothetical voting choices if an election were to be held on the day of the questionnaire. The actual voter choices can be obtained for the Parliamentary elections held in years 2006 (waves 1-3), 2010 (waves 4-5) and 2012 (waves 6-9). Although the data correspond to a panel structure, the DPM model defined in the upcoming sections does not make use of this structure. Instead, a sample of all relevant variables is obtained for the years 2010 and 2012 separately, which are the only two years for which the complete data set is available at the time of research[8]. As explained in Section 10, the model is then estimated for each year separately. This way, model performance can be assessed over multiple years.

The response variable and covariates are selected based on variables used in previous literature (Paap et al., 2005; Quinn et al., 1999). The, by far, dominating approach is to use hypothetical voting choices on the day of the questionnaire as response variable and to use sociodemographic respondent characteristics to model party preferences. The following explanatory variables are selected: Age, Individual Income, Gender, Education, Civil Status, Home Ownership, Urban character of place of residence, Primary Occupation and Religion. All sociodemographic variables but one are obtained

[8]Each variable is obtained from the wave of data collected closest to the relevant election date. For all covariates and voting intentions this corresponds to the most recent wave prior to the election date, while for the actual voting choices this corresponds to the wave right after the election date.

from the BV questionnaire. Only the variable Religion is obtained from the RE questionnaire.

Paap et al. (2005) show that modeling state dependence (by including previous election choices in the model) improves the predictive power of the model. This observation is supported by the data selected for this study. Therefore, the party choice of the individuals in the previous election is also added to the list of explanatory variables. Given the model specification, state dependence can be added to the model by viewing this variable as an ordinary categorical variable using dummies. Another method would be to use alternative specific parameters that account for whether the individual has chosen the alternative in the previous election or not. Although the latter would be preferred in the sense that it leads to a smaller model, the former option showed much better results with the data at hand. This is in line with the findings of Paap et al. (2005). Hence, state dependence is modeled by means of dummy variables only.

In the interest of reducing the size of the model, the response variable is transformed into fewer number of alternatives. In line with the approach taken in previous literature, only the four parties that received the largest party shares in 2010 are listed separately. All other parties are lumped together into alternative 'Other'. The options 'Do not know' (DNK) or 'Prefer not to say' (PNTS) are merged into one category DNK/PNTS. Eventually, the voter preferences are described by alternatives VVD, PvdA, PVV, CDA, Other, DNK/PNTS, No Vote. Some studies choose to drop all cases of DNK/PNTS and No Vote and proceed the analysis with the remaining part of the data. Disregarding this information could, however, lead to selection bias. Although this is a serious issue and much literature is devoted to it, this is not the main concern of this study. Therefore, it is decided to keep these alternatives in the model at an attempt to resolve this issue to some degree.

Some of the sociodemographic variables are altered as well. Civil Status and Home Ownership are recoded into binary variables as most previous literature do not include any other options either. Urban character of place of residence is recoded into a four-category variable as one of the categories showed no explanatory power at all. Primary Occupation originally measures 14 different categories that aim to describe the occupation of respondents rather detailed. Previous literature mainly includes a variable regarding whether the individual is (un)employed. Therefore it has been recoded into options Employed (paid), Unemployment and Other. Lastly, the variable Religion originally specifies 14 different religions. The variable is recoded to only specify religions that are adopted by a significant amount of respondents and/or seem to have an effect on voting choices[9]. Lastly, the variables Age and Income are log-transformed as their distribution is skewed. This way the data conform more closely to the normal distribution, which simplifies its prior specification in the upcoming sections.

5 Dirichlet Process Mixture Models

This section is divided into four parts that discuss a set of definitions and concepts related to the Dirichlet Process Mixture model. The first section gives a formal definition of the Dirichlet prior and explains its function in the DPM model. The second and third sections are different representations off the DPM model that are necessary to develop the intuition in the workings of the model. Furthermore, the latter representation seems to be particularly interesting when deriving the Gibbs sampler. Finally, the Gibbs sampler that is necessary to sample the parameters of the DPM model used in this paper is presented.

[9]The (magnitude of the) effect of religious beliefs on voting choices differs for each party. Religions that seem to have a strong effect or an effect on most parties are kept in the data set.

5.1 Dirichlet Process and General DPM Model

For non-parametric Bayesian analysis, prior distributions are desired to have two properties: they should have a large support and the resulting posterior should be analytically manageable (Ferguson, 1973). Even though these two properties are difficult to be obtained at the same time, based on the conjugacy of the Dirichlet and Multinomial distribution, Ferguson (1973) developed a distribution, the Dirichlet process, that meets these criteria.

The Dirichlet process (DP) is a stochastic process: a distribution over functions. In this case, these functions happen to be probability distributions. Due to this, the DP is a distribution over distributions, which has properties that are useful in non-parametric Bayesian analysis (Teh, 2010). In particular, it is used as a prior for the mixing proportions in an infinite mixture model.

Formally, the DP is defined as a distribution of probability measures on a partition of some space (sigma-algebra) in the following way. Let G_0 be a distribution over $\mathcal{X}$ and α be a positive, real number. Furthermore, let $(A_1, \cdots, A_k)$ be an arbitrary, measurable partition of $\mathcal{X}$. Then, by the definition of the DP, if G is a DP denoted as $G \sim DP(\alpha, G_0)$, the following must hold (Ferguson, 1973; Rossi, 2014)

$$(G(A_1), \cdots, G(A_k)) \sim Dir(\alpha G_0(A_1), \cdots, \alpha G_0(A_k)) \tag{5}$$

Hence, the DP has marginal distributions that are Dirichlet distributed, based on which the distribution obtains its name (Hjort et al., 2010; Teh, 2010). As can be seen from (5), the DP has two parameters. The following two derived properties indicate the interpretation of these parameters. For any measurable set A it holds that[10]

$$\begin{aligned} E[G(A)] &= \frac{\alpha G_0(A)}{\alpha} = G_0(A) \\ V[G(A)] &= \frac{\alpha G_0(A)(\alpha - \alpha G_0(A))}{\alpha^2(\alpha+1)} = \frac{G_0(A)(1-\alpha G_0(A))}{\alpha+1} \end{aligned} \tag{6}$$

This implies that the parameter G_0, the base measure, can be interpreted as the mean or location of the DP. On the other hand, α is a tightness (or inverse variance) parameter that determines how tightly the DP is distributed around the base measure. The larger α is, the smaller the variance with which the DP is dispersed around its base measure or mean. Due to this, when α approaches infinity, $G(A)$ will approach $G_0(A)$ (Rossi, 2014; Teh, 2010).

Since draws from a DP are distributions, they can be used to describe random variables. This random variable[11] can be defined as

$$\begin{aligned} \theta|G &\sim G \\ G|\alpha, G_0 &\sim DP(\alpha, G_0) \end{aligned} \tag{7}$$

This feature is exploited in a DPM model where each data point d_i has a distribution F depending

[10] These derivation are based on the properties of a Dirichlet distribution given the fact that $(G(A_1), \cdots, G(A_k))$ are Dirichlet distributed.

[11] As G is defined on $\mathcal{X}$, the random variable θ takes values in this set too.

on the data point specific parameter θ_i which is a DP random variable

$$\begin{aligned} d_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G|\alpha, G_0 &\sim DP(\alpha, G_0) \end{aligned} \tag{8}$$

F and G_0 may depend on more hyperparameters, which can be given priors as well. This model can be seen as the infinite limit of the finite mixture model. This property is explored in the next section.

5.2 Relation to Finite Mixture Model

The finite mixture model as shown in (2) under an infinite limit was initially explored, for example by Neal (1992), to avoid determining the number of components and to be able to let the number of components grow with sample size (Teh, 2010). Nowadays, it is widely known as the DPM model. Using the finite mixture model representation, we can show the relationship between the two models, which gives insight into the characteristics of a DPM model. As stated in section 3, the Bayesian representation of the finite mixture model is[12]

$$\begin{aligned} d_i|z_i, \theta_j^* &\sim F(\theta_{z_i}^*) \\ z_i|\pi &\sim Mult(\pi) \\ \theta_j^*|G_0 &\sim G_0 \\ \pi|\alpha &\sim Dir(\frac{\alpha}{K}, \cdots, \frac{\alpha}{K}) \end{aligned}$$

Given this model definition, the posterior distribution of the classification of a particular data point i can be derived by integrating out the parameters. For this we first augment the probability of interest, $p(z_i = j|z_1, \cdots, z_{i-1})$, with the parameters π

$$p(z_i = j|z_1, \cdots, z_{i-1}) = \frac{p(z_1, \cdots, z_{i-1}, z_i = j)}{p(z_1, \cdots, z_{i-1})} = \frac{\int p(z_1, \cdots, z_{i-1}, z_i = j, \pi)d\pi}{\int p(z_1, \cdots, z_{i-1}, \pi)d\pi}$$

By Bayes' rule this is equal to

$$p(z_i = j|z_1, \cdots, z_{i-1}) = \frac{\int p(\pi)p(z_1, \cdots, z_{i-1}, z_i = j|\pi)d\pi}{\int p(\pi)p(z_1, \cdots, z_{i-1}|\pi)d\pi}$$

This expression can be specified according to the generative model definition

[12]Two arguments apply for the manner in which the Dirichlet distribution is parametrized in the prior distribution. First, the strength of the Dirichlet prior depends on the number of components. More components, inherently give more strength to the prior due to the flexibility they create. For this reason, the sum of the parameters of a Dirichlet prior is often seen as the strength of the prior. Therefore, the prior's strength should be adjusted for accounting for the number of components (Teh, 2013). Secondly, as clarified in the remainder of this section, this particular symmetric parametrization ensures that the number of components used to model is independent from K in its infinite limit as α/K approaches zero when K approaches infinity (Neal, 2000).

$$\frac{\int p(\pi)p(z_1,\cdots,z_{i-1},z_i=j|\pi)d\pi}{\int p(\pi)p(z_1,\cdots,z_{i-1}|\pi)d\pi} = \frac{\int \Gamma(\alpha)\Gamma(\alpha/K)^{-K}\pi_1^{\alpha/K-1}\cdots\pi_K^{\alpha/K-1}\pi_{z_1}\cdots\pi_{z_{i-1}}\pi_j d\pi}{\int \Gamma(\alpha)\Gamma(\alpha/K)^{-K}\pi_1^{\alpha/K-1}\cdots\pi_K^{\alpha/K-1}\pi_{z_1}\cdots\pi_{z_{i-1}} d\pi}$$

and further simplified to

$$\frac{\int \Gamma(\alpha)\Gamma(\alpha/K)^{-K}\pi_1^{\alpha/K+n_1-1}\cdots\pi_K^{\alpha/K+n_j}\cdots\pi_K^{\alpha/K+n_K-1} d\ \pi}{\int \Gamma(\alpha)\Gamma(\alpha/K)^{-K}\pi_1^{\alpha/K+n_1-1}\cdots\pi_K^{\alpha/K+n_K-1} d\pi}$$

where n_k is the number of z_k equal to k for $k \neq j$. As can be seen from this expression, the two distributions have the kernel of a Dirichlet distribution, which can be integrated out after accounting for the proper normalizing constant, so that the expression finally simplifies to

$$\frac{\prod_{k\neq j}\Gamma(\alpha/K+n_k)\Gamma(\alpha/K+n_j+1)\Gamma(\alpha+i+1)^{-1}}{\prod_{k=1}^{K}\Gamma(\alpha/K+n_k)\Gamma(\alpha+i)^{-1}} = \frac{n_j+\alpha/K}{i-1+\alpha}$$

When we let K go to infinity, the relevant probabilities become

$$\begin{aligned} p(z_i=j|z_1,\cdots,z_{i-1}) &= \frac{n_j}{i-1+\alpha} \\ p(z_i\neq j|z_1,\cdots,z_{i-1}) &= \frac{\alpha}{i-1+\alpha} \end{aligned} \tag{9}$$

Hence, for large K, the number of components used to model N data items is independent of K. This implies that the model remains well defined as K approaches infinity. These probabilities are sufficient to define the model in the sense that the categorical variables z_i are only important in whether or not they are equal to other z_i (Neal, 2000).

The equivalence between the two 'different' models can be shown by making use of

$$\begin{aligned} \theta_i|G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{10}$$

By integrating G out of the model, we can derive the Blackwell-MacQueen Polya Urn representation; the distribution of a draw θ_i conditional on previous draws. In order to arrive to this distribution, we start by considering the posterior $G|\theta_1$ after one draw. Bayesian theory dictates that the posterior distribution $p(G|\theta_1)$ is proportional to prior times the likelihood, $p(G)p(\theta_1|G)$. Under a given partition of $\mathcal{X}$, G is Dirichlet distributed (see (5)). Due to the fact that DP random variables are inherently clustered and the fact that the probability of each cluster given G is fixed, $P(\theta_1 \in A_j|G) = G(A_j)$, $\theta_1|G$ has a categorical distribution with the aforementioned cluster probabilities. Since the posterior distribution given a partition of $\mathcal{X}$ is proportional to the product of a Dirichlet and a categorical distribution, we can make use of the Dirichlet-Multinomial conjugacy. Hence, the posterior for a given partition of $\mathcal{X}$ is

$$(G(A_1), \cdots, G(A_k))|\theta_1 \sim Dir(\alpha G_0(A_1) + n_1, \cdots, \alpha G_0(A_k) + n_k)$$

Here, n_i is the number of draws in subset A_i. Using (10), this implies that

$$G|\theta_1 \sim Dir(\alpha + 1, \frac{\alpha G_0 + \delta_{\theta_1}}{\alpha + 1})$$

where δ_{θ_1} is a point mass. To arrive at the distribution of the next draw θ_2 conditional on θ_1, we integrate G out of the generative model

$$\theta_2|G, \theta_1 \sim G|\theta_1$$

Now we can marginalize G out given that $P(\theta_1 \in A_j) = E[G(A_j)] = \frac{\alpha G_0 + \delta_{\theta_1}}{\alpha+1}$ (Teh, 2010)

$$\theta_2|\theta_1 \sim \frac{\alpha G_0 + \delta_{\theta_1}}{\alpha + 1}$$

Repeating this idea for multiple draws implies

$$(G(A_1), \cdots, G(A_k))|\theta_1, \cdots, \theta_n \sim Dir(\alpha G_0(A_1) + n_1, \cdots, \alpha G_0(A_k) + n_k) \tag{11}$$

As this is true for any arbitrary partition of $\mathcal{X}$, the posterior of G must be a DP. In fact, given that the $\sum_{i=1}^{n} \alpha G_0(A_i) + n_i = \alpha + n$, the posterior is a DP with concentration parameter $\alpha + n$. Moreover, the base distribution of the constructed DP is $\frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n}$, a result that can be achieved by factoring out $\alpha + n$ from all $\alpha G_0(A_i) + n_i$. This implies the result shown in (12).

$$G|\theta_1, \cdots, \theta_n \sim Dir(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}) \tag{12}$$

and the Blackwell-MacQueen representation

$$\theta_n|\theta_1, \cdots, \theta_{n-1} \sim \frac{1}{n - 1 + \alpha} \sum_{i=1}^{n-1} \delta(\theta_i) + \frac{\alpha}{n - 1 + \alpha} G_0 \tag{13}$$

where $\delta(\theta_i)$ is the probability mass concentrated at θ_i. Hence, the predictive distribution of θ_i is a mixture of the empirical distribution of all previously observed values, $\delta(\theta_i) : i \in 1, \cdots, n-1$ and the base measure G_0. As the sequence of draws is exchangeable, the labeling of the draws in the derivation does not matter; the same result applies for any draw.

The Polya Urn representation of the DPM model in (13) obtains its name from its interpretation based on an urn metaphor. For this interpretation, we start out with an empty urn and draw a ball from G_0, the value of which represents a colour. This ball is then dropped into the urn. After this, subsequently new balls are drawn. With probability proportional to the number of balls of colour x in the urn, a ball with colour x is drawn. If this is the case, the drawn ball is replaced and another ball with colour x is added to the urn. With probability proportional to α a ball with a new colour is drawn from G_0 and one ball of that colour is dropped in the urn (Teh, 2010). This scheme has been used to show that the DP prior exists by Blackwell and MacQueen (1973).

Using the results in (9) and (13), we can now see the correspondence between the limiting probability of the finite mixture model and the DPM model; the limiting probabilities in (9) are the Blackwell-Macqueen conditional probabilities of either drawing from an existing cluster (previous draws of θ_i) or creating a new cluster for the new draw from G_0. The limit of the finite mixture model becomes equivalent to the DPM model when we take θ_i from (13) equal to $\theta^*_{z_i}$ from (9) (Neal, 2000). The DP can be constructed from the infinite limit of the finite mixture model (taking $K \to \infty$) of the random probability measure $\sum_{k=1}^{K} \pi_k \delta_{\theta^*_k}$ (Teh, 2010).

The result in (13) implies several important properties of the DP distribution. First, we can infer that the resulting posterior can be discrete no matter how smooth G_0 is. For a long enough sequence of draws, a value will be repeated. Secondly, larger clusters have a larger posterior probability of being drawn (Rossi, 2014). This is called the rich-gets-richer property (Teh, 2013). Lastly, as can be seen from the posterior, the larger α, or as previously mentioned, the strength parameter is, the more informative the prior becomes. Hence, when the number of observations grows, the empirical distribution dominates the posterior. This indicates that the posterior approaches the true distribution for large number of observations (Teh, 2010).

Besides the representation as the infinite limit of the finite mixture model, the DPM model has two more representations, the Chinese Restaurant Process and the stick-breaking representation. The latter proves to be more useful in the practical application of the model and is explored in the next section.

5.3 Stick-breaking representation

The previous section has shown that there is a direct link between the finite mixture model and the DPM model. This equivalence occurs when we take the number of components in a finite mixture model to approach infinity. In theory, this explanation is valid. In practice, however, constructing a DPM model in this way is not feasible since one cannot draw an infinite number of mixing probabilities from the finite mixture representation of the mixture model.

We can, however, make use of a different construction of the Dirichlet distribution called the stick-breaking construction. This construction makes use of the fact that if $\pi = (\pi_1, \cdots, \pi_k)$ is Dirichlet distributed, the marginal of π_1 has a Beta distribution and the remaining π_j conditional on π_1 are Dirichlet distributed (Broderick, 2015). More precisely, if $\pi = (\pi_1, \cdots, \pi_k) \sim Dir(\alpha_1, \cdots, \alpha_K)$, then

$$\pi_1 \sim Beta(\alpha_1, \sum_{j=2}^{K} \alpha_j)$$

$$\frac{(\pi_2, \cdots, \pi_k)}{1 - \pi_1} \sim Dir(\alpha_2, \cdots, \alpha_K)$$

Sethuraman (1994) uses this idea to show that the DP $G \sim DP(\alpha, G_0)$ can be constructed from a series of Beta random variables. The metaphorical meaning of the stick-breaking construction follows from interpreting the sum of cluster probabilities $\sum_{k=1}^{\infty} \pi_k$ as a the length of a stick. Starting from the first iteration $i = 1$, we draw a Beta random variable to determine π_1 and break off a piece of the stick of length π_1. The piece of the stick that remains is used to determine the following cluster probabilities; we recursively draw β_i and break off π_i by means of the following identities (Teh, 2010)

$$\begin{aligned} \beta_k &\sim Beta(1, \alpha) \\ \pi_k &= \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \end{aligned}$$

A DP can be constructed from the cluster probabilities in the following way

$$\begin{aligned} \theta_k^* &\sim G_0 \\ G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \end{aligned}$$

The distribution of the sequence of $(\pi_1, \pi_2, \cdots)$ is called the GEM distribution, named after Griffiths, Engen and McCloskey (Ishwaran and Zarepour, 2002). By making use of the GEM distribution the DPM model in (8) can be expressed equivalently as

$$\begin{aligned} \pi|\alpha &\sim GEM(\alpha) \\ z_i|\pi &\sim Cat(\pi) \\ \theta_k^*|G_0 &\sim G_0 \\ d_i|z_i, \theta_{z_i}^* &\sim F(\theta_{z_i}^*) \end{aligned} \tag{14}$$

with $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$ and $\theta_i = \theta_{z_i}^*$. This representation makes way for an intuitive interpretation of the manner in which a DP operates. The DP is a distribution that is constructed from a countable infinite number of atoms by combining two steps. Firstly, the random weights of the distribution π_k are obtained from a GEM distribution. The result is a discrete distribution that puts probability mass π_k on location k for $k \in \mathcal{N}$; this determines the scaling of the DP. Secondly, the draws from base distribution G_0, which may be a continuous distribution, determines the location of each π_k; it transforms the discrete distribution draw from the GEM distribution which assigns locations $1, 2, ...$ to a distribution with locations $\theta_1^*, \theta_2^*, ...$ (Broderick, 2015). Due to this, as compared to the finite mixture model representation, the DP combines the clustering probabilities, cluster assignment, and the cluster parameters into one random variable.

5.4 Gibbs Sampler

Many Gibbs sampling schemes for the model in (8) have been determined, with the most obvious being the sampler from the complete conditional posterior $\theta_j|\theta_{-j}, D$, i.e. the parameter of each observation is sampled one by one. As draws from a Dirichlet Process (DP) have a clustering property, one can define a more efficient algorithm by exploiting this feature. To do this, the stick-breaking representation of the DPM model is used.

$$\begin{aligned} \pi|\alpha &\sim GEM(\alpha) \\ z_i|\pi &\sim Cat(\pi) \\ \theta_k^*|\lambda &\sim G_0(\lambda) \\ d_i|z_i, \theta_{z_i}^* &\sim F(\theta_{z_i}^*) \end{aligned} \tag{15}$$

The θ_i parameters in (5) carry the same information as the cluster indicators z_i combined with the relevant cluster parameter θ_k^* (15). Hence, by writing the model in stick-breaking representation one can obtain a sample from the joint distribution in two steps. First, the z_i are sampled for all observations. Second, using this clustering scheme, θ_k^* are sampled for all clusters.

As π is difficult to sample due to its infinite size, it is integrated out. The respective conditional distributions of this sampler are derived as follows. First the distribution of z_i conditional on all except π is derived. Let λ represent all hyperparameters, D the data points d_i for $i = 1, \cdots, n$ and j a cluster observed among z_{-i}, then[13]

$$\begin{aligned} p(z_i = j|z_{-i}, D, \{\theta_k^*\}_{k=1}^K, \alpha, \lambda) &\propto p(z_i = j|z_{-i}, \alpha) p(D|z_{-i}, z_i = j, \{\theta_k^*\}_{k=1}^K, \alpha, \lambda) \\ &= p(z_i = j|z_{-i}, \alpha) \prod_{l=1}^{n} p(d_l|z_{-i}, z_i = j, \{\theta_k^*\}_{k=1}^K, \alpha, \lambda) \\ &= \frac{n_{j,-i}}{n + \alpha - 1} p(d_i|\theta_j^*) \end{aligned} \tag{16}$$

Similarly, the probability that z_i starts a new cluster is

$$\begin{aligned} & p(z_i = K + 1|z_{-i}, D, \{\theta_k^*\}_{k=1}^K, \alpha, \lambda) \\ &= p(z_i = K + 1|z_{-i}, \alpha) \prod_{l=1}^{n} p(d_l|z_{-i}, z_i = K + 1, \{\theta_k^*\}_{k=1}^K, \alpha, \lambda) \\ &= \frac{\alpha}{n + \alpha - 1} p(d_i|\lambda) \\ &= \frac{\alpha}{n + \alpha - 1} \int p(d_i|\theta) p(\theta|\lambda) d\theta \end{aligned} \tag{17}$$

The cluster parameters θ_k^* for all K existing clusters can be sampled from

$$\begin{aligned} p(\theta_k^*|\theta_{-k}^*, z, D, \alpha, \lambda) &\propto p(\theta_k^*|\lambda) p(D|z, \{\theta_k^*\}_{k=1}^K, \alpha, \lambda) \\ &= p(\theta_k^*|\lambda) \prod_{i=1}^{N} p(d_i|\theta_{z_i}^*) \\ &\propto p(\theta_k^*|\lambda) \prod_{z_i=k} p(d_i|\theta_{z_i}^*) \end{aligned} \tag{18}$$

[13]The probabilities $p(z_i = j|z_{-i}, \alpha)$ and $p(z_i = K + 1|z_{-i}, \alpha)$ are derived in section 5.2 and do not depend on cluster parameters.

The integral in $\int p(d_i|\theta)p(\theta|\lambda)d\theta$ is hard to evaluate in cases where the prior $p(\theta|\lambda)$ and the likelihood are not conjugate. Neal (2000) proposes to use auxiliary variables to circumvent having to calculate this integral. His algorithm can be summarized in the following way.

Neal's Algorithm 8 Let $(z_1, \cdots, z_n)$ and $\theta^* = (\theta^*_k : k \in \{z_1, \cdots, z_n\})$ describe the state of the Markov chain. The algorithm alters the previously derived Gibbs sampler by creating temporary variables and augmenting θ^* with m additional parameters drawn from G_0.

- For $i = 1, \cdots, n$, let K^- be the number of clusters after z_i is removed from the state. Relabel z_{-i} (if necessary) such that they take value in $1, \cdots, K^-$.
  - If $z_i = z_j$ for some $i \neq j$ draw new values for θ^*_h for $K^- < h < k^- + m$ from G_0.
  - If $z_i \neq z_j$ for all $i \neq j$, set $\theta^*_{K^-+1} = \theta^*_{z_i}$ and draw new values for θ^*_h for $K^-+1 < h < k^-+m$ from G_0.

  Then, draw a new value for z_i from

$$p(z_i = k|z_{-i}, d_i, \theta^*_1, \cdots, \theta^*_{K^-+m}) \propto \begin{cases} \frac{n_{-i,k}}{\alpha+n-1} p(d_i|\theta^*_k) & \text{for } 1 \leq k \leq K^-, \\ \frac{\alpha/m}{\alpha+n-1} p(d_i|\theta^*_k) & \text{for } K^- < k \leq K^- + m, \end{cases} \tag{19}$$

  where $n_{-i,k}$ is the number of $z_j = k$ for all $j \neq i$. Conclude this step by updating the state by altering z to contain z_i and θ^* to only contain cluster parameters that are associated with z_i for all $i = 1, \cdots, n$.
- For all $k \in \{z_1, \cdots, z_n\}$ draw a new value for θ^*_k from $\theta^*_k|\{d_i : z_i = k\}$
- A final step can potentially contain updates of hyperparameters.

6 Adapted Model Specification

Thus far, a very general specification of the DPM model is given. In this section, the general DPM model is modified to make it fit for the data, keeping in mind the aims of this analysis. In the next subsection, the model is adapted to facilitate the inference of the conditional distribution. After this adaptation, the likelihood and prior specifications are discussed.

6.1 Generative vs. Discriminative Approach

So far, the DPM model is defined for a general variable vector d. In terms of the individual-level election data, data point d denotes the variable bundle (x, y) where x are the sociodemographic variables, while y denotes the response variable. There are two main ways of adapting the DPM model to facilitate conditional distribution inference: the generative and the discriminative approach. Using the generative approach the joint distribution of (x, y) is modeled and $p(y|x)$ is inferred implicitly. On the other hand, the discriminative approach models the distribution of $p(y|x)$ directly.

At first glance, the discriminative approach seems the simpler alternative. There are plenty of models that aim to describe the relationship between a multinomial response variable and covariates that can easily be used here. Another benefit is the fact that the dimension of the problem is much smaller

when the response variable only is explicitly modeled. This implies that the discriminative approach would be computationally less expensive compared to the joint approach. While this argument holds for many other models, these observations are not necessarily true for the non-parametric approach taken in this analysis. To comply with the non-parametric feature of the DPM model, one must find a model specification that ensures that the covariates enter the distribution of the response variable in a flexible way. For this method to be truly non-parametric, the covariates should be at least able to influence the mixing probabilities (Rossi, 2014). An example of this is the approach taken by Geweke and Keane (2007). They aim to do non-parametric analysis with a finite mixture model where the following discriminative approach is taken

$$y|x \sim \sum_{k=1}^{K} \pi_k(x) f(x^T \beta_k, \sigma_k)$$

While this type of specification would indeed allow covariates to influence mixing probabilities, it does not facilitate a general form of heterogeneity. For example, heteroskedasticity is only accounted for implicitly through the influence of the covariates on the mixing probabilities. This inflexibility in the model has two consequences. First, the number of components needed to model the data is likely to increase as the sampler will try to accommodate a good fit. Furthermore, despite the increased number of components, the model may still perform poorly as it is build on the assumption of homoskedasticity (Villani et al., 2009). The disadvantages of this approach do not limit themselves to model performance. Due to the particular model specification, (conditional) conjugacy in the model may no longer hold, which complicates the sampling procedure (Rossi, 2014).

Clearly, the discriminative approach is not an elegant solution for a non-parametric model. The joint approach, on the other hand, inherently takes into account the distribution of the covariates and its effect on the model and does not require for these effects to be modeled explicitly. Due to this, it allows for all types of conditional heterogeneity (Rossi, 2014). Most DPM regression models make use of the generative approach (see for instance (Taddy and Kottas, 2010; Hannah et al., 2011; Shahbaba and Neal, 2009)).

In principle, using the generative approach, the model specification does not need to accommodate the inference of the conditional distribution. The DPM model can be specified in the following way

$$\begin{aligned} x_i|\theta_{x,i} &\sim F_y(\theta_{x,i}) \\ y_i|\theta_{y,i} &\sim F_x(\theta_{y,i}) \\ \theta_i = (\theta_{x,i}, \theta_{y,i})|G &\sim G \\ G|\alpha, G_0 &\sim DP(\alpha, G_0) \end{aligned} \tag{20}$$

where the response and the covariates are assumed to be independent. This model is not popular in the literature due to its poor performance in predicting the response variable. This result can be attributed to the weight of the response variable in the posterior. As the number of covariates grow, the weight of the response variable shrinks quickly and makes it less important for prediction. For this reason, this method is not considered here.[14].

[14]When explicitly modeling the conditional distribution, an increase in the number of covariates is accompanied by an increase in the number of parameters related to the response variable. This makes this approach more resistant to dimensionality compared to the model presented above.

Due to above mentioned reasons, the generative approach is taken. The DPM model is adapted as follows

$$\begin{aligned} y_i|x_i,\theta_{y,i} &\sim F_y(x_i^T\theta_{y,i}) \\ x_i|\theta_{x,i} &\sim F_x(\theta_{x,i}) \\ \theta_i = (\theta_{x,i},\theta_{y,i})|G &\sim G \\ G|\alpha,G_0 &\sim DP(\alpha,G_0) \end{aligned} \tag{21}$$

where the joint distribution of (x_i, y_i) is modeled using the the decomposition $p(x_i, y_i|\theta_i) = p(y_i|x_i, \theta_i^y)p(x_i|\theta_i^x)$.

6.2 Likelihood Specification

The next step towards completing the model specification is determining the likelihood distributions. The response variable y_i in the voting choice model concerns categorical data with J alternatives labeled $1, \cdots, J$. These voter choices are predicted using the covariate vector x_i. The covariates are either continuous or categorical. For notation purposes, let $x_i = (x_{i,1}^{\text{con}}, \cdots, x_{i,L^{\text{con}}}^{\text{con}}, x_{i,1}^{\text{cat}}, \cdots, x_{i,L^{\text{cat}}}^{\text{cat}})^T$. Using the adapted DPM model, for all $i = 1, \cdots, n$ it holds that

$$\begin{aligned} y_i|x_i,\theta_i^y &\sim Cat(p_i(x_i)) \\ x_i|\theta_i^x &\sim F_x(\theta_i^x) \\ \theta_i|G = (\theta_i^x,\theta_i^y)|G &\sim G \\ G &\sim DP(\alpha,G_0) \end{aligned} \tag{22}$$

The marginal distributions of the covariates, which together form F_x, are assigned in accordance with the nature of the data.

$$\begin{aligned} x_i^{\text{con}}|\mu,\Sigma &\sim N(\mu,\Sigma) \\ x_{i,l}^{\text{cat}}|\pi_l &\sim Cat(\pi_l) \quad \text{for l} = 1, \cdots, L^{\text{cat}} \end{aligned} \tag{23}$$

where $x_i^{\text{con}} = (x_{i,1}^{\text{con}}, \cdots, x_{i,L^{\text{con}}}^{\text{con}})^T$ and $x_i^{\text{cat}} = (x_{i,1}^{\text{cat}}, \cdots, x_{i,L^{\text{con}}}^{\text{cat}})^T$. Here, the continuous variables x_i^{con} and categorical variables x_i^{cat} are assumed to be independent from each other. Continuous variables $x_{i,l}^{con}$ can be correlated among each other due to the multivariate normal likelihood specification. Each categorical variable $x_{i,l}^{cat}$, however, is independent from all other categorical variables. Although variable Age is strictly speaking count data and therefore not continuous, it is categorized as continuous data to not complicate the model any further.

Popular choices for the conditional relationship between y_i and x_i for a multinomial response variable are the probit and the multinomial logit (MNL) specification. Using the probit specification in a DPM model makes the sampling procedure more tricky as the sampling algorithm requires the likelihood to be evaluated. Numerical approximation of the likelihood could be a solution, but is not desirable

as it makes an already computationally expensive sampling procedure more lengthy. Therefore, the MNL specification is chosen to model the conditional relationship between the response variable and covariates. Hence, the probabilities for each party are defined as follows. Let $\tilde{x_i}$ be the regressor vector, then

$$P(y_i = j|x_i, \beta_i) = \frac{\exp(\tilde{x_i}^T \beta_i^j)}{\sum_{l=1}^{J} \exp(\tilde{x_i}^T \beta_i^l)} \tag{24}$$

which is the j^{th} element of $p_i(x_i))$ in (22). Due to the coefficients being alternative specific, parameter β_i captures the collection of all alternative specific coefficients $\{\beta_i^j\}_{j=1}^{J-1}$ with $\beta_i^j = (\beta_{i,0}^j, \cdots, \beta_{i,L^*}^j)'$. Without loss of generality, β^J is set to zero to keep the model identified under addition[15].

6.3 Prior Specification

In the adapted DPM model described in (22), θ_i represents all parameters, where $\theta_i^y = \{\beta_i^j\}_{j=1}^J$ and $\theta_i^x = (\mu_i, \Sigma_i, \{\pi_{i,l}\}_{l=1}^{L^{Cat}})$. These parameters are DP distributed with base measure G_0 and concentration parameter α. Specifying a prior in this model corresponds to choosing the base distribution and concentration parameter.

On the one hand, the choice of prior distribution in a DPM model seems of less importance compared to regular Bayesian analysis due the DP prior. Particularly since the strength parameter α can be chosen such that the prior does not influence the posterior much (low values of alpha). While this is true to some extent, there are some important considerations regarding G_0.

Priors are chosen to impart existing knowledge on the model. When there is little existing knowledge, the prior is specified such that it has a low impact relative to data. This is typically realized by using diffuse prior settings. In this research, imparting little knowledge is especially relevant as the DPM model is chosen for its non-parametric properties. Although this seems to suggest that a diffuse prior is preferred, a diffuse prior can have undesired effects on the outcome. Rossi (2014) discusses the role of hyperparameters λ on the posterior distribution. The conditional posterior distribution of θ can be written as

$$p(\theta_i|\theta_{-i}, d_i) \propto q_0 \alpha G_i(\theta_i|\lambda) + \sum_{j \neq i} q_j \delta(\theta_j)$$

where $q_0 = \int p(d_i|\theta_i)p(\theta_i|\lambda)d\theta_i$ and $q_j = p(d_i|\theta_j)$. $G_i(\lambda)$ is the posterior for θ_i under prior G_0 updated with observation d_i. This conditional distribution can be interpreted as drawing θ_i from N possible 'models'. θ_i is either equal to one of the $N-1$ existing values in θ_{-i} or a new value drawn from $G_i(\lambda)$. The weights q_0 and q_j are the marginal probabilities under each of the N models and determine the likelihood with which a model is chosen. The influence of λ on the posterior becomes clear from q_0. The marginal likelihood under 'model 0', q_0, can be made smaller by choosing a more diffuse specification for G_0. This decreases the probability that a new cluster is opened. Hence, a very diffuse base distribution results in a posterior with a small number of components. Even though very diffuse

[15] $L^* = 1 + L^{con} + \sum_{l=1}^{L^{cat}} (M_l - 1)$, with M_l the number of categories of covariate x_l^{cat}.

priors (such as flat priors) are common to use in Bayesian analysis to express little prior beliefs about parameters, they are not desirable in the DPM model due to the aforementioned result.

Instead, Rossi (2014) suggests using a conjugate Normal-Inverse Wishart prior specification[16]

$$\begin{aligned} \mu|\Sigma &\sim N(\mu_0, c^{-1}\Sigma) \\ \Sigma &\sim IW(\nu, \nu\upsilon I) \end{aligned} \tag{25}$$

with proper flat priors on the hyperparameters. The Normal-Inverse Wishart (NIW) specification is common in Bayesian and DPM model literature as it is easy to sample from. It also sidesteps the previously mentioned issue as the diffuseness of the prior specification can be governed by the researcher. A (flat) prior on the hyperparameters λ allows the researcher to express uncertainty about the prior specification and the data to influence the values of the hyperparameters. However, the downside of this specification is that the covariance structure of μ is heavily restricted. On top of this, this prior specification gives a lot of weight to small clusters with small variance in determining μ (West et al., 1994). Görür and Rasmussen (2010) put the NIW specification in a DPM model to a test by comparing its performance to the less restrictive conditionally conjugate model. The latter prior specification is similar to (25) with the only difference being that the distribution of μ does not depend on Σ. They find that conjugate NIW specification typically uses more components to model the same data. More importantly, they find the predictive performance of the conditionally conjugate model to be better, where the difference in predictive performance grows with dimensionality.

The choice of base distribution and distributions for the corresponding hyperparameters is mainly driven by these two arguments. The base measure is defined as follows

$$\begin{aligned} \mu|\mu_\mu, \Sigma_\mu &\sim N(\mu_\mu, \Sigma_\mu) \\ \Sigma|\nu_\Sigma, \upsilon_\Sigma &\sim IW(\nu_\Sigma, \nu_\Sigma \upsilon_\Sigma I_{L^{con}}) \\ \pi_l|a_{0,l} &\sim Dir(\frac{a_{0,l}}{M_l}, \cdots, \frac{a_{0,l}}{M_l}) \quad \text{for l} = 1, \cdots, L^{\text{cat}} \\ \beta|\bar{\beta}, B &\sim N(\bar{\beta}, B) \end{aligned} \tag{26}$$

where Dir and IW represent the Dirichlet and Inverse-Wishart distribution respectively. Here, the categorical covariates are given a Dirichlet prior, due to the fact that it has the appropriate support for parameter π_l an its conjugacy with the categorical distribution. The symmetric parametrization for its hyperparameter $a_{0,l}$ is primarily chosen as it reflects the a priori believe that the relative likelihood of each category occurring is not known without looking at the data. Additionally, it simplifies its interpretation. The prior distributions of the hyperparameters are

[16] Rossi (2014) argues this specific parametrization of the Wishart distribution is chosen as an attempt to isolate the tightness and location of the distribution. In principle, the distribution can also be parametrized as $IW(\nu, \tilde{\upsilon} I)$ or, even more general, $IW(\nu, \Psi)$ to maintain conjugacy and achieve the same results. The difference here is only in interpretation. In fact, even though this parametrization is less common in Bayesian literature, it makes the interpretation of the parameters more intuitive. In this case, the interpretation of the tightness parameter relates to the amount of information captured in the prior, where ν is the number of observations. υ can be interpreted as the variance in the sample of ν observations (Gelman et al., 2014; Nydick, 2012).

$$
\begin{aligned}
\alpha|a,b &\sim Ga(a,b) \\
\mu_{\mu,l}|\mu_{\mu,l}^{lo},\mu_{\mu,l}^{up} &\sim unif(\mu_{\mu,l}^{lo},\mu_{\mu,l}^{up}) \quad \forall l=1,\cdots,L^{Con} \\
\Sigma_\mu|\nu_\mu,\upsilon_\mu &\sim IW(\nu_\mu,\nu_\mu\upsilon_\mu I_{L^{Con}}) \\
\upsilon_\mu|\upsilon_\mu^{lo},\upsilon_\mu^{up} &\sim unif(\upsilon_{\mu,l},\upsilon_\mu^{up}) \\
\upsilon_\Sigma|\upsilon_\Sigma^{lo},\upsilon_\Sigma^{up} &\sim unif(\upsilon_\Sigma^{lo},\upsilon_\Sigma^{up}) \\
a_{0,l}|a_{0,l}^{lo},a_{0,l}^{up} &\sim unif(a_{0,l}^{lo},a_{0,l}^{up}) \quad \forall l=1,\cdots,L^{Cat} \\
B|\nu_\beta,\upsilon_\beta &\sim IW(\nu_\beta,\nu_\beta\upsilon_\beta I_{L^*}) \\
\bar{\beta}_l|\bar{\beta}_l^{lo},\bar{\beta}_l^{up} &\sim unif(\bar{\beta}_l^{lo},\bar{\beta}_l^{up}) \quad \forall l=1,\cdots,L^* \\
\upsilon_\beta|\upsilon_\beta^{lo},\upsilon_\beta^{up} &\sim unif(\upsilon_\beta^{lo},\upsilon_\beta^{up})
\end{aligned} \tag{27}
$$

where *uni* and Ga represent the uniform and gamma distribution respectively. All hyperparameter distributions except the distribution of α are based on the suggestion of Rossi (2014). Adopting a gamma distribution for α is common practice in DPM models. The approach is first introduced by Escobar and West (1995) and makes use of auxiliary variable Gibbs sampling to simplify sampling from its posterior distribution.

7 Hybrid MCMC Sampler

The Gibbs sampler scheme in section 5.4 can be applied to the voter choice model by writing the model as

$$
\begin{aligned}
x_i,y_i|\theta_i &\sim F(\theta_i) \\
\theta_i|G &\sim G \\
G|\alpha,G_0 &\sim DP(\alpha,G_0)
\end{aligned} \tag{28}
$$

and setting d_i in (5) equal to x_i, y_i in (28). Step 1 of Neal's Algorithm 8 boils down to computing the joint likelihood given the data point specific parameter $\theta_i = (\beta_i, \mu_i, \Sigma_i, \{\pi_{i,l}\}_{l=1}^{L^{cat}})$. Using the model specification in (23) and (24), the joint likelihood for individual i is

$$
\begin{aligned}
p(x_i,y_i|\theta_i) &= p(y_i|x_i,\beta_i)p(x_i^{\text{con}}|\mu_i,\Sigma_i)p(x_i^{\text{cat}}|\{\pi_{i,l}\}_{l=1}^{L^{cat}}) \\
&= \prod_{j=1}^{J} p_{j,i}(x_i,\beta_i^j)^{I(y_i=j)} N(x_i^{con}|\mu_i,\Sigma_i) \prod_{l=1}^{L^{cat}} \prod_{m=1}^{M_l} \pi_{i,l,m}^{I(x_{i,l}^{cat}=m)}
\end{aligned} \tag{29}
$$

M_l the number of categories in x_l^{cat}. Step 2 of the algorithm requires the computation of the posterior distributions of $\theta_k^* = (\beta_k^*, \mu_k^*, \Sigma_k^*, \{\pi_{l,k}^*\}_{l=1}^{L^{cat}})$. From the derivation in (17) we know that

$$
p(\theta_k^*|z,y,x,\lambda) \propto p(\theta_k^*|\lambda) \prod_{z_i=k} p(x_i,y_i|\theta_{z_i}^*) \tag{30}
$$

where

$$p(\theta_k^*|\lambda) \propto N(\beta_k^*|\bar{\beta}, B)N(\mu_k^*|\mu_\mu, \Sigma_\mu)IW(\Sigma_k^*|\nu_\Sigma, \upsilon_\Sigma) \prod_{l=1}^{L^{cat}} Dir(\pi_l^*|\frac{a_{0,l}}{M_l}, \cdots, \frac{a_{0,l}}{M_l}) \tag{31}$$

due to the specification in (26). The conditional posterior distribution for each parameter are derived from this expression. The next subsections provide an overview of these conditional posteriors, while the derivation can be found in the appendix.

7.1 Sampling θ_i^x

From the latter expressions we can obtain the conditional posterior distributions of the distinct parameters μ_k^*, Σ_k^* and $\pi_{l,k}^*$ by dropping the irrelevant terms. The conditional posteriors are

$$\begin{aligned}
\mu_k^*|\mu_\mu, \Sigma_\mu, \{x_k^{con}\}_{z_i=k} &\sim N((\Sigma_\mu + n_k\Sigma_k^*)^{-1}(\Sigma_\mu\mu_\mu + n_k\Sigma_k^*\bar{x}^{con}), (\Sigma_\mu + n_k\Sigma_k^*)^{-1}) \\
\Sigma_k^*|\nu_\Sigma, \upsilon_\Sigma, \mu_k^*, \{x_k^{con}\}_{z_i=k} &\sim IW(n_k + \nu_\Sigma, \nu_\Sigma\upsilon_\Sigma I_{L^{con}} + S_x) \\
\pi_{l,k}^*|a_{0,l}, M_l, \{x_k^{cat}\}_{z_i=k} &\sim Dir(\frac{a_{0,l}}{M_l} + \sum_{z_i=k} I(x_{i,l}=1), \cdots, \frac{a_{0,l}}{M_l} + \sum_{z_i=k} I(x_{i,l}=M_l))
\end{aligned}$$

where $S_x = \sum_{z_i=k}(x_i^{con} - \mu_k^*)(x_i^{con} - \mu_k^*)^T)$. The complete derivation can be found in A.1.

7.2 Sampling θ_i^y

Expressions (29) to (31) imply that the conditional posterior of β_k only depends on its prior and on the likelihood of all $y_i|x_i$ in the same cluster, where

$$p(\beta_k^*|z, x, y, \lambda_\beta) \propto N(\beta_k^*|\lambda_\beta) \prod_{z_i=k} \prod_{j=1}^{J} p_{j,i}(x_i, \beta_k^{*j})^{I(y_i=j)}$$

with $\lambda_\beta = (\bar{\beta}, B)$. This means that, when sampling β_k^*, we deal with an MNL model with nonrandom covariates. As the prior-likelihood pair is not conjugate, these parameters cannot be sampled with a Gibbs step. Instead, a Metropolis-Hastings (MH) sampler can be used to obtain posterior samples from the conditional posterior distribution of β_k^*. In such a hybrid MCMC (or sometimes MH-within-Gibbs) sampling scheme, the proposal and target distributions may depend on other parameters in the model. Due to the hierarchical Bayesian approach, this step depends on parameters $(z, \bar{\beta}, B)$. To properly fuse the remainder of the Gibbs sampler scheme with the MH-step, the values of these parameters are updated with their most recent draw to achieve convergence to the correct stationary distribution (Rossi et al., 2005).

There are numerous samplers that can be used to sample this part of the model. Each of these have their own advantages and disadvantages. The next bit of writing aims at giving an overview of these samplers and discuss the extent to which they are suitable for the DPM voter choice model.

7.2.1 Samplers for the multinomial logit model

The likelihood of the data under an MNL specification usually has a good asymptotic normal approximation. Commonly used MH algorithms make use of this feature. Examples of these are the independence sampler with multivariate-t proposal distribution and the random walk sampler with Normal proposal distribution, both presented in Rossi et al. (2005). Other, more recent examples using this principle are data augmented samplers such as (Holmes et al., 2006; Scott, 2011; Frühwirth-Schnatter and Frühwirth, 2010, 2012). Following the ideas presented in Albert and Chib (1993) and McCulloch and Rossi (1994), the latter samplers use the random utility model (RUM) and take a normal approximation to $\epsilon_{j,i}$. In the RUM representation, $\epsilon_{j,i}$ is defined in the following way[17]

$$
\begin{aligned}
&y^u_{j,i} = x_i^T \beta_j + \epsilon_{j,i}{}^{18} \quad j = 1, ..., J-1 \\
&y^u_{J,i} = \epsilon_{J,i} \\
&y_i = j \iff y^u_{j,i} = \max(y^u_{1,i}, \cdots, y^u_{J,i})
\end{aligned}
$$

The aforementioned samplers, and more, are all considered good candidates. A good starting point for picking one out of the many candidates is the comparative study on various MH samplers for the multinomial logit model carried out by Frühwirth-Schnatter and Frühwirth (2010). This study includes the previously mentioned samplers and more (with and without data augmentation). Using five well-known datasets with different characteristics, the samplers are compared based on their performance in total CPU time, acceptance rate and (in)efficiency based on the empirical correlation in the MCMC draws. The study finds that the data augmented samplers perform best overall. However, the performance of each sampler is specific to each data set. Therefore, several samplers are selected and compared in their performance based on the data set at hand.

Frühwirth-Schnatter and Frühwirth find that, among the MH samplers without data augmentation, the independence sampler clearly outperforms all others in total CPU time and efficiency. Compared to the data augmented samplers, the independence sampler is less efficient in most cases, yet again faster in total CPU time. Due its speed, the independence sampler seems an attractive choice for the voter choice model. When using this sampler, the posterior is approximated as

$$
p(\beta|y, X) \propto |H|^{\frac{1}{2}} \exp\{-\frac{1}{2}(\beta - \hat{\beta})^T H (\beta - \hat{\beta}))\}
$$

Rossi et al. (2005) discuss many possibilities for $\hat{\beta}$ and H and eventually choose the MLE for $\hat{\beta}$ and the expected Hessian of the negative log-likelihood evaluated at $\hat{\beta}$ for H. This is not feasible in the DPM model, as the number of data points assigned to a given cluster can be lower than the dimension of the β parameter. The distribution can, however, be centered on the mode of the posterior distribution

[17] The superscript * and subscript k in β^*_k are dropped in the remainder of the section to simplify notation.
[18] Taking $\epsilon_{j,i} \sim EV$ in the RUM model leads to the MNL specification in (24).

and H can be defined as the expected Hessian of the negative log-posterior. The mode of the posterior distribution can be obtained by numerically optimizing the log-posterior distribution.

The advantage the independence sampler has to offer is the 'tuning' parameter ν. This parameter can be adjusted based on the acceptance rate. In a MH sampler, this is especially important as it is desirable that the proposal distribution has fatter tails than the target distribution. When this condition is not met, the chain can move to the tails and repeat values to build up mass due to the high relative mass at the tails. Small values of ν imply fat tails, which can avoid a high rejection rate. On the other hand, too small values cause the peak of proposal distribution to be very 'slim', which causes the sampler to build mass in the 'shoulders' of the Normal distribution (Rossi et al., 2005).

Frühwirth-Schnatter and Frühwirth recommend their own data augmented sampler based the sampler proposed by Scott (2011) as first choice due to the combination of its simplicity, relative efficiency and speed. The proposal distribution in this MH sampler is the posterior distribution implied by the differenced RUM specification

$$z_{j,i} = x_i^T \beta_j + \varepsilon_{j,i} \qquad j = 1, ..., J-1$$

$$y_i = \begin{cases} j \neq J, & \text{if } z_{j,i} = \max(z_{1,i}, \cdots, z_{J-1,i}) > 0 \\ J, & \text{if } \max(z_{1,i}, \cdots, z_{J-1,i}) < 0 \end{cases}$$

where $z_{j,i} = y_{j,i}^u - y_{J,i}^u$, $\varepsilon_{j,i} = \epsilon_{j,i} - \epsilon_{J,i}$ and $\varepsilon_{j,i}$ has a multivariate logistic distribution. Following Scott (2011), a normal approximation for the error term is used where $\varepsilon_{j,i} \sim N(0, R)$[19]. This sampler outperforms the original sampler in terms of acceptance rate as the logistic distribution is closer to the normal distribution than the extreme value distribution is. Using this augmented specification, the auxiliary variables $z_{j,i}$ and parameters are sampled sequentially. The $z_{j,i}$ are derived using the formula

$$y_{j,i}^u = -\log(-\frac{\log(U_i)}{\sum_{l=1}^J \exp(x_i^T \beta_l)} - \frac{\log(V_{j,i})}{\exp(x_i^T \beta_j)} I(y_i \neq j))$$

where U_i and $V_{j,i}$ are uniform random variables in [0,1]. The $\beta|z$ parameters under the dRUM specification do not have a closed form conditional posterior. Based on Scott's algorithm, the conditional posterior under the normal approximation is used as a proposal distribution for the MH algorithm. Under the normal approximation, $\beta|z$ is sampled from a multivariate regression model with equicorrelated errors

$$L_R^T z_i = L_R^T X_i \beta + L_R^T \varepsilon_i$$

where X_i is the regressor matrix for individual i. To circumvent the issue of correlated errors, the left- and right-hand side of the stacked regression equation with L_R^T which is determined using the

[19] $R = \frac{\pi^2}{6}(I + \iota\iota^T)$ where ι is a vector of ones.

Cholesky decomposition of $R = L_R L_R^T$. The posterior distribution of the transformed model has the well-known closed form $\beta|z \sim N(b_N, B_N)$ where

$$B_N = (B^{-1} + \sum_{i=1}^{n} X_i^T R^{-1} X_i)^{-1}$$
$$b_N = B_N(B^{-1}\bar{\beta} + \sum_{i=1}^{n} X_i^T R^{-1} z_i)$$

Both of these MH algorithms can be useful in sampling the β coefficients of the DPM model due to the aforementioned reasons and are probably the best pick for many datasets/models. When used on the voter choice data, however, several issues present themselves. The data augmented sampler by Frühwirth and Frühwirth-Schnatter remains stuck in the initial conditions and has a acceptance rate of practically zero. In fact, when applied to the full data set without assuming a DPM model, the acceptance rate is exactly zero. This result can most likely be attributed to the size of the model. In the voter choice model, the response variable can take seven different values, the number of regressors is 29 and the number of data points is larger than 2000. The data sets used in Frühwirth-Schnatter and Frühwirth (2010) are much smaller in terms of all three of these factors. Although each data set is different, there seems to be a general tendency where complexity of each model in terms of the previously mentioned factors correlate with lower acceptance rates and efficiency.

The independence sampler by Rossi does relatively better on this front with an additional benefit of the possibility to tune the proposal distribution for better results. The numerical optimization that is required for the sampler, however, proves to be a big problem. Due to the hybrid MCMC sampling scheme, this optimization must be performed every iteration to account for the changes in $(z, \bar{\beta}, B)$. Taking into consideration the dimensionality of the data set, using the independence sampler turns out to be not feasible.

Based on these results, other samplers have been selected to test on the data at hand. The results of the independence sampler suggest using a t-distribution may be more suitable. Due to this, a doubly data augmented sampler taken from Frühwirth-Schnatter and Frühwirth (2012) is explored. This sampler uses a multivariate t-distribution to approximate the multivariate logistic distribution in the differenced RUM specification. Exploiting the fact that the t-distribution is a scale mixture of normal distribution, this sampler uses a similar strategy to the sampling scheme of Frühwirth-Schnatter and Frühwirth (2010) explained above. Unfortunately, its performance is no different than the original data augmented sampler. For the sake of completeness, the data augmented sampler proposed by Scott (2011) is tested as well, with the exact same results.

Due to the bad performance of the preferred samplers, the random walk sampler is used to sample from the posterior of the voter choice model. The proposal distribution in this sampling scheme is

$$\beta_{t+1}|\beta_t \sim N(\beta_t, s^2 H^{-1})$$

where s is the scaling factor and H represents the hessian. Although this method is much more inefficient and relatively slow, it is the only option that is simple to implement and feasible. This way, the numerical optimization issue is circumvented. Furthermore, the acceptance rates are mostly about 20 - 25 %. These acceptance rates are realized under asymptotically optimal scaling where $s = 2.382/L^*$ (Roberts et al., 2001).

7.3 Sampling α

The Gibbs sampling scheme for concentration parameter α is taken from Escobar and West (1995). This is the most commonly used method to update the concentration parameter α during the step-by-step sampling procedure.

To sample α in a Gibbs step, Escobar and West (1995) assume α has a Gamma distribution. They augment the sampler scheme with an auxiliary variable that allows α to be sampled from analytically tractable conditional distribution. As shown by Antoniak (1974), the likelihood of the number of clusters K in a DPM model has a likelihood that only depends on α and the number of observations n.

$$p(k|\alpha, n) \propto \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}$$

Using this likelihood and the fact that $\Gamma(x) = \frac{\Gamma(x+1)}{x+1}$ we can write the posterior of α as follows

$$\begin{aligned} p(\alpha|k, n) &\propto p(\alpha)p(k|\alpha) \\ &= p(\alpha)\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \\ &= p(\alpha)\alpha^k \frac{\Gamma(\alpha + 1)}{\alpha} \frac{\alpha + n}{\Gamma(\alpha + n + 1)} \\ &= p(\alpha)\alpha^{k-1} \frac{\alpha + n}{\Gamma(n)} \frac{\Gamma(\alpha + 1)\Gamma(n)}{\Gamma(\alpha + n + 1)} \end{aligned}$$

Recognizing that the last factor is the reciprocal of the normalizing constant of a Beta($\alpha + 1$, n) distribution, allows for the following rewrite

$$p(\alpha|k, n) \propto p(\alpha)\alpha^{k-1} \frac{\alpha + n}{\Gamma(n)} \int_0^1 \eta^\alpha (1 - \eta)^{n-1} d\eta$$

This result implies

$$\begin{aligned} p(\alpha|k, n) &\propto \int_0^1 p(\alpha)\alpha^{k-1}(\alpha + n)\eta^\alpha (1 - \eta)^{n-1} d\eta \\ &\propto \int_0^1 p(\alpha, \eta|k, n) d\eta \end{aligned}$$

Hence, the posterior of α is the marginal of the joint distribution of α and some Beta distributed auxiliary variable η. This property is exploited in the Gibbs sampler scheme. Assuming a Gamma prior on α, i.e. $p(\alpha) \sim G(a, b)$, the following holds

$$
\begin{aligned}
p(\alpha|k,n) &\propto p(\alpha,\eta|k,n) \\
&\propto \frac{b^a}{\Gamma(a)}\alpha^{a-1}e^{-b\alpha}\alpha^{k-1}(\alpha+n)\eta^{\alpha} \\
&\propto \alpha^{a+k-2}e^{-b\alpha}(\alpha+n)e^{\alpha log(\eta)} \\
&\propto (\alpha+n)\alpha^{a+k-2}e^{\alpha(-b+log(\eta))} \\
&\propto \alpha^{a+k-1}e^{-\alpha(b-log(\eta))} + n\alpha^{a+k-2}e^{-\alpha(b-log(\eta))}
\end{aligned} \tag{32}
$$

The last expression is a mixture of two Gamma distributions with weights $c_\eta = \frac{a+k-1}{a+k-1+n(b-log(\eta))}$

$$
\alpha|k,n \sim c_\eta G(a+k, b-log(\eta)) + (1-c_\eta)nG(a+k-1, b-log(\eta))
$$

The auxiliary variable η has distribution Beta($\alpha+1$, n)

$$
\begin{aligned}
p(\eta|k,n) &\propto p(\alpha,\eta|k,n) \\
&\propto \alpha^{k-1}(\alpha+n)\eta^{\alpha}(1-\eta)^{n-1}
\end{aligned} \tag{33}
$$

In conclusion, the concentration parameter α can be updated in each Gibbs iteration by first sampling η from (33) and using this to sample α from the conditional distribution in (32).

7.4 Sampling the hyperparameters

To determine the posterior of hyperparameters λ, the conditional independence implied by the stick-breaking construction is used (Müller and Quintana, 2004).

$$
p(\lambda|\{\mu_k^*, \Sigma_k^*, \{\pi_{l,k}^*\}_{l=1}^{L^{cat}}, \beta_k^*\}_{k=1}^{K}) \propto p(\lambda)\prod_{k=1}^{K} p(\mu_k^*, \Sigma_k^*, \{\pi_{l,k}^*\}_{l=1}^{L^{cat}}, \beta_k^*|\lambda)
$$

As the priors on the hyperparameters are specified independently from each other, $p(\lambda)$ is simply their product. Together with

$$
p(\mu_k^*, \Sigma_k^*, \{\pi{*}_{l,k}\}_{l=1}^{L^{cat}}, \beta_k^*|\lambda) = p(\mu_k^*, \Sigma_k^*|\lambda_\mu, \lambda_\Sigma)\prod_{l=1}^{L^{cat}} p(\pi_{l,k}^*|\lambda_{\pi_l})p(\beta_k^*|\lambda_\beta)
$$

these expressions imply that the conditional posteriors of $(\lambda_{\mu_x}, \lambda_{\Sigma_x})$, λ_{π_l} for $l = 1, \cdots, L^{cat}$ and λ_β can be derived separately. This derivation can be found in sections A.2 - A.4 in the appendix.

8 Convergence and Forecasting

Once a sample from the joint posterior distribution is obtained, the simulated draws can be used to compute point estimates for posterior prediction. The second part of this section covers this topic. Before one can use the output of the simulation, however, the draws must be assessed for convergence. In the next section, the approach in assessing convergence is discussed.

8.1 Detecting Convergence

One of the challenges in using MCMC draws is determining whether the Markov chain in question has converged to the target distribution. As the course of the chain is affected by the initial starting point of the algorithm, if the simulation is not run long enough, the draws may not be representative of the stationary distribution. To decrease the influence of the initial starting point a number of initial draws is discarded. This burn-in sample should be large enough to shake off the influence of the initial condition (Gelman et al., 2014).

This problem is magnified due to the fact that MCMC draws exhibit serial correlation. Partially, this correlation stems from repeated values in the MH step. On top of that, cross-correlation between variables contributes to the serial correlation, as highly cross-correlated variables cause the steps taken in the parameter space to be small (Lunn et al., 2013). Consequently, the Markov chain needs more time to be able to navigate through the entire sample space. This complicates the task of assessing convergence as it decreases the speed with which the effects of the initial condition fade. Aside from the convergence issues, the inefficiency causes inference to be less accurate.

There are numerous ways in which one can approach the problem of detecting convergence. An informal, but popular way is visual convergence detection by means of traceplots. In this case, a line fitted through successive draws of the sampler is plotted against the iteration number. When the chain is converged, the plot should look like a random scatter around a stable mean (Lunn et al., 2013). Furthermore, there are various formal convergence diagnostics. The diagnostics proposed by Geweke (1992) and Gelman and Rubin (1992) stand out due to their simplicity and wide applicability. Especially the Gelman-Rubin diagnostic is popular in the MCMC literature. Below follows an outline of these methods.

Geweke's convergence diagnostic is essentially build on the idea behind traceplots. When a chain is converged, the sample mean of the draws (or any function thereof) computed using a part of the chain with a sufficient window length should be the same for any part of the chain. In particular, this holds for the sample mean computed using the beginning of the converged chain compared to the last part of the chain. Geweke's statistic formalizes this idea by performing a Z-test on the equality of these two sample means. As MCMC draws are correlated, the variance cannot be estimated by the sample variance. Therefore, the statistic estimates the variance from the spectral density at zero. However, convergence can only be detected in this way if the sampler has sufficiently navigated through sample space. This reveals the problem with this diagnostic (and, essentially, with traceplots as well). In some cases, convergence may be so slow that it appears as if the chain is converged (Lunn et al., 2013). Hence, Geweke's method does not address the issue of whether navigation is complete (Rossi et al., 2005).

A possible solution to this problem is to compare multiple chains started on different initial points. If these end up in the same stable mean, the chain should be converged (Lunn et al., 2013). Gelman and Rubin's statistic is based on this idea. Their statistic can be broken down into two steps. The

first step is to find an overdispersed estimate of the target distribution that is centered about the target's mode(s). From this estimate, several points are sampled that serve as starting points for the Markov chains. The second step is to make use of the information provided by the multiple chains. Suppose that the variable of interest has variance σ^2 under the target distribution. Then, Gelman and Rubin (1992) propose an overestimate ($\hat{V}$) for this variance using the between-chain and within-chain variance. $\hat{V}$ is then compared to the pooled within-chain variance. This statistic, the potential scale reduction factor (PSRF), is defined as [20]

$$\sqrt{\hat{R}} = \sqrt{\frac{df+3}{df+1}\frac{\hat{V}}{W}}$$

where df is the estimated degrees of freedom and W is the pooled within-chain variance. The rationale behind this statistic is as follows. When the chain is not converged, W underestimates σ^2. Furthermore, $\hat{V}$ has been defined such that it overestimates the variance under overdispersed starting points. Therefore, the PSRF will be large when the chains are not converged. When the chains approach convergence, the PSRF approaches 1. This way, the statistic gives an indication of how close the chain is to convergence (Gelman and Rubin, 1992; Cowles and Carlin, 1996). Brooks and Gelman (1998) propose a PSRF value of smaller than 1.2 as a general rule for approximate convergence.

Officially, all these methods should be applied to each scalar in the model. For models with many parameters, such as the model in this writing, this is highly impractical. Although these diagnostics have initially been proposed as a univariate statistics, they can easily be extended to multidimensional problems by applying them on a function of all parameters. A convenient choice is the log posterior evaluated at each draw. This way, one can assess whether the joint posterior distribution has converged. Using the conditional posterior $p(\theta_i|\theta_{-i}, d_i)$, the joint posterior of the DPM model can be determined with chainrule as follows

$$p(\theta_1, \theta_2, \cdots, \theta_n) = p(\theta_1)\prod_{i=2}^{n} p(\theta_i|\theta_1, \cdots, \theta_j, d_i : \forall j < i)$$

For nonconjugate models, this expression is hard to evaluate due to analytically intractable integrals. In section 5.4 the stick-breaking construction is used to split this posterior into two parts to form a Gibbs sampler. Due to this, the second part of this sampler, where the unique cluster parameters are sampled, simplifies to

$$\prod_{k=1}^{K} p(\theta_k^*|z, D, \alpha, \lambda) \propto \prod_{k=1}^{K} p(\theta_k^*|\lambda) \prod_{z_i=k} p(d_i|\theta_{z_i}^*) \tag{34}$$

This is a simple expression to evaluate. Moreover, it is a relatively good candidate for monitoring convergence (at least partially), since it is a function of most variables of interest. It only lacks in

[20] Original definition of the statistic in (Gelman and Rubin, 1992) is different from the formula as presented in this section. This formula led to several issues as it was defined incorrectly. Brooks and Gelman (1998) address this issue and conclude that the correction factor in the formula should be adjusted to $\frac{df+3}{df+1}$.

directly accounting for the partition of the data in clusters. This partition is determined by α, the likelihood of the data and the prior distribution. Expression (34) contains the latter two components. Hence, although not ideal, (34) and the α draws combined form a good basis for determining convergence for the entire model. Traceplots, Geweke's statistic and Gelman-Rubin's statistic are all used for this.

The downside of the Gelman-Rubin statistic, especially for high dimensional problems, is that finding good overdispersed initial conditions is rather complicated (Rossi et al., 2005). This is particularly true for the DPM voter choice model defined in the previous sections. Therefore, an attempt to find overdispersed initial conditions is made by finding crude estimates obtained from the hyperparameters (Gelman et al., 2014). Most hyperparameters are specified with a uniform distribution. Two initial starting points can be obtained from this by initializing all hyperparameters on either the lower bound or the upper bound. The initial condition of all other parameters is based on the value of the hyperparameters. The mode implied by the initial conditions for the hyperparameters is taken as a base and multiplied by a factor to make it more overdispersed[21]. For these two initial conditions, all data points are given their own cluster with a large value for α to encourage a large number of clusters in the initial iterations. Besides this, another chain is run with hyperparameters initialized in the middle of their range. All data points are assigned to the same cluster with a very low value for α. All other parameters are sampled from the resulting distributions.

8.2 Forecasting

Forecasting in Bayesian analysis boils down to computing the predictive density of a future observation given the training data. In terms of the general DPM model specification, this density can be denoted as $p(d_{n+1}|D_{\text{train}})$. This expression is defined as follows

$$\begin{aligned} p(d_{n+1}|D_{\text{train}}) &= \int p(d_{n+1}|\theta_{n+1}) p(\theta_{n+1}|D_{\text{train}}) d\theta_{n+1} \\ &= \int p(d_{n+1}|\theta_{n+1}) \int \cdots \int p(\theta_{n+1}|\theta_1, \cdots, \theta_n) p(\theta_1, \cdots, \theta_n|D_{\text{train}}) d\theta_1, \cdots, d\theta_{n+1} \\ &= \int \cdots \int p(d_{n+1}|\theta_{n+1}) p(\theta_{n+1}|\theta_1, \cdots, \theta_n) p(\theta_1, \cdots, \theta_n|D_{\text{train}}) d\theta_1, \cdots, d\theta_{n+1} \end{aligned}$$

where $p(\theta_{n+1}|D_{\text{train}})$ is obtained from the posterior distribution of the parameters θ_i for the training data. For the model at hand, this integral is not analytically tractable and is approximated using simulation. Thus, $p(d_{n+1}|D_{\text{train}})$ can be estimated by Monte Carlo integration. For R posterior draws, this estimate is

$$\hat{p}(d_{n+1}|D_{\text{train}}) = \frac{1}{R} \sum_{r=1}^{R} p(d_{n+1}|\theta_{n+1}^r)$$

Draws from $p(\theta_1, \cdots, \theta_n|D_{\text{train}})$ are obtained from the hybrid MCMC sampler discussed in previous sections. Given each posterior draw, the distribution of θ_{n+1}^r is essentially a weighted average of the prior with weight $\frac{\alpha}{\alpha+n}$ and the empirical posterior distribution of the parameters with weight $\frac{n}{\alpha+n}$ (Rossi, 2014). A draw θ_{n+1}^r is obtained from constructing this weighted average.

[21] This factor ranges between $20 - 100$ for the upper bound and $1/100 - 1/20$ for the lower bound depending on the parameter in question.

This forecasting scheme can be easily adapted to the voter choice model. In this model, we are interested in a forecasts for the party choices of an individual $n+i$, which is based on posterior predictive probabilities $p(y_{n+i} = j|x_{n+i}, D_{\text{train}})$. These probabilities can be directly obtained computing the Monte Carlo estimate for $p(y_{n+i} = j|x_{n+i}, D_{\text{train}})$ using the conditional distribution in equation (24). Furthermore, using the definition of conditional probability, the probability can be indirectly derived from the joint distribution using

$$\hat{p}(y_{n+i} = j|x_{n+i}, D_{\text{train}}) = \frac{\hat{p}(x_{n+i}, y_{n+i} = j|D_{\text{train}})}{\hat{p}(x_{n+i}|D_{\text{train}})} \tag{35}$$

The latter option has the advantage that it exploits the information the covariates of individual $n+i$ carry when deciding on the previously mentioned weights (Shahbaba and Neal, 2009). Both methods are used to generate forecasts.

Finally, the forecast for the election outcome, defined as the share of votes for each party, is obtained by taking the average of the posterior predictive probabilities for each individual in (35)

$$\hat{p}(y = j|D_{\text{train}}) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \hat{p}(y_{n+i} = j|x_{n+i}, D_{\text{train}}) \tag{36}$$

9 Selecting Parameter Values For Hyperparameter Distributions

To finalize the model, the parameters of the hyperparameter distributions are specified. First, a baseline prior specification is chosen. This is the main focus of this section. From this point, more sets of prior specification are determined by adopting small differences to the baseline prior specification.

To simplify the prior specification, the continuous data is standardized. Due to this, the prior of μ_μ is centered on 0 and while still given a relative wide range to reflect uncertainty caused by the DPM specification. The hyperparameters of Σ_μ are chosen such that the mode of the draws range between approximately $(0.5, 1.4)$ while the corresponding expected variance ranges between approximately $(0.4, 3.5)$. This reflects the belief that the variance of the mean is likely to be lower than the variance of the data, while still leaving room for the prior to become less strong (higher variance) in case necessary.

$$\begin{aligned} \mu_\mu|-2,2 &\sim unif(-2,2) \\ \nu_\mu &= 25 \\ \upsilon_\mu|0.5,1.5 &\sim unif(0.5,1.5) \end{aligned}$$

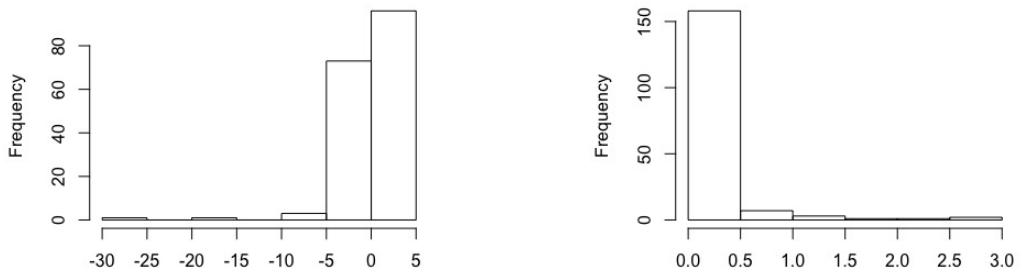
The prior specification of Σ is finalized based on more or less the same principles. The a priori expected rage for the mode of these draws contains the variance of the standardized continuous data,

but maintains the possibility of other values. The tightness parameter here is lower than ν_μ to reflect the fact that there is more uncertainty about the distribution of Σ compared to μ.

$$\nu_\Sigma = 10$$
$$\upsilon_\Sigma | 1, 2 \sim unif(1, 2)$$

Due to the symmetric specification of the Dirichlet prior, the parameter $a_{0,l}$ only determines whether the prior distribution a) is unimodal with the mode located on $(1/M_l, \cdots, 1/M_l)$ and b) how peaked this mode is. When $a_{0,l} < M_l$, the distribution has no mode. In this case, the prior puts most mass on the edges of the parameters space. Hence, $a_{0,l} < M_l$ implies that, a priori, it is believed that the probabilities are not uniform; some categories are a priori more likely than others. It is not possible to make these kinds of assumptions without examining the data first. Therefore the prior distribution of a_0 is specified such that it accounts for all these possibilities.

$$a_{0,l} | 1, 1000 \sim unif(1, 1000) \quad \forall l = 1, \cdots, L^{Cat}$$



(i) Histogram of β_{ML} coefficients.

(ii) Histogram of the variance of β_{ML} coefficients.

Figure 1: Histograms of the ML coefficients of the β parameters and their variances. In the histogram of the variances two observations are excluded. These values fall outside of the domain in the histogram, making the figure unclear.

The ML coefficients of the β parameters and their variances are used to specify the distributions of λ_β. Figure 1 shows a histogram of these. The aim here is to get an indication of the scale of the β parameter, which serves as a reference point for the boundaries of $\bar{\beta}$. Most coefficients have values ranging between -5 and 5, which determines the choice of prior specification for $\bar{\beta}$. A similar strategy is adopted for the variance parameters. The respective ML variances are taken as a rough lowerbound to the location of variance distribution in the DPM model. Most MLE variances fall in a 0 to 3 region. Hence the range for υ_β is specified such that it accounts for modes $(2.6, 10.4)$. The tightness parameter is chosen such that it is not too low as this leads to a more right skewed distribution and drives the a priori expected value of B up. At the same time it is chosen not too high to reflect the uncertainty in the prior specification and give the data more power to influence the posterior.

$$\begin{aligned}
&\bar{\beta}_l| - 5, 5 \sim unif(-5, 5) \quad \forall l = 1, \cdots, L^* \\
&\nu_\beta = 190 \\
&\upsilon_\beta|5, 20 \sim unif(5, 20)
\end{aligned}$$

The prior distribution of α is determined based on the implied prior probabilities on the number of clusters. α is important to the model since it reflects the strength of prior. Furthermore, it is strongly related to the number of clusters. Low values of α imply little confidence in the prior and a small number of clusters. As the number of underlying components is unknown, it is desirable to specify a prior distribution that puts mass on a large range of values of α. At the same time, to reflect the uncertainty in the prior specification, small values of α are desired. This implies that the distribution should have a relatively large variance and small mean. Based on these requirements, the following prior is chosen

$$\alpha|2, 1 \sim Ga(2, 1)$$

which puts substantial mass on $\alpha \in (0, 4)$. Antoniak (1974) shows the relationship between α and the number of clusters. Using his results, the prior probability for a given amount of clusters can be expressed as

$$P(K = k|\alpha, N) = |S_{n,k}|N!\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \tag{37}$$

where $S_{n,k}$ represents the Stirling numbers of first kind. Using this expression prior probabilities are calculated. Table 10 in Section B shows the prior probabilities for different values of α. These figures support the choice for $Ga(2, 1)$ as the prior allows for a relatively large number of clusters (over 50), but puts most mass on a moderate number of clusters.

By making small changes to the above mentioned baseline, 19 prior specifications are obtained, which are labeled as prior specifications A-S. These are displayed in Table 3. Model D is equivalent to the above discussed baseline. The table also shows additional prior specifications T-Z. As discussed in Section 10, these have been selected after examining the results of the 2010 election and are only used to forecast the 2012 election outcome.

Table 3: Prior Specifications A - Z

	$\mu_{\mu,l}$[1]	υ_μ	ν_μ	υ_Σ	ν_μ	$a_{0,l}$[2]	$\bar{\beta}_l$[3]	υ_β	ν_β	α
A	(-2,2)	(0.5,1.5)	25	(1,2)	10	(1,1000)	(-5,5)	(0.5,5)	250	(2,1)
B	(-2,2)	(0.5,1.5)	25	(1,2)	10	(1,1000)	(-5,5)	(0.5,5)	190	(2,1)
C	(-2,2)	(0.5,1.5)	25	(1,2)	10	(1,1000)	(-5,5)	(5,20)	250	(2,1)
D	(-2,2)	(0.5,1.5)	25	(1,2)	10	(1,1000)	(-5,5)	(5,20)	190	(2,1)
E	(-2,2)	(0.5,1.5)	25	(1,2)	10	(1,1000)	(-5,5)	(10,30)	190	(2,1)
F	(-2,2)	(0.5,1.5)	15	(1,2)	10	(1,1000)	(-5,5)	(5,20)	190	(2,1)
G	(-2,2)	(2,10)	25	(1,2)	10	(1,1000)	(-5,5)	(5,20)	190	(2,1)
H	(-2,2)	(0.5,1.5)	25	(1,2)	5	(1,1000)	(-5,5)	(5,20)	190	(2,1)
I	(-2,2)	(0.5,1.5)	25	(5,10)	10	(1,1000)	(-5,5)	(5,20)	190	(2,1)
J	(-2,2)	(0.5,1.5)	25	(1,2)	10	(1,1000)	(-5,5)	(5,20)	190	(10,1)
K	(-2,2)	(5,10)	15	(10,20)	5	(1,1000)	(-5,5)	(5,20)	190	(2,1)
L	(-2,2)	(2,10)	15	(1,2)	10	(1,1000)	(-5,5)	(5,20)	190	(2,1)
M	(-2,2)	(2,10)	25	(5,10)	5	(1,1000)	(-5,5)	(5,20)	190	(2,1)
N	(-2,2)	(2,10)	15	(5,10)	5	(1,1000)	(-5,5)	(5,20)	190	(2,1)
O	(-2,2)	(5,10)	10	(10,15)	5	(1,1000)	(-5,5)	(5,20)	190	(2,1)
P	(-2,2)	(2,10)	15	(5,10)	5	(1,100)	(-5,5)	(5,20)	190	(10,1)
Q	(-2,2)	(2,10)	15	(5,10)	5	(1,1000)	(-5,5)	(5,20)	190	(10,1)
R	(-10,10)	(2,10)	15	(5,10)	5	(1,1000)	(-20,20)	(5,20)	190	(2,1)
S	(-2,2)	(2,10)	15	(5,10)	5	(1,1000)	(-5,5)	(5,20)	190	(100,1)
T	(-2,2)	(5,10)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(2,1)
U	(-2,2)	(7,12)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(2,1)
V	(-2,2)	(5,10)	15	(15,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(2,1)
W	(-2,2)	(7,12)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(2,1)
X	(-2,2)	(5,10)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(1,1)
Y	(-2,2)	(7,12)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(1,1)
Z	(-2,2)	(5,10)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	190	(0.5,1)

[1] For all $\mu_{\mu,l}$ parameters with $l \in (1, \cdots, L^{Con})$
[2] For all $a_{0,l}$ parameters with $l \in (1, \cdots, L^{Cat})$
[3] For all $\bar{\beta}_l$ parameters with $l \in (1, \cdots, L^*)$

10 Results

The primary interest of this research is to evaluate the model in terms of predictive performance of election outcomes. To this end, the sum squared errors (SSE) is calculated for each election outcome forecast. This statistic provides a good summary of the distance between the predicted party shares and the actual election outcome. Using (36), the SSE is computed in two distinct ways

$$SSE_i = \sum_{j \in \mathcal{J}_i} (P(y = j|D_{\text{train}}) - ps_j)^2$$

for $i \in (1, 2)$ where ps_j is the actual party share of party j. Here, $\mathcal{J}_1$ is defined as the set of (VVD, PvdA, PVV, CDA, Other), while $\mathcal{J}_2$ denotes the set of (VVD, PvdA, PVV, CDA, Other, No Vote).

To obtain the results showcased below, three chains are run for each DPM model for 40000 iterations. These chains are initialized as described in Section 8.1. The number of auxiliary variables used in Neal's Algorithm 8 is set to 3. Due limited storage, thinning is performed on the output. Only 1 out of 4 draws is retained. This results in a sample of 10000 draws per chain. The burn-in sample is established using the convergence diagnostics discussed in Section 8. These are reported and evaluated in Section 10.1.

Four benchmark models are selected to generate forecasts that serve as a comparison to the results of the DPM models. The first benchmark forecast (VI) is generated by aggregating the voting intentions in each sample into party shares[22]. Besides this, three models are selected based on commonly used approaches to multinomial logit models, which are the maximum likelihood model (ML) and two Bayesian models. Similar to the DPM models, the Bayesian models make use of a Random Walk sampler under asymptotically efficient scaling. The two Bayesian models differ only in their prior specification. Whereas both are assigned a multivariate normal prior $N(0, B)$, the Bayes 1 model is specified with a small prior variance $B = 1.5I_{L^*}$ and the Bayes 2 model is specified with a large variance $B = 100I_{L^*}$. The Bayesian benchmark models are each run for 40000 iterations, where the initial 3000 are discarded as burn-in sample[23].

All models are run using both the 2010 and 2012 data set. To evaluate whether the DPM model is subject to overfitting, the data is split into a training and a test data set. The training data set amounts to 2000 randomly chosen individuals, which roughly splits the data set in half. The previously mentioned statistics are calculated for both years and both types of data sets. The results of the 2010 data set are used to formulate additional models for the 2012 data set to attempt to improve upon the prior knowledge and thus forecasting performance.

The remainder of this section is organized as follows. First, practical matters pertaining to computation time are discussed in section 10.1. After this, the convergence statistics of the DPM models are assessed in section 10.2. The forecasting performance of all models is evaluated in section 10.3. Finally, some posterior results are discussed in section 10.4.

10.1 Computation Time

A major disadvantage of the DPM voter choice model is the computation time required to sample sufficient draws. There are several factors that contribute to this result.

Firstly, the number of data points in the training sample is strongly related to the computation time. The reason for this is that the cluster assignment of each data point is determined using the cluster assignments of all other data points. Due to this, sequential sampling where all computations have to be repeated for each data point, is necessary. In practice, this takes a relatively long time. The second important factor that contributes to the long computation time is the dimension of the DPM voter choice model. It takes much longer to sample from and evaluate densities of parameters with a large dimension. Especially the computations related to the β parameter take much time. On top of this, due to the large number of parameters that need to be saved, each simulation uses substantial working memory. Lastly, the chains of the DPM models exhibit high correlation due to which large samples of posterior draws are required for both convergence detection and accurate inference. These two factors increase the total computation time of the simulation as well.

Applying the hybrid sampler programmed in R (Version 3.5.0) on the data at hand results in a computation time of 2-3 days for 10000 iterations[24]. Clearly, this is undesired and makes running multiple chains for a large number of iterations unfeasible. Part of the problem is caused by the programming language. R is known to be very efficient with matrix computations, but can be slow in on other aspects. The Rcpp package offers a (partial) solution to this problem. This package

[22] For example, the forecasted party share of party j is obtained by computing $\frac{\sum_{i=1}^{n} I(y_i=j)}{\sum_{j \in \mathcal{J}_1} \sum_{i=1}^{n} I(y_i=j)}$.

[23] This was sufficient to achieve convergence of the likelihood based on visual expectation and the Geweke statistic with values ranging from -0.5 to 0.3.

[24] All figures are based on computations carried out using a desktop with an Intel® Core™ i7-7700K processor (overclocked to 4.7 GHz).

offers an integration of the C++ language into the R environment. C++ is known to be able to perform computations up to 50-100 times faster compared to R. Using this package, some functions are rewritten in C++ and integrated into the original code. This results in a computation speed of 1000 iterations per approximately 25 minutes while running three chains parallel to each other. The written code, which consist of both R and C++ functions, is consolidated and turned into the R-package *DPMMmlogit*. This package is made publicly available and can be obtained from https://github.com/banuatav/DPMMmlogit.

10.2 Convergence Diagnostics

Tables 4 and 5 present the convergence diagnostics for each simulation. The burn-in column shows how many draws are discarded for each chain. The sample size column displays the number of draws that are retained to use for forecasting purposes. To establish a forecast for each model, all three chains are combined after discarding the burn-in sample. For all models except model G, the initial run of 40000 iterations was sufficient to establish convergence. Due to high autocorrelation, convergence in model G turned out to be slower. For this reason, the chain is run for an additional 16000 iterations.

The tables show the Geweke statistic computed for each chain after discarding the burn-in sample. Given that the Geweke statistic concerns a z-test on the equality of two means computed with different parts of the chain, its null hypothesis corresponds to convergence of the chain (equal means). Hence, absolute values greater than 1.96 indicate 95% confidence in rejection of this null hypothesis. The tables also report the Gelman-Rubin statistic; both as point estimate and upper limit of the confidence interval (Upper CI). These are computed using all three chains after discarding the burn-in sample. Values of smaller than 1.2 are considered acceptable to conclude convergence. Both statistics are computed for the variable α and the log conditional posterior likelihood (LCPL).

A visual inspection of the traceplots served to generate an initial proposal for the burn-in sample. The experience is that the chains display high autocorrelation as a result of the Random walk sampler, with acceptance rates of about $20 - 25$ %, and the high dimensionality. However, the traceplots did show a clear indication of convergence. In general, the chains tend to highly deviate from the eventual stable mean in the first $2000 - 4000$ draws.

After the initial proposal, the Geweke and Gelman-Rubin statistics are used to finalize the decision on the number of draws to discard. In most cases, small adjustments to the initial proposal are made to arrive at the results shown below. The Geweke and Gelman-Rubin statistics do not always lead to the same conclusion. Due to the high amount of autocorrelation in the chains, the Geweke statistic is very sensitive to the window of draws that is used. Therefore, the Gelman-Rubin statistic served as the decisive criterion when the two methods disagreed. Due to this, some of the results below show Geweke statistics that are a bit higher than the critical value 1.96.

10.3 Forecasting Performance

Table 6 displays the SSE for all models applied to the 2010 and 2012 training data respectively. The table is split in two parts. The leftmost columns display the SSE computed based on the actual (population) election outcome in 2010 and 2012. The rightmost columns show the SSE based on the election outcome in the respective sample. A subscript 1 in the table denotes the SSE computed for the errors in the party shares of VVD, PvdA, PVV, CDA and Other (the No Vote option is not included), while a subscript 2 denotes the SSE computed based on the previous mentioned parties

and No Vote. The SSE of the DPM models are obtained in two distinctive ways: based on the forecasted conditional probability implied by the MNL model only, denoted with SSE C, or based on the forecasted conditional probability implied by the joint distribution, denoted with SSE J. Tables 11 and 12 display the party share forecasts for all models and forecasting approaches. These can be found in appendix C.

The table shows that, among the benchmark models, the ML and VI models have similar forecasts. The Bayesian benchmark models generate forecasts that are different in the sense that the share of the Other option is predicted to be lower, which improves the forecasts of the party shares of all other options as well. The election outcome forecasts clearly benefit from the Bayesian approach, which are computed while accounting for the uncertainty in the posterior distributions, whereas the ML and VI forecasts completely ignore this uncertainty. All in all, the Bayes 2 model performs best compared to all other benchmark models.

Table 4: Convergence diagnostics of all DPM model for the 2010 data set.

				Geweke Statistic			**Gelman-Rubin Statistic**	
	Burn-in[1]	Sample size[2]		Chain 1	Chain 2	Chain 3	Point Estimate	Upper CI
A	4000	6000	α	0.301	-0.967	1.343	1.007	1.025
			LCPL	1.584	-0.150	1.656	1.026	1.081
B	4500	5500	α	1.609	0.338	1.458	1.009	1.030
			LCPL	-1.537	-0.178	0.673	1.017	1.052
C	5000	5000	α	-0.903	-1.407	-0.707	1.002	1.009
			LCPL	-0.940	0.610	1.913	1.010	1.029
D	4000	6000	α	-0.264	-0.824	1.427	1.016	1.053
			LCPL	-0.117	0.995	-1.210	1.004	1.013
E	4750	5250	α	1.011	-0.055	0.810	1.008	1.028
			LCPL	1.956	-0.303	1.126	1.006	1.017
F	6750	3250	α	0.248	-0.809	1.904	1.003	1.010
			LCPL	1.160	0.524	-0.736	1.006	1.020
G	8250	5250	α	0.313	-0.335	-2.482	1.003	1.011
			LCPL	-0.573	-1.584	0.971	1.001	1.002
H	2000	8000	α	0.639	0.656	0.671	1.001	1.002
			LCPL	0.445	0.503	-1.010	1.003	1.008
I	2250	7750	α	0.525	1.607	0.199	1.000	1.000
			LCPL	-1.472	-0.267	-1.297	1.000	1.001
J	5750	4250	α	2.350	1.143	0.802	1.009	1.032
			LCPL	-1.870	-0.698	0.302	1.000	1.001
K	2750	7250	α	-1.216	1.170	0.267	1.000	1.001
			LCPL	1.257	-1.239	-1.265	1.001	1.002
L	4000	6000	α	1.244	-1.744	0.081	1.011	1.038
			LCPL	1.694	0.587	-1.652	1.019	1.055
M	4500	5500	α	-0.127	1.420	1.821	1.000	1.001
			LCPL	0.226	0.303	-0.550	1.002	1.002
N	5000	5000	α	0.325	1.421	1.234	1.000	1.001
			LCPL	0.090	-0.426	0.056	1.001	1.001
O	3000	7000	α	0.004	1.013	1.002	1.000	1.001
			LCPL	0.765	0.565	0.600	1.018	1.021
P	3500	6500	α	0.100	-1.646	2.484	1.005	1.019
			LCPL	1.132	0.929	-1.513	1.002	1.008
Q	4500	5500	α	1.045	1.096	-0.322	1.028	1.069
			LCPL	-0.660	-0.207	0.129	1.005	1.008
R	7500	2500	α	0.695	-0.101	0.429	1.000	1.000
			LCPL	0.797	-0.395	-0.655	1.004	1.005
S	5500	4500	α	-0.514	1.304	0.244	1.000	1.000
			LCPL	-0.810	-1.044	-1.403	1.003	1.010

Upper CI = Upper bound of the 95% confidence interval of the Gelman-Rubin statistic, LCPL = Log conditional posterior likelihood

[1] Burn-in sample of each chain after thinning.

[2] Remaining sample size after discarding burn-in and after thinning.

Table 5: Convergence diagnostics of all DPM model for the 2012 data set.

				Geweke Statistic			Gelman-Rubin Statistic	
	Burn-in[1]	Sample size[2]		Chain 1	Chain 2	Chain 3	Point Estimate	Upper CI
A	6000	4000	α	1.553	0.206	0.292	1.009	1.032
			LCPL	0.467	-0.516	-0.675	1.023	1.072
B	5000	5000	α	0.563	0.864	0.564	1.003	1.008
			LCPL	0.107	1.081	0.448	1.005	1.016
C	4000	6000	α	0.954	1.597	-1.770	1.005	1.020
			LCPL	-0.346	-1.848	1.047	1.002	1.006
D	4250	5750	α	-1.861	1.272	0.297	1.000	1.000
			LCPL	0.346	0.560	0.046	1.002	1.007
E	5750	4250	α	1.064	-1.039	0.257	1.001	1.004
			LCPL	-0.364	-0.036	0.613	1.004	1.013
F	4000	6000	α	-0.276	0.210	0.819	1.025	1.081
			LCPL	-0.437	0.433	1.270	1.030	1.087
G	4000	6000	α	0.962	0.897	0.843	1.001	1.002
			LCPL	-1.364	-0.054	0.059	1.010	1.031
H	5250	4750	α	-0.430	0.742	0.475	1.033	1.115
			LCPL	-0.312	-0.087	1.036	1.032	1.109
I	4000	6000	α	-0.276	0.210	0.819	1.025	1.081
			LCPL	-0.437	0.433	1.270	1.030	1.087
J	7000	3000	α	-0.531	-1.086	-0.032	1.004	1.014
			LCPL	0.544	0.074	1.255	1.007	1.025
K	4000	6000	α	-0.619	1.301	-0.796	1.001	1.003
			LCPL	1.641	-1.563	1.198	1.065	1.076
L	2250	7750	α	-1.771	0.956	-1.254	1.005	1.017
			LCPL	1.048	-1.036	1.718	1.004	1.016
M	3500	6500	α	0.897	-0.773	1.516	1.001	1.004
			LCPL	0.660	-0.622	-0.565	1.008	1.013
N	6500	3500	α	-0.463	0.789	-0.464	1.001	1.002
			LCPL	-1.015	-1.350	-0.955	1.013	1.015
O	5500	4500	α	0.097	1.903	-0.985	1.040	1.132
			LCPL	0.841	-0.435	-0.479	1.040	1.042
P	6750	3250	α	0.319	-0.397	1.883	1.000	1.000
			LCPL	-1.199	0.679	-0.597	1.002	1.004
Q	4250	5750	α	0.871	0.373	-1.503	1.000	1.000
			LCPL	0.257	0.704	1.718	1.005	1.009
R	6000	4000	α	-1.115	0.008	1.375	1.034	1.114
			LCPL	-0.297	-0.276	-0.698	1.044	1.080
S	3000	7000	α	-0.852	-0.302	-0.985	1.000	1.001
			LCPL	0.915	-0.700	-0.312	1.000	1.001
T	3750	6250	α	-0.960	0.984	0.174	1.001	1.003
			LCPL	1.541	-0.296	1.267	1.019	1.020
U	3750	6250	α	-1.077	-0.625	0.527	1.036	1.123
			LCPL	0.092	0.602	0.059	1.047	1.054
V	5500	4500	α	-1.798	0.183	0.344	1.001	1.003
			LCPL	0.994	-1.508	-1.237	1.059	1.064
W	2000	8000	α	1.358	0.071	0.194	1.000	1.000
			LCPL	-1.044	-0.874	-0.027	1.036	1.037
X	2250	7750	α	1.195	-0.706	1.109	1.000	1.001
			LCPL	0.258	0.417	-1.006	1.020	1.021
Y	2250	7750	α	1.892	-1.117	0.789	1.001	1.002
			LCPL	-1.075	0.646	1.241	1.057	1.059
Z	4750	5250	α	-1.385	-1.435 -	1.318	1.001	1.003
			LCPL	0.785	1.833	0.295	1.123	1.124

Upper CI = Upper bound of the 95% confidence interval of the Gelman-Rubin statistic, LCPL = Log conditional posterior likelihood

[1] Burn-in sample of each chain after thinning.

[2] Remaining sample size after discarding burn-in and after thinning.

For the DPM models, it holds that SSE C is generally lower compared to SSE J. This is especially true for SSE_1. All models perform badly in forecasting the actual No Vote share. Partially, this is due to the fact that the No Vote response is highly underestimated in the data at hand, which distorts the shares of other parties too. The different prior specifications in the DPM models seem to mostly affect the forecasts obtained from the conditional distribution. Forecasts obtained from the joint distribution are much less reactive to changes in prior specification.

All models perform rather badly at predicting the actual population outcome. This can be seen from

the high SSE reported in the first four columns of Table 6. This is because, even though the panel respondents are selected such that they are representative of the Dutch population, the actual voter choices in the resulting data set after collecting the data and combining the variables, deviates from the population outcome. Part of this is caused by the DNK/PNTS option. Yet, this option cannot explain the complete deviation as it only amounts to about 3-4% in the actual voter choices. Another source of error is the fact that many respondents misremember or misreport their actual election choice. This is evident from comparing their reported voting choice for the same election over several years. The remainder of the error is likely to be the result of selection bias in the used sample.

10.3.1 2010 Forecasts

The 2010 forecasts obtained from the joint distribution all heavily underestimate the shares of parties VVD and PvdA, while the share of Other is heavily overestimated. The forecasts obtained from the conditional distribution do better on this front. They, however, heavily underestimate the No Vote option. Furthermore, although all models and forecasting approaches overestimate the share of PVV and CDA, the forecasts obtained from the conditional distribution overestimate these share more. The forecasts of the benchmark models behave similarly to forecasts obtained from the joint approach.

The DPM models generally perform better than the benchmark models. A clear exception to this rule are models A, B and C. These tend to have a higher SSE compared to the Bayes 2 model. Compared to all other DPM models, these models give the prior distribution for the β parameter more weight as they correspond to lower υ_β (prior 'variance') and/or higher ν_β (degrees of freedom). The improvement in forecasting performance from choosing a less influential prior for β is greater for the forecasts obtained from the conditional distribution compared to the forecasts obtained from the joint distribution. This improvement in SSE C continues with further increasing υ_β when moving from model D to E. The 'better fitting' β parameters could arise from two factors. First, an increase in prior variance gives more weight to the data to determine β and decreases the weight of the prior distribution. Second, this change affects the clustering of the data points. Better fitting β parameters imply a greater influence of the density of the response variable relative to the density of the covariates. Clustering dominated by the β parameters is likely to improve the fit of these parameters as well.

Forecasts obtained from the joint distribution perform generally best under a prior specification that puts mass on smaller (co)variance for the continuous variables and medium to high (co)variance for the β parameters. Differences in other parameters do not affect the forecasts much. These prior specifications result, a posteriori, in medium levels of alpha (1.0-1.2), a high number of clusters (8-10) and very high values of $a_{0,l}$ for some of the categorical covariates relative to other models. The posterior distribution of $a_{0,l}$ puts larger mass on high values off $a_{0,l}$ when the corresponding $\pi_{l,m}$ are more uniform. This is likely to be a result of the clustering being dominated by the continuous covariates due to their small variance.

Under the conditional approach, prior specification K performs best. It has a relatively low SSE_2 (only 0.0002 higher than the lowest), and the lowest SSE_1 among all models. A posteriori, model K has a larger variance for the continuous covariates and the β parameters. Furthermore, it exhibits a relatively low α with mean 0.73 and, due to this, a relatively small number of clusters (on average 5). Its α parameter is lower compared to models with the same prior specification on this parameter caused by the large variance as predicted on beforehand.

Table 6: SSE of all models applied to the 2010 and 2012 training data.

2010

	Deviation from actual election outcome				**Deviation from sample election outcome**			
	SSE_1		SSE_2		SSE_1		SSE_2	
VI	0.0318		0.0483		0.0226		0.0231	
ML	0.0317		0.0482		0.0226		0.0231	
Bayes 1	0.0284		0.0443		0.0202		0.0206	
Bayes 2	0.0276		0.0434		0.0197		0.0201	
	SSE J_1	SSE C_1	SSE J_2	SSE C_2	SSE J_1	SSE C_1	SSE J_2	SSE C_2
A	0.0237	0.0296	0.0419	0.0515	0.0169	0.0203	0.0178	0.0220
B	0.0234	0.0211	0.0414	0.0510	0.0167	0.0154	0.0174	0.0198
C	0.0202	0.0125	0.0372	0.0449	0.0145	0.0128	0.0150	0.0182
D	0.0195	0.0102	0.0362	0.0426	0.0140	0.0085	0.0145	0.0139
E	0.0200	0.0097	0.0360	0.0422	0.0144	0.0080	0.0148	0.0134
F	0.0204	0.0104	0.0370	0.0416	0.0146	0.0091	0.0151	0.0140
G	0.0201	0.0107	0.0368	0.0425	0.0144	0.0092	0.0149	0.0143
H	0.0200	0.0094	0.0373	0.0419	0.0143	0.0083	0.0149	0.0137
I	0.0222	0.0126	0.0386	0.0418	0.0160	0.0087	0.0165	0.0129
J	0.0201	0.0096	0.0375	0.0452	0.0143	0.0081	0.0149	0.0148
K	0.0220	0.0119	0.0383	0.0408	0.0156	0.0075	0.0160	0.0115
L	0.0200	0.0097	0.0366	0.0404	0.0143	0.0081	0.0149	0.0128
M	0.0216	0.0132	0.0379	0.0408	0.0154	0.0084	0.0158	0.0120
N	0.0215	0.0138	0.0377	0.0424	0.0153	0.0090	0.0158	0.0129
O	0.0217	0.0143	0.0378	0.0419	0.0155	0.0091	0.0160	0.0127
P	0.0212	0.0124	0.0375	0.0422	0.0152	0.0086	0.0157	0.0130
Q	0.0213	0.0124	0.0377	0.0412	0.0153	0.0080	0.0158	0.0120
R	0.0218	0.0121	0.0391	0.0375	0.0155	0.0085	0.0161	0.0113
S	0.0221	0.0102	0.0386	0.0461	0.0158	0.0085	0.0163	0.0154

2012

	Deviation from actual election outcome				**Deviation from sample election outcome**			
	SSE_1		SSE_2		SSE_1		SSE_2	
VI	0.0351		0.0533		0.0287		0.0288	
ML	0.0351		0.0533		0.0287		0.0288	
Bayes 1	0.0319		0.0500		0.0262		0.0263	
Bayes 2	0.0310		0.0486		0.0255		0.0256	
	SSE J_1	SSE C_1	SSE J_2	SSE C_2	SSE J_1	SSE C_1	SSE J_2	SSE C_2
A	0.0292	0.0194	0.0490	0.0473	0.0239	0.0114	0.0242	0.0134
B	0.0288	0.0172	0.0484	0.0478	0.0236	0.0102	0.0239	0.0130
C	0.0247	0.0128	0.0432	0.0461	0.0208	0.0101	0.0210	0.0137
D	0.0251	0.0149	0.0438	0.0486	0.0212	0.0132	0.0214	0.0169
E	0.0248	0.0141	0.0434	0.0496	0.0210	0.0128	0.0212	0.0172
F	0.0249	0.0129	0.0435	0.0470	0.0208	0.0102	0.0210	0.0141
G	0.0255	0.0115	0.0441	0.0449	0.0214	0.0094	0.0216	0.0131
H	0.0251	0.0128	0.0434	0.0475	0.0213	0.0096	0.0214	0.0137
I	0.0273	0.0087	0.0457	0.0327	0.0227	0.0065	0.0228	0.0076
J	0.0249	0.0124	0.0439	0.0477	0.0208	0.0105	0.0211	0.0148
K	0.0268	0.0095	0.0452	0.0302	0.0223	0.0069	0.0224	0.0073
L	0.0253	0.0127	0.0439	0.0462	0.0214	0.0107	0.0216	0.0144
M	0.0260	0.0097	0.0447	0.0350	0.0216	0.0072	0.0218	0.0085
N	0.0258	0.0109	0.0443	0.0363	0.0215	0.0085	0.0217	0.0099
O	0.0268	0.0096	0.0451	0.0329	0.0225	0.0068	0.0227	0.0077
P	0.0259	0.0111	0.0444	0.0384	0.0217	0.0085	0.0219	0.0103
Q	0.0263	0.0092	0.0447	0.0351	0.0219	0.0067	0.0221	0.0081
R	0.0265	0.0111	0.0453	0.0376	0.0220	0.0092	0.0223	0.0108
S	0.0261	0.0126	0.0448	0.0469	0.0218	0.0108	0.0220	0.0147
T	0.0272	0.0088	0.0452	0.0297	0.0228	0.0062	0.0229	0.0067
U	0.0260	0.0103	0.0443	0.0334	0.0217	0.0080	0.0219	0.0089
V	0.0274	0.0096	0.0458	0.0300	0.0228	0.0068	0.0230	0.0072
W	0.0274	0.0091	0.0458	0.0291	0.0228	0.0055	0.0229	0.0059
X	0.0274	0.0083	0.0456	0.0288	0.0228	0.0053	0.0229	0.0058
Y	0.0270	0.0107	0.0452	0.0303	0.0225	0.0077	0.0227	0.0080
Z	0.0279	0.0099	0.0464	0.0303	0.0231	0.0078	0.0233	0.0083

The benchmark forecasts are obtained as follows: VI = Benchmark forecast obtained by aggregating voting intentions, ML = Maximum likelihood model, Bayes 1 = Bayesian model using prior $N(0, B)$ with $B = 1.5I_{L^*}$, Bayes 2 = Bayesian model using prior $N(0, B)$ with $B = 100I_{L^*}$.

For most models, it hold that SSE C decreases with a larger Σ (often at the same time with an increase in Σ_μ) and lower ν_β. Due the fact that the conditional approach tends to produce better results altogether, additional prior specification are formulated for the 2012 data using this result. To test whether the improvement in model performance is caused by the decrease in the α parameter (and not the increased variance), prior specifications that put more mass on lower values of α have been added as well. A summary of the parameter settings of these models can be found in table 7.

Table 7: Prior Specifications T - Z.

	$\mu_{\mu,l}$[1]	υ_μ	ν_μ	υ_Σ	ν_μ	$a_{0,l}$[2]	$\bar{\beta}_l$[3]	υ_β	ν_β	α
T	(-2,2)	(5,10)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(2,1)
U	(-2,2)	(7,12)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(2,1)
V	(-2,2)	(5,10)	15	(15,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(2,1)
W	(-2,2)	(7,12)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(2,1)
X	(-2,2)	(5,10)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(1,1)
Y	(-2,2)	(7,12)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	185	(1,1)
Z	(-2,2)	(5,10)	15	(10,20)	5	(1, 1000)	(-5, 5)	(5, 20)	190	(0.5,1)

[1] For all $\mu_{\mu,l}$ parameters with $l \in (1, \cdots, L^{Con})$
[2] For all $a_{0,l}$ parameters with $l \in (1, \cdots, L^{Cat})$
[3] For all $\bar{\beta}_l$ parameters with $l \in (1, \cdots, L^*)$

10.3.2 2012 Forecasts

The forecasting performance of all models for the 2012 data follow similar general patterns displayed by the 2010 forecasting performance. The best performing models have an even lower SSE for this election. This is likely to be caused due to the fact that the stated voting preferences for the 2012 election are closer to election date compared to the 2010 voting preferences. The DPM models perform better in predicting the No Vote option for this data set. On the other hand, the forecasted No Vote share is further from the actual election outcome of 25.4%. This, however, is caused by a larger gap between the sample and actual election.

The difference compared to the previous election is the fact that forecasts obtained from the joint distribution completely underestimate the PvdA party share. On the other hand, the forecasts of the party shares of VVD, CDA and PVV are much closer to the election outcome.

As expected, models with a lower ν_β perform much better using the conditional approach. This can be seen from comparing the results of model T with K. This change mainly improves the predicted share of CDA. In combination with larger υ_Σ and υ_μ in model W, the improvement in forecasting performance continues by causing a decrease in the predicted PVV party share. However, the best performing model using the conditional approach is model X with lower ν_β and lower α compared to model K. This implies that improved forecasts result from a decrease in both ν_β, which results in a larger β covariance a posteriori, and smaller α.

The DNK/PNTS option has a significant influence on the forecasting results of all models. While in the stated voting intentions the share of this option is 15.4% and 18.4% for 2010 and 2012 respectively, this share is only 4.1% and 3.1% in the election outcome. The forecast of this share is relatively close to the share in the stated voting intentions for all models. When determining the election outcome, its forecast is based on the alternatives VVD, PvdA, PVV, CDA, Other and No Vote only. Therefore, these shares are normalized based on their relative value after discarding the share of DNK/PNTS. Thus, we implicitly assume that discarding DNK/PNTS in this way does not change the election outcome. In other words, we assume that among the subsample of voters that have stated the

DNK/PNTS option in their voting intentions, the shares of the remaining alternatives are distributed according to the shares of the rest of the sample. This assumption, however, does likely not hold.

Table 8 displays the cross-tabulated shares of intended and actual vote. Each percentage is calculated based on the total intended votes per party alternative. The figures show great similarities over the two years. It also shows that the DNK/PNTS option is divided among all parties in a way that does not comply with the sample election outcome. For example, the share of DNK/PNTS respondents that vote for PVV is lower and the share that does not vote is higher compared to the rest of the sample. Due to this, the shares of the alternatives PVV and No Vote are overestimated and underestimated respectively when discarding the DNK/PNTS option while computing forecasts. This tendency sheds some explanation on the reason as to why the models perform badly in forecasting these options.

Table 8: Cross-tabulated shares of intended and actual vote. Each percentage is calculated based on the total intended votes per party alternative.

2010

	VVD	PvdA	PVV	CDA	Other	DNK/PNTS	No Vote
VVD	78.5%	2.0%	3.2%	3.5%	5.2%	0.3%	7.3%
PvdA	3.3%	78.8%	1.8%	0.9%	8.2%	0.9%	6.1%
PVV	20.9%	2.6%	54.5%	2.0%	8.5%	2.0%	9.6%
CDA	12.0%	3.4%	2.4%	67.2%	6.4%	3.2%	5.4%
Other	7.3%	17.4%	4.1%	3.2%	60.4%	1.9%	5.8%
DNK/PNTS	15.9%	14.3%	4.9%	9.4%	24.1%	14.8%	16.6%
No Vote	11.1%	8.1%	7.9%	3.5%	11.1%	4.6%	53.7%

2012

	VVD	PvdA	PVV	CDA	Other	DNK/PNTS	No Vote
VVD	81.3%	2.2%	1.5%	2.5%	7.0%	0.6%	5.1%
PvdA	2.4%	83.8%	0.0%	1.1%	6.2%	1.4%	5.1%
PVV	15.5%	6.1%	49.2%	0.6%	10.3%	1.5%	16.7%
CDA	10.5%	4.0%	0.0%	70.3%	9.8%	0.7%	4.7%
Other	3.3%	24.9%	1.5%	3.6%	59.1%	1.6%	5.9%
DNK/PNTS	17.1%	21.4%	4.4%	8.5%	22.4%	10.6%	15.6%
No Vote	9.1%	12.1%	8.3%	1.6%	15.1%	3.0%	50.8%

10.3.3 Training vs Test Data

Table 9 shows the SSE computed based on the sample election outcome for the 2010 and 2012 test data sets. The election outcomes of these data sets can be found in appendix C, Tables 13 and 14.

Overall, the SSE of the test data shows similar patters to the SSE of the training data. There are some differences in the order of forecasting performance, but these are not large, especially for forecasts obtained from the joint distribution. Tables 13 and 14 indicate that the DPM election forecasts obtained directly from the conditional distribution are less reactive to the change in data set compared to the joint approach or benchmark models. Mostly, the difference in the share of intended No Vote between the test and training samples are reflected in their respective forecasts forecast. This result is consistent with both years.

Most DPM models for 2010 have an SSE almost equal or lower for the test data compared to the

SSE of the training data, while most of the benchmark models have a higher SSE. For the joint approach, the SSE stays rather the same under both the training and test data set. For the conditional approach, on the other, hand, the SSE is much lower for most models. However, as these forecast are so unreactive to the change in data set, changes in the sample shares of the voting intentions are not completely reflected in the forecasts. Therefore, the decline is mainly caused due to the fact that, relative to the training sample election outcome, the test sample party shares are closer to the forecasted shares.

Table 9: SSE of all models applied to the 2010 and 2012 test data.

	Deviation from sample election outcome							
	SSE_1		SSE_2		SSE_1		SSE_2	
VI	0.0222		0.0223		0.0202		0.0206	
ML	0.0244		0.0246		0.0245		0.0249	
Bayes 1	0.0216		0.0218		0.0225		0.0228	
Bayes 2	0.0208		0.0210		0.0215		0.0217	
	SSE J_1	SSE C_1	SSE J_2	SSE C_2	SSE J_1	SSE C_1	SSE J_2	SSE C_2
A	0.0175	0.0227	0.0180	0.0237	0.0211	0.0109	0.0215	0.0130
B	0.0174	0.0152	0.0179	0.0180	0.0208	0.0100	0.0213	0.0127
C	0.0143	0.0101	0.0147	0.0137	0.0172	0.0105	0.0175	0.0140
D	0.0137	0.0065	0.0140	0.0102	0.0173	0.0144	0.0176	0.0180
E	0.0139	0.0058	0.0141	0.0095	0.0169	0.0139	0.0172	0.0181
F	0.0142	0.0070	0.0145	0.0103	0.0174	0.0108	0.0177	0.0145
G	0.0143	0.0070	0.0146	0.0105	0.0171	0.0105	0.0174	0.0141
H	0.0141	0.0060	0.0145	0.0096	0.0181	0.0103	0.0183	0.0143
I	0.0161	0.0077	0.0163	0.0104	0.0186	0.0074	0.0188	0.0085
J	0.0142	0.0057	0.0146	0.0105	0.0168	0.0114	0.0172	0.0154
K	0.0157	0.0069	0.0158	0.0096	0.0181	0.0069	0.0184	0.0074
L	0.0141	0.0060	0.0144	0.0090	0.0173	0.0115	0.0176	0.0152
M	0.0152	0.0080	0.0154	0.0102	0.0168	0.0083	0.0172	0.0096
N	0.0151	0.0082	0.0153	0.0108	0.0169	0.0094	0.0172	0.0107
O	0.0154	0.0089	0.0156	0.0111	0.0178	0.0074	0.0181	0.0083
P	0.0150	0.0078	0.0152	0.0107	0.0171	0.0093	0.0174	0.0111
Q	0.0152	0.0074	0.0154	0.0100	0.0172	0.0077	0.0175	0.0093
R	0.0153	0.0077	0.0157	0.0092	0.0173	0.0102	0.0177	0.0118
S	0.0155	0.0064	0.0158	0.0113	0.0172	0.0120	0.0174	0.0157
T					0.0186	0.0068	0.0188	0.0073
U					0.0172	0.0084	0.0174	0.0093
V					0.0188	0.0069	0.0191	0.0074
W					0.0185	0.0051	0.0187	0.0056
X					0.0185	0.0055	0.0187	0.0061
Y					0.0185	0.0077	0.0187	0.0081
Z					0.0189	0.0078	0.0192	0.0083

The benchmark forecasts are obtained as follows: VI = Benchmark forecast obtained by aggregating voting intentions, ML = Maximum likelihood model, Bayes 1 = Bayesian model using prior $N(0, B)$ with $B = 1.5I_{L^*}$, Bayes 2 = Bayesian model using prior $N(0, B)$ with $B = 100I_{L^*}$.

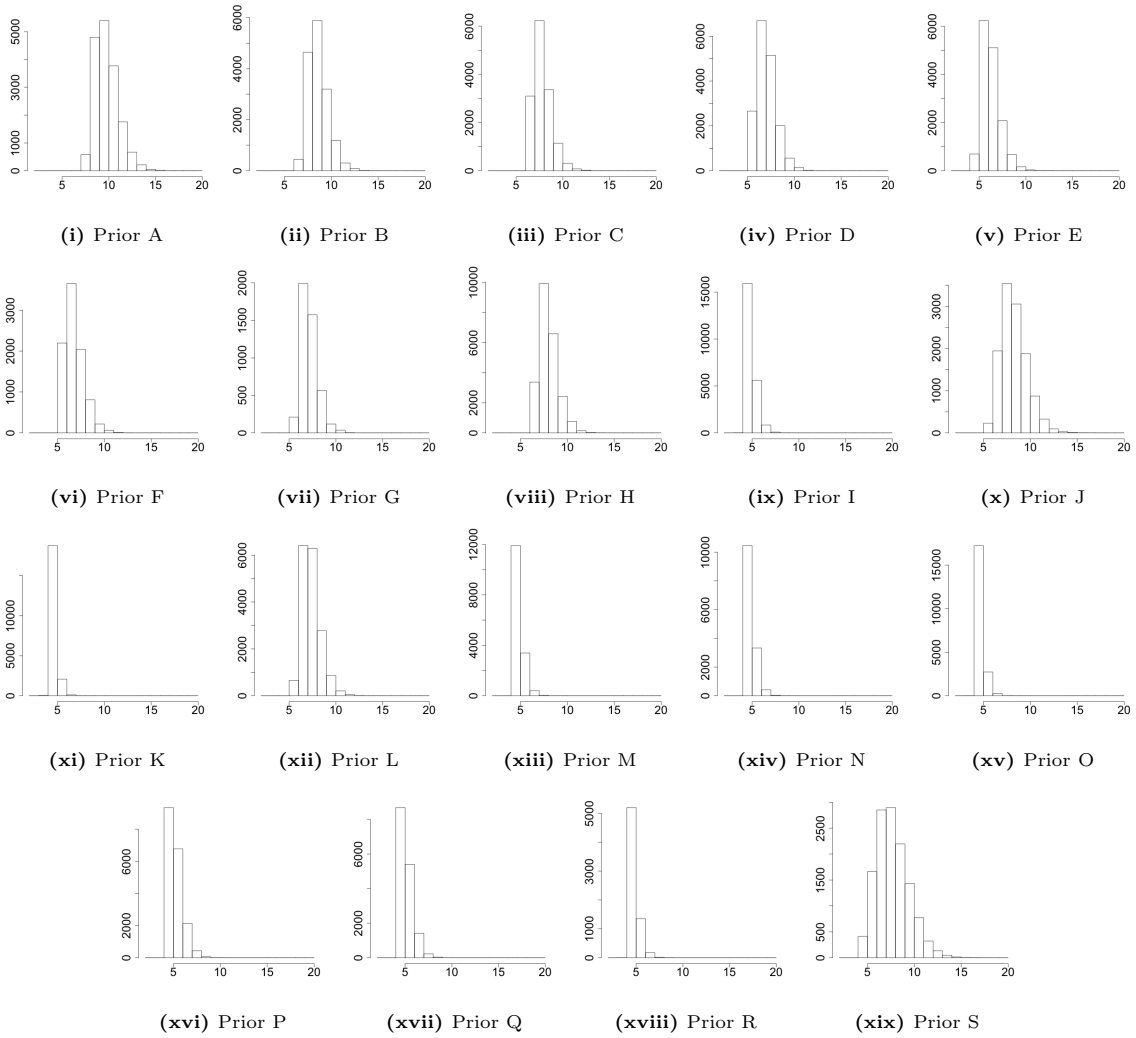
For the 2012 data, the test SSE is lower for all benchmark models and forecasts obtained from the joint distribution. This result is caused due to the fact that, even though the reported intended PvdA vote share is higher for the training sample, the realized PvdA share is lower for the test sample. As this approach tends to largely underestimate the PvdA share, these changes cause a significant improvement of the SSE of the test sample. Forecasts generated using the conditional approach have a similar or slightly higher SSE for the test data.

Overall, there is no indication of a clear pattern of overfitting in the DPM models. This is especially true when compared to the benchmark models.

10.4 Number Of Clusters And Estimates Of β

Figure 2 displays the histograms of the number of clusters K of each DPM model for the 2010 election. The complete set of histograms for both elections can be found in appendix D.

The number of clusters range from about 3 to 15 for all prior specifications. Despite the fact that the prior specification of the parameter α is $(2, 1)$ for most models, the distribution of the number of clusters differs a lot. As previously discussed, this is due to the effect of the hyperparameters λ on the clustering procedure. In particular, Section 6.3 argued that more dispersed specifications of prior distribution lead to a lower number of clusters. This is evident from the histograms as well.



(i) Prior A **(ii)** Prior B **(iii)** Prior C **(iv)** Prior D **(v)** Prior E

(vi) Prior F **(vii)** Prior G **(viii)** Prior H **(ix)** Prior I **(x)** Prior J

(xi) Prior K **(xii)** Prior L **(xiii)** Prior M **(xiv)** Prior N **(xv)** Prior O

(xvi) Prior P **(xvii)** Prior Q **(xviii)** Prior R **(xix)** Prior S

Figure 2: Histograms of the number of clusters K in the posterior draws for the 2010 election under prior specification A-S.

The figures show that, under a given prior specification, the distributions of K are very similar for both election years. The similarities in the posterior distribution of K between the 2010 and 2012 elections imply that the posterior distributions of related parameters, which are the dispersion parameters and α, may exhibit similarities too. This supports the notion that recent previous elections carry some information that is useful for predicting current elections.

Combined with the forecasting performance, the histograms imply that a smaller number of clusters (about 3-5) is preferred for the conditional approach, but a relatively large number of clusters (about 5-12) for the joint approach. The differences in estimating the probability $p(y_{n+1}|x_{n+1}, D_{\text{train}})$ under both approaches can be derived using the forecasting procedure described in Section 10.3. Under the conditional approach, this probability is estimated in each iteration t as

$$\sum_{k=1}^{K} \frac{1}{K+\alpha} p(y_{n+1}|x_{n+1}, \beta_k^{t,*}) + \frac{\alpha}{K+\alpha} p(y_{n+1}|x_{n+1}, \beta_0^t)$$

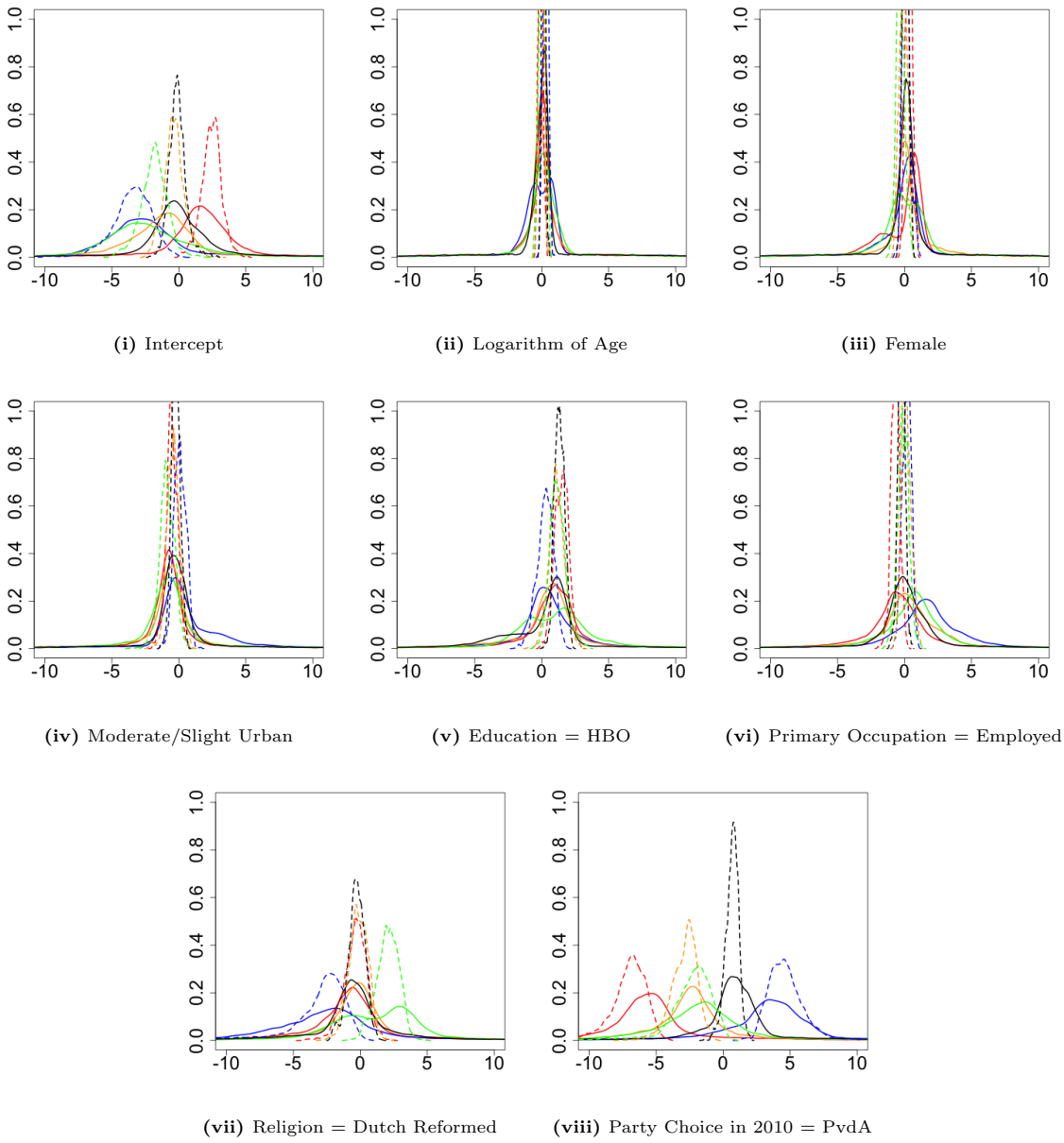
where $p(y_{n+1}|x_{n+1}, \beta)$ is obtained from the multinomial logit specification and β_0^t is a draw from the base distribution of β parametrized using its hyperparameters drawn in iteration t. Using the joint approach, however, this probability is calculated as

$$\sum_{k=1}^{K} \frac{p(x_{n+1}|\mu_k^{t,*}, \Sigma_k^{t,*}, \pi_k^{t,*})}{c(x_{n+1})} p(y_{n+1}|x_{n+1}, \beta_k^{t,*}) + \frac{\alpha p(x_{n+1}|\mu_0^t, \Sigma_0^t, \pi_0^t)}{c(x_{n+1})} p(y_{n+1}|x_{n+1}, \beta_0^t)$$

where $c(x_{n+1}) = \sum_{k=1}^{K} p(x_{n+1}|\mu_k^{t,*}, \Sigma_k^{t,*}, \pi_k^{t,*}) + \alpha p(x_{n+1}|\mu_0^t, \Sigma_0^t, \pi_0^t)$ and μ_0^t, Σ_0^t, π_0^t are drawn from their respective base distributions at iteration t. These expressions show that the difference between these two types of forecasts can only be attributed to the weights assigned to each cluster. This implies that a smaller number of clusters may be preferred under the conditional approach, because it merely results in a more optimal weight for $p(y_{n+1}|x_{n+1}, \beta_0^t)$.

Due to the size of the model, it is not feasible to showcase the parameter estimates of all models. For this reason, only the posterior results of the β coefficient of DPM model X for the 2012 election, which is the best performing DPM model under the conditional approach, and the results of the best performing benchmark model, Bayes 2, are elaborated upon. Figure 3 shows the posterior distributions of selected β parameters. Here solid lines denote the results of the DPM model, while dashed lines denote the results of Bayes 2. The different colours indicate different party choices, where red = VVD, blue = PvdA, orange = PVV, green = CDA and black = Other. In appendix E, the density plots of all β can be found.

From the plots we can conclude that the posterior densities of β under the DPM and benchmark model have similar locations. Overall, the DPM model maintains the same relative results among party choice alternatives as the Bayes 2 model. The main difference is that the posterior densities of the DPM model are more dispersed compared to the benchmark model. These observations can be seen in, for instance, plots 3i and 3viii. Another difference is that the β parameters of the benchmark model tend to be more peaked close to location 0. This explains the tendency of this model to predict a higher No Vote share compared to the DPM models as the β coefficients of the No Vote option are set to zero. Furthermore, the posterior densities of the DPM model tend to be more skewed (see Figures 3i and 3iv, are evidently multimodal for some of the coefficients (see Figures 3ii, 3iii and 3vii) and may exhibit fat tails (see Figures 3iv and 3v).



(i) Intercept **(ii)** Logarithm of Age **(iii)** Female

(iv) Moderate/Slight Urban **(v)** Education = HBO **(vi)** Primary Occupation = Employed

(vii) Religion = Dutch Reformed **(viii)** Party Choice in 2010 = PvdA

Figure 3: Posterior density plots of several β coefficients under DPM model X and benchmark model Bayes 2. A solid line indicates draws of model X, while the dashed line corresponds to draws from Bayes 2. The different colours denote different party choices, where red = VVD, blue = PvdA, orange = PVV, green = CDA and black = Other.

Due to its flexibility, the DPM model picks up on patters that the ordinary Bayesian model does not pick up on. For instance, the variable Logarithm of Age seems to have little to no effect on voting preferences when viewed from the Bayes 2 results. However, Figure 3ii shows that its posterior density for PvdA is bimodal under the DPM model. There seem to be two subpopulations which exhibit two different relationships between age and the preference for voting on PvdA. The same pattern can be seen in Figure 3iii for the Female coefficient of VVD. The benchmark model disregards this and smooths over the two modes, which makes it seem like these variables have no effect at all. In Figure 3vii, the negligence of the bimodality of the Religion = Dutch Reformed coefficient under Bayes 2

underestimates its effect on the preference for CDA for one of the subpopulations, while overestimating it for the other.

Another example of the difference between the two models can be observed from the Moderate/Slightly Urban coefficient for PvdA. The DPM model estimates a posterior density that is right-skewed with a fat right tail, while the benchmark model shows a posterior density that is fairly symmetric with a location near zero.

11 Conclusion

The previous sections have attempted to shed light on current issues in election forecasting, formulate a new model that can be a potential solution to some of the problems and put this model to a test. To this end, a DPM model is proposed and applied on individual-level data. The rationale behind this strategy is as follows. First, individual-level data are a great source of information for election forecasting as they can capture many aspects desired in election forecasting. Examples of these are voting intentions used by aggregate models, sociodemographic characteristics that are linked to voting preferences and even more structural variables pertaining to political preferences or economic indicators, which are data sources usually explored by fundamental models. On top of this, the usage of voting intentions on an individual level allows for a more general model specification, which expands the track record of the model. The latter is useful for model selection.

Secondly, due to its Bayesian nature, the proposed model facilitates a natural way in which past election information can be included to improve on present election forecasts. While doing so, it is not necessary to focus on all (or a large part of) past information in the way fundamental models are necessitated to do. The cross-sectional source of data allows for enough information for an election to be forecasted using the present data alone. Rather, it is possible to include past information that is relevant only.

Lastly, the non-parametric characteristic of the DPM model reduces the problem of model selection. While each election may be influenced with past information, the model is still flexible enough to adapt itself to the relationship of the present election, regardless of whether this is similar to the past relationship or not. This reduces the problem of idiosyncrasy to some extent.

The results show that the DPM model has great potential to improve on the election forecasts of the benchmark models. Under the right prior settings, the DPM model outperforms the other models. These prior settings are not hard to establish. Models that are formulated in a moderately flexible way perform well. The hyperparameter specification of the DPM model proposed here, makes this task relatively easy. Under this DPM approach, learning from previous years is indeed possible and proves to be beneficial for the 2012 election. A comparison of posterior parameter estimates indicates that the DPM model can pick up on certain patters, such as asymmetry, fat tails and multimodality, that the benchmark models cannot pick up on. Lastly, even though the DPM model has many parameters due to its clustering feature, there is no indication of consistent overfitting.

One of the challenges of this approach, however, is finding a sample that is representative of the population. The data at hand showed large deviations in the final election outcome from the population election outcome. For each year, there was a consistent bias in selected parties for all forecasts and, for both years, forecasting the share of the population that did not vote proved to be difficult. Although, the share of the population that does not vote is strictly speaking not part of the election outcome as elections are held to determine party shares only, this is still a problem. This is because errors in

forecasting the No Vote option carry over to errors in forecasting the shares of all other parties.

Next to this, the implementation of the DPM model is subject to practical difficulties. Due to the large dimension of the problem and the sequential sampling made necessary by the Gibbs sampler, computation time is substantially higher than rival methods. As a result, proper sensitivity analysis is very time consuming.

Evidently, the analysis in this writing is subject to many limitations. In terms of model specification, the simplification of the Age variable as continuous variable makes this part of the model not accurate to the characteristics of the data. Furthermore, the model is largely dependent on categorical covariates, which are, due to the lack of good alternative for the nature of this data, specified under independent Categorical distributions with a Dirichlet prior. Although, for example, it has proven to be beneficial to alter the scale of the continuous covariate distributions, this specification is too restricted to allow for this. This gives the forecaster less control over the relative influence of the categorical covariates compared to the remaining data, which is especially important as it dictates clustering in the DPM model. Another serious drawback is related to the manner in which the forecasts are obtained. While generating the election outcome forecasts the share of DNK/PNTS is discarded under the assumption that this does not change the election outcome. This is not likely to be true.

Despite its caveats, the DPM model shows potential to be a good addition to the broad range of election forecasting models. It offers a solution to some of the difficulties faced by commonly used models and may be a good starting point for synthesizers or forecasts combinations in general. The latter touches upon a subject that can be a topic for any future research. In fact, the DPM model and its findings offer a great variety of extensions, such as the possibility to take advantage of the panel structure, to use other types of variables besides sociodemographic characteristics and to apply the model on many other (types of) elections. Lastly, although it would be interesting to perform a comparison on the forecasting performance of the DPM model and the many other types of models discussed in Section 2, this task was beyond the scope of this research. As most of these models are defined for the US presidential elections, perhaps the DPM model could be applied using US individual-level data to make this comparison possible.

12 Acknowledgments

In this paper, data of the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands) are used.

A MCMC Sampler

In the subsequent sections the derivation of the posterior conditional distributions of the sampled parameters θ_i^x and hyperparameters λ is shown.

A.1 Conditional Posterior θ_i^x

Dropping the irrelevant terms from equation (30) using (31) and (29) and some simple algebra give the following results.

Conditional posterior μ_k^*

$$
\begin{aligned}
& p(\mu_k^*|z, x, \Sigma_\mu^*, \Sigma_k^*, \lambda_\mu) \\
\propto\ & N(\mu_k^*|\mu_\mu, \Sigma_\mu^* \prod_{z_i=k} N(x_i^{con}|\mu_k^*, \Sigma_k^*) \\
\propto\ & \exp(-\frac{1}{2}(\mu_k^* - \mu_\mu)^T \Sigma_\mu^{*^{-1}} (\mu_k^* - \mu_\mu)) \prod_{z_i=k} \exp(-\frac{1}{2}(x_i^{con} - \mu_k^*)^T \Sigma_k^{*^{-1}} (x_i^{con} - \mu_k^*)) \\
\propto\ & \exp(-\frac{1}{2}(\mu_k^{*T} \Sigma_\mu^{*^{-1}} \mu_k^* - 2\mu_k^{*T} \Sigma_\mu^{*^{-1}} \mu_\mu)) \exp(-\frac{1}{2}(n_k \mu_k^{*T} \Sigma_k^{*^{-1}} \mu_k^* - 2\mu_k^{*T} \Sigma_k^{*^{-1}} n_k \bar{x}^{con})) \\
\propto\ & \exp(-\frac{1}{2}(\mu_k^{*T} (\Sigma_\mu^{*^{-1}} + n_k \Sigma_k^{*^{-1}}) \mu_k^* - 2\mu_k^{*T} (\Sigma_\mu^{*^{-1}} \mu_\mu + n_k \Sigma_k^{*^{-1}} \bar{x}^{con}))) \\
\propto\ & \exp(-\frac{1}{2}(\mu_k^{*T} (\Sigma_\mu^{*^{-1}} + n_k \Sigma_k^{*^{-1}}) \mu_k^* - 2\mu_k^{*T} (\Sigma_\mu^{*^{-1}} + n_k \Sigma_k^{*^{-1}}) (\Sigma_\mu^{*^{-1}} + n_k \Sigma_k^{*^{-1}})^{-1} (\Sigma_\mu^{*^{-1}} \mu_\mu + n_k \Sigma_k^{*^{-1}} \bar{x}^{con})))
\end{aligned}
$$

where n_k is the number of $z_i : z_i = k$ and $\bar{x}^{con} = \frac{1}{n_k} \sum_{z_i=k} x_i$. The last line is the kernel of $N((\Sigma_\mu^{*^{-1}} + n_k \Sigma_k^{*^{-1}})^{-1} (\Sigma_\mu^{*^{-1}} \mu_\mu + n_k \Sigma_k^{*^{-1}} \bar{x}^{con}), (\Sigma_\mu^{*^{-1}} + n_k \Sigma_k^{*^{-1}})^{-1})$.

Conditional posterior Σ_k^*

$$
\begin{aligned}
& p(\Sigma_k^*|z, x, \lambda_\Sigma, \lambda_\mu) \\
\propto\ & IW(\Sigma_k^*|\nu_\Sigma, \nu_\Sigma \upsilon_\Sigma I_{L^{con}}) \prod_{z_i=k} N(x_i^{con}|\mu_k^*, \Sigma_k^*) \\
\propto\ & |\Sigma_k^*|^{-\frac{n_k + \nu_\Sigma + L^{con} + 1}{2}} \exp(-\frac{1}{2} tr\{\nu_\Sigma \upsilon_\Sigma I_{L^{con}} \Sigma_k^{*^{-1}}\}) \exp(-\frac{1}{2} tr\{\sum_{z_i=k} (x_i^{con} - \mu_k^*)^T \Sigma_k^{*^{-1}} (x_i^{con} - \mu_k^*)\}) \\
\propto\ & |\Sigma_k^*|^{-\frac{n_k + \nu_\Sigma + L^{con} + 1}{2}} \exp(-\frac{1}{2} tr\{\nu_\Sigma \upsilon_\Sigma I_{L^{con}} \Sigma_k^{*^{-1}}\}) \exp(-\frac{1}{2} tr\{\sum_{z_i=k} (x_i^{con} - \mu_k^*)(x_i^{con} - \mu_k^*)^T \Sigma_k^{*^{-1}}\}) \\
\propto\ & |\Sigma_k^*|^{-\frac{n_k + \nu_\Sigma + L^{con} + 1}{2}} \exp(-\frac{1}{2} tr\{(\nu_\Sigma \upsilon_\Sigma I_{L^{con}} + \sum_{z_i=k} (x_i^{con} - \mu_k^*)(x_i^{con} - \mu_k^*)^T) \Sigma_k^{*^{-1}}\})
\end{aligned}
$$

The last line is the kernel of a $IW(n_k + \nu_\Sigma, \nu_\Sigma \upsilon_\Sigma I_{L^{con}} + \sum_{z_i=k} (x_i^{con} - \mu_k^*)(x_i^{con} - \mu_k^*)^T)$ distribution.

Conditional posterior π_k^*

Lastly, as $p(\pi_{l,k}^*|\pi_{-l,k}^*, z, x, \lambda_{\pi_l}) = p(\pi_{l,k}^*|z, x, \lambda_{\pi_l})$ the conditional posterior of $\pi_{l,k}^*$ is

$$
\begin{aligned}
p(\pi^*_{l,k}|z,x,\lambda_{\pi_l}) &\propto Dir(\pi^*_{l,k}|\frac{a_{0,l}}{M_l},\cdots,\frac{a_{0,l}}{M_l})\prod_{z_i=k} Cat(x^{\text{cat}}_{i,l}|\pi^*_{l,k}) \\
&\propto \prod_{m=1}^{M_l} {\pi^*_{l,k,m}}^{\frac{a_{0,l}}{M_l}-1} \prod_{z_i=k}\prod_{m=1}^{M_l} {\pi^*_{l,k,m}}^{I(x^{\text{cat}}_{i,l}=m)} \\
&\propto \prod_{m=1}^{M_l} {\pi^*_{l,k,m}}^{\frac{a_{0,l}}{M_l}+\sum_{z_i=k} I(x^{\text{cat}}_{i,l}=m)-1}
\end{aligned}
$$

Hence, $\pi_{l,k}$ has the posterior conditional distribution $Dir(\frac{a_{0,l}}{M_l}+\sum_{z_i=k} I(x^{\text{cat}}_{i,l}=1),\cdots,\frac{a_{0,l}}{M_l}+\sum_{z_i=k} I(x^{\text{cat}}_{i,l}=M_l))$.

A.2 Conditional Posterior $(\lambda_\mu, \lambda_\Sigma)$

$$
\begin{aligned}
p(\mu_\mu,\upsilon_\mu,\upsilon_\Sigma|\{\mu^*_k,\Sigma^*_k\}_{k=1}^K,\nu_\mu,\nu_\Sigma) &\propto I(\mu_\mu)I(\upsilon_\mu)I(\upsilon_\Sigma)IW(\Sigma_\mu)\prod_{k=1}^K p(\mu^*_k,\Sigma^*_k|\mu_\mu,\nu_\mu,\upsilon\mu,\nu_\Sigma,\upsilon_\Sigma) \\
&\propto I(\mu_\mu)I(\upsilon_\Sigma)I(\upsilon_\Sigma)IW(\Sigma_\mu)\prod_{k=1}^K N(\mu^*_k|\mu_\mu,\Sigma_\mu)IW(\Sigma^*_k|\nu_\Sigma,\upsilon_\Sigma)
\end{aligned}
$$

Conditional Posterior μ_μ

$$
\begin{aligned}
p(\mu_\mu|\{\mu^*_k,\Sigma_\mu\}_{k=1}^K) &\propto I(\mu_\mu)\prod_{k=1}^K N(\mu^*_k|\mu_\mu,\Sigma_\mu) \\
&\propto I(\mu_\mu)\exp(-\frac{1}{2}(\sum_{k=1}^K(\mu^*_k-\mu_\mu)^T\Sigma_\mu^{-1}(\mu^*_k-\mu_\mu))) \\
&\propto I(\mu_\mu)\exp(-\frac{1}{2}(\mu_\mu^T(K\Sigma_\mu^{-1})\mu_\mu-2\mu_\mu^T(\sum_{k=1}^K\Sigma_\mu^{-1}\mu^*_k))) \\
\mu_\mu|c,\{\mu^*_k,\Sigma^*_k\}_{k=1}^K &\sim I(\mu_\mu)N(K^{-1}\Sigma_\mu(\sum_{k=1}^K\Sigma_\mu^{-1}\mu^*_k),K^{-1}\Sigma_\mu))
\end{aligned}
$$

where $I(\mu_\mu)$ is 1 when $\mu_{\mu,l}\in[\mu^{lo}_{\mu,l},\mu^{up}_{\mu,l}]$ for all $l\in\{1,\cdots,L^{Con}\}$ and zero otherwise.

Conditional Posterior Σ_μ

$$
\begin{aligned}
&p(\Sigma_\mu|\mu_\mu, \upsilon_\mu, \{\mu_k^*\}_{k=1}^K) \\
&\propto IW(\Sigma_\mu|\nu_\mu, \nu_\mu \upsilon_\mu I_{L^{con}}) \prod_{k=1}^K N(\mu_k^*|\mu_\mu, \Sigma_\mu) \\
&\propto |\Sigma_\mu|^{-\frac{\nu_\mu + L^{con}+1}{2}} \exp(-\frac{1}{2} tr\{\nu_\mu \upsilon_\mu I_{L^{con}} \Sigma_\mu^{-1}\}) |\Sigma_\mu|^{-\frac{K}{2}} \prod_{k=1}^K \exp(-\frac{1}{2} tr\{(\mu_k^* - \mu_\mu)^T \Sigma_\mu^{-1} (\mu_k^* - \mu_\mu)\}) \\
&\propto |\Sigma_\mu|^{-\frac{\nu_\mu + K + L^{con}+1}{2}} \exp(-\frac{1}{2} tr\{(\nu_\mu \upsilon_\mu I_{L^{con}} + \sum_{k=1}^K (\mu_k^* - \mu_\mu)(\mu_k^* - \mu_\mu)^T) \Sigma_\mu^{-1}\}) \\
&\Sigma_\mu|\mu_\mu, \upsilon_\mu, \{\mu_k^*\}_{k=1}^K) \sim IW(\nu_\mu + K, \nu_\mu \upsilon_\mu I_{L^{con}} + \sum_{k=1}^K (\mu_k^* - \mu_\mu)(\mu_k^* - \mu_\mu)^T)
\end{aligned}
$$

Conditional Posterior υ_μ

$$
\begin{aligned}
p(\upsilon_\mu|\mu_\mu, \Sigma_\mu, \{\mu_k^*\}_{k=1}^K) &\propto I(\upsilon_\mu) IW(\Sigma_\mu|\upsilon_\mu, \nu_\mu) \\
&\propto I(\upsilon_\mu) |\upsilon_\mu \nu_\mu I_{L^{con}}|^{\frac{\nu_\mu}{2}} \exp(-\frac{1}{2} tr\{\upsilon_\mu \nu_\mu I_{L^{con}} \Sigma_\mu^{-1}\}) \\
&\propto I(\upsilon_\mu) \upsilon_\mu^{\frac{\nu_\mu L^{con}}{2}} \exp(-\frac{1}{2} tr\{\upsilon_\mu \nu_\mu \Sigma_\mu^{-1}\}) \\
&\propto I(\upsilon_\mu) \upsilon_\mu^{\frac{\nu_\mu L^{con}}{2}} \exp(-\upsilon_\mu (\frac{1}{2} \nu_\mu tr\{\Sigma_\mu^{-1}\})) \\
\upsilon_\mu|\Sigma_\mu, \nu_\mu, \{\mu_k^*\}_{k=1}^K &\sim I(\upsilon_\mu) Ga(\frac{\nu_\mu L^{con}}{2} + 1, \frac{1}{2} \nu_\mu tr(\Sigma_\mu^{-1}))
\end{aligned}
$$

Conditional Posterior υ_Σ

$$
\begin{aligned}
p(\upsilon_\Sigma|\nu_\Sigma, \{\Sigma_k^*\}_{k=1}^K) &\propto I(\upsilon_\Sigma) \prod_{k=1}^K IW(\Sigma_k^*|\nu_\Sigma, \upsilon_\Sigma) \\
&\propto I(\upsilon_\Sigma) \prod_{k=1}^K |\upsilon_\Sigma \nu_\Sigma I_{L^{con}}|^{\nu_\Sigma/2} \exp(-\frac{1}{2} tr(\nu_\Sigma \upsilon_\Sigma I_{L^{con}} \Sigma_k^{*^{-1}}) \\
&\propto I(\upsilon_\Sigma) \upsilon_\Sigma^{\frac{\nu_\Sigma L^{con} K}{2}} \exp(-\frac{1}{2} tr(\nu_\Sigma \upsilon_\Sigma \sum_{k=1}^K \Sigma_k^{*^{-1}})) \\
&\propto I(\upsilon_\Sigma) \upsilon_\Sigma^{\frac{\nu_\Sigma L^{con} K}{2}} \exp(-\upsilon_\Sigma (\frac{1}{2} \nu_\Sigma \sum_{k=1}^K tr(\Sigma_k^{*^{-1}}))) \\
\upsilon_\Sigma|\nu_\Sigma, \{\Sigma_k^*\}_{k=1}^K &\sim I(\upsilon_\Sigma) Ga(\frac{\nu_\Sigma L^{con} K}{2} + 1, \frac{1}{2} \nu_\Sigma \sum_{k=1}^K tr(\Sigma_k^{*^{-1}})))
\end{aligned}
$$

A.3 Conditional Posterior λ_{a_0}

For all $l \in \{1, \cdots, L^{Cat}\}$

Conditional Posterior a_0

$$
\begin{aligned}
p(a_{0,l}|\pi^*_{l,k}) &\propto I(a_{0,l}) \prod_{k=1}^{K} p(\pi^*_{l,k}|a_{0,l}) \\
&\propto I(a_{0,l}) \prod_{k=1}^{K} \prod_{m=1}^{M_l} {\pi^*_{l,k,m}}^{\frac{a_{0,l}}{M_l}-1} \\
&\propto I(a_{0,l}) \prod_{k=1}^{K} \prod_{m=1}^{M_l} \exp(\log({\pi^*_{l,k,m}}^{\frac{a_{0,l}}{M_l}-1})) \\
&\propto I(a_{0,l}) \prod_{k=1}^{K} \prod_{m=1}^{M_l} \exp(\frac{a_{0,l}}{M_l}\log(\pi^*_{l,k,m})) \\
&\propto I(a_{0,l}) \exp(-a_{0,l}(\frac{1}{M_l}\sum_{k=1}^{K}\sum_{m=1}^{M_l} -\log(\pi^*_{l,k,m}))) \\
p(a_{0,l}|\pi^*_{l,k}) &\sim I(a_{0,l}) Exp(\frac{1}{M_l}\sum_{k=1}^{K}\sum_{m=1}^{M_l} -\log(\pi^*_{l,k,m}))
\end{aligned}
$$

where $I(a_{0,l})$ is 1 when $a_{0,l} \in [a^{lo}_{0,l}, a^{up}_{0,l}]$ and zero otherwise.

A.4 Conditional Posterior λ_β

$$
p(\bar{\beta}, B, \upsilon_\beta | \{\beta^*_k\}_{k=1}^K) \propto I(\bar{\beta}) I(\upsilon_\beta) IW(B|\nu_\beta, \nu_\beta \upsilon_\beta I_{L^*}) \prod_{k=1}^{K} N(\beta^*_k|\bar{\beta}, B)
$$

where $I(\bar{\beta})$ and $I(\upsilon_\beta)$ are 1 when their respective argument is in the region specified in (27) and zero otherwise.

Conditional Posterior B

$$
\begin{aligned}
& p(B|\bar{\beta}, \upsilon_\beta, \{\beta^*_k\}_{k=1}^K) \\
\propto\ & IW(B|\nu_\beta, \nu_\beta \upsilon_\beta I_{L^*}) \prod_{k=1}^{K} N(\beta^*_k|\bar{\beta}, B) \\
\propto\ & |B|^{-\frac{\nu_\beta + L^* + 1}{2}} \exp(-\frac{1}{2} tr\{\nu_\beta \upsilon_\beta I_{L^*} B^{-1}\}) |B|^{-\frac{K}{2}} \prod_{k=1}^{K} \exp(-\frac{1}{2} tr\{(\beta^*_k - \bar{\beta})^T B^{-1} (\beta^*_k - \bar{\beta})\}) \\
\propto\ & |B|^{-\frac{\nu_\beta + K + L^* + 1}{2}} \exp(-\frac{1}{2} tr\{(\nu_\beta \upsilon_\beta I_{L^*} + \sum_{k=1}^{K} (\beta^*_k - \bar{\beta})(\beta^*_k - \bar{\beta})^T) B^{-1}\}) \\
& B|\bar{\beta}, \upsilon_\beta, \{\beta^*_k\}_{k=1}^K) \sim IW(\nu_\beta + K, \nu_\beta \upsilon_\beta I_{L^*} + \sum_{k=1}^{K} (\beta^*_k - \bar{\beta})(\beta^*_k - \bar{\beta})^T)
\end{aligned}
$$

Conditional Posterior $\bar{\beta}$

$$
\begin{aligned}
p(\bar{\beta}|B,\{\beta_k^*\}_{k=1}^K) &\propto I(\bar{\beta})\prod_{k=1}^K N(\beta_k^*|\bar{\beta},B) \\
&\propto I(\bar{\beta})\exp(-\frac{1}{2}(\sum_{k=1}^K(\beta_k^*-\bar{\beta})^T B^{-1}(\beta_k^*-\bar{\beta}))) \\
&\propto I(\bar{\beta})\exp(-\frac{1}{2}(\bar{\beta}^T K B^{-1}\bar{\beta}-2\bar{\beta}^T B^{-1}(\sum_{k=1}^K\beta_k^*))) \\
\bar{\beta}|B,\{\beta_k^*\}_{k=1}^K &\sim I(\bar{\beta})N(\frac{1}{K}\sum_{k=1}^K\beta_k^*, K^{-1}B)
\end{aligned}
$$

Conditional Posterior υ_β

$$
\begin{aligned}
p(\upsilon_\beta|\bar{\beta},B,\{\beta_k^*\}_{k=1}^K) &\propto I(\upsilon_\beta)IW(B|\upsilon_\beta,\nu_\beta) \\
&\propto I(\upsilon_\beta)|\upsilon_\beta\nu_\beta I_{L^*}|^{\frac{\nu_\beta}{2}}\exp(-\frac{1}{2}tr\{\upsilon_\beta\nu_\beta I_{L^*}B^{-1}\}) \\
&\propto I(\upsilon_\beta)\upsilon_\beta^{\frac{\nu_\beta L^*}{2}}\exp(-\frac{1}{2}tr\{\upsilon_\beta\nu_\beta B^{-1}\}) \\
&\propto I(\upsilon_\beta)\upsilon_\beta^{\frac{\nu_\beta L^*}{2}}\exp(-\upsilon_\beta(\frac{1}{2}\nu_\beta tr\{B^{-1}\})) \\
\upsilon_\beta|B,\nu_\beta,\{\beta_k^*\}_{k=1}^K &\sim I(\upsilon_\beta)Ga(\frac{\nu_\beta L^*}{2}+1,\frac{1}{2}\nu_\beta tr(B^{-1}))
\end{aligned}
$$

B Table prior probabilities of the number of clusters

Table 10: Prior probabilities of number of clusters K for different values of α.

	α					
K	1	2	3	4	5	10
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.002	0.000	0.000	0.000	0.000	0.000
3	0.009	0.000	0.000	0.000	0.000	0.000
4	0.026	0.000	0.000	0.000	0.000	0.000
5	0.052	0.000	0.000	0.000	0.000	0.000
6	0.086	0.001	0.000	0.000	0.000	0.000
7	0.117	0.002	0.000	0.000	0.000	0.000
8	0.136	0.005	0.000	0.000	0.000	0.000
9	0.139	0.010	0.000	0.000	0.000	0.000
10	0.127	0.018	0.000	0.000	0.000	0.000
11	0.104	0.030	0.000	0.000	0.000	0.000
12	0.077	0.044	0.001	0.000	0.000	0.000
13	0.052	0.060	0.002	0.000	0.000	0.000
14	0.033	0.076	0.004	0.000	0.000	0.000
15	0.019	0.089	0.007	0.000	0.000	0.000
16	0.011	0.097	0.012	0.000	0.000	0.000
17	0.005	0.099	0.018	0.001	0.000	0.000
18	0.003	0.095	0.026	0.001	0.000	0.000
19	0.001	0.086	0.036	0.002	0.000	0.000
20	0.001	0.074	0.046	0.003	0.000	0.000
21	0.000	0.061	0.057	0.005	0.000	0.000
22	0.000	0.047	0.066	0.008	0.000	0.000
23	0.000	0.035	0.074	0.012	0.000	0.000
24	0.000	0.025	0.079	0.017	0.001	0.000
25	0.000	0.017	0.081	0.023	0.001	0.000
26	0.000	0.011	0.079	0.030	0.002	0.000
27	0.000	0.007	0.075	0.037	0.003	0.000
28	0.000	0.004	0.068	0.045	0.005	0.000
29	0.000	0.002	0.059	0.053	0.008	0.000
30	0.000	0.001	0.050	0.059	0.011	0.000
31	0.000	0.001	0.041	0.065	0.015	0.000
32	0.000	0.000	0.033	0.068	0.019	0.000
33	0.000	0.000	0.025	0.070	0.025	0.000
34	0.000	0.000	0.019	0.069	0.031	0.000
35	0.000	0.000	0.013	0.067	0.037	0.000
36	0.000	0.000	0.009	0.062	0.043	0.000
37	0.000	0.000	0.006	0.057	0.049	0.000
38	0.000	0.000	0.004	0.050	0.054	0.000
39	0.000	0.000	0.003	0.043	0.058	0.000
40	0.000	0.000	0.002	0.036	0.061	0.000

C Results: Election Forecasts

Table 11: Forecasts of the election outcome for the 2010 training data.

	VVD	PVDA	PVV	CDA	Other	No Vote	VVD	PVDA	PVV	CDA	Other	No Vote
Actual party shares	0.155	0.148	0.116	0.103	0.233	0.246	0.155	0.148	0.116	0.103	0.233	0.246
Sample party shares	0.177	0.172	0.101	0.116	0.295	0.139	0.177	0.172	0.101	0.116	0.295	0.139
VI	0.103	0.097	0.147	0.149	0.386	0.118	0.103	0.097	0.147	0.149	0.386	0.118
ML	0.103	0.098	0.147	0.149	0.386	0.118	0.103	0.098	0.147	0.149	0.386	0.118
Bayes 1	0.104	0.102	0.148	0.149	0.376	0.120	0.104	0.102	0.148	0.149	0.376	0.120
Bayes 2	0.107	0.101	0.149	0.149	0.374	0.120	0.107	0.101	0.149	0.149	0.374	0.120
	Based on joint distribution						**Based on conditional distribution**					
A	0.113	0.108	0.157	0.150	0.362	0.111	0.109	0.112	0.160	0.136	0.385	0.098
B	0.113	0.108	0.156	0.151	0.360	0.112	0.128	0.119	0.179	0.154	0.347	0.073
C	0.117	0.111	0.154	0.153	0.349	0.116	0.147	0.146	0.194	0.166	0.282	0.066
D	0.119	0.111	0.155	0.152	0.347	0.117	0.155	0.152	0.175	0.163	0.288	0.066
E	0.116	0.111	0.153	0.152	0.348	0.120	0.162	0.151	0.171	0.165	0.286	0.066
F	0.116	0.110	0.154	0.152	0.350	0.117	0.159	0.144	0.176	0.165	0.288	0.069
G	0.116	0.112	0.155	0.153	0.348	0.117	0.160	0.142	0.174	0.167	0.289	0.068
H	0.117	0.113	0.155	0.153	0.348	0.115	0.165	0.152	0.176	0.160	0.282	0.066
I	0.114	0.107	0.154	0.151	0.356	0.118	0.148	0.137	0.169	0.157	0.316	0.075
J	0.117	0.112	0.154	0.153	0.349	0.114	0.169	0.155	0.176	0.161	0.282	0.057
K	0.113	0.109	0.152	0.151	0.356	0.118	0.142	0.149	0.162	0.155	0.316	0.076
L	0.116	0.112	0.155	0.153	0.348	0.117	0.161	0.146	0.173	0.161	0.288	0.071
M	0.115	0.109	0.152	0.153	0.354	0.118	0.137	0.145	0.163	0.152	0.324	0.080
N	0.115	0.109	0.153	0.151	0.354	0.119	0.140	0.139	0.163	0.159	0.323	0.077
O	0.114	0.109	0.153	0.152	0.355	0.119	0.137	0.140	0.165	0.150	0.329	0.080
P	0.115	0.109	0.154	0.151	0.353	0.119	0.137	0.153	0.169	0.156	0.312	0.073
Q	0.115	0.109	0.154	0.152	0.353	0.118	0.142	0.145	0.164	0.156	0.317	0.076
R	0.116	0.109	0.155	0.151	0.355	0.115	0.137	0.143	0.167	0.152	0.315	0.087
S	0.113	0.109	0.154	0.152	0.356	0.117	0.159	0.159	0.178	0.162	0.286	0.057

The benchmark forecasts are obtained as follows: VI = Benchmark forecast obtained by aggregating voting intentions, ML = Maximum likelihood model, Bayes 1 = Bayesian model using prior $N(0, B)$ with $B = 1.5I_{L^*}$, Bayes 2 = Bayesian model using prior $N(0, B)$ with $B = 100I_{L^*}$.

Table 12: Forecasts of the election outcome for the 2012 training data.

	VVD	PVDA	PVV	CDA	Other	No Vote	VVD	PVDA	PVV	CDA	Other	No Vote
Actual party shares	0.198	0.185	0.075	0.063	0.224	0.254	0.198	0.185	0.075	0.063	0.224	0.254
Sample party shares	0.213	0.227	0.070	0.082	0.278	0.132	0.213	0.227	0.070	0.082	0.278	0.132
VI	0.209	0.103	0.100	0.079	0.389	0.120	0.209	0.103	0.100	0.079	0.389	0.120
ML	0.209	0.103	0.100	0.079	0.389	0.120	0.209	0.103	0.100	0.079	0.389	0.120
Bayes 1	0.208	0.106	0.105	0.081	0.380	0.120	0.208	0.106	0.105	0.081	0.380	0.120
Bayes 2	0.207	0.107	0.104	0.083	0.377	0.122	0.207	0.107	0.104	0.083	0.377	0.122
	Based on joint distribution						**Based on conditional distribution**					
A	0.211	0.111	0.110	0.083	0.372	0.114	0.179	0.177	0.116	0.089	0.353	0.087
B	0.211	0.112	0.110	0.083	0.371	0.114	0.181	0.180	0.124	0.096	0.340	0.079
C	0.208	0.114	0.115	0.090	0.355	0.118	0.182	0.191	0.140	0.137	0.279	0.072
D	0.207	0.114	0.115	0.090	0.356	0.118	0.170	0.204	0.147	0.149	0.260	0.071
E	0.209	0.113	0.115	0.091	0.354	0.118	0.184	0.197	0.147	0.150	0.257	0.066
F	0.208	0.114	0.114	0.091	0.356	0.118	0.181	0.196	0.139	0.141	0.274	0.070
G	0.209	0.113	0.113	0.089	0.358	0.118	0.187	0.202	0.139	0.138	0.262	0.072
H	0.208	0.112	0.116	0.090	0.355	0.119	0.178	0.203	0.140	0.136	0.275	0.068
I	0.209	0.111	0.111	0.087	0.364	0.118	0.183	0.204	0.120	0.131	0.263	0.099
J	0.210	0.114	0.113	0.090	0.356	0.117	0.189	0.198	0.145	0.139	0.263	0.066
K	0.208	0.112	0.110	0.089	0.363	0.119	0.174	0.196	0.111	0.134	0.275	0.111
L	0.208	0.113	0.115	0.090	0.357	0.118	0.184	0.194	0.144	0.138	0.269	0.071
M	0.209	0.113	0.112	0.088	0.360	0.118	0.176	0.207	0.126	0.128	0.268	0.095
N	0.209	0.113	0.112	0.088	0.359	0.118	0.173	0.199	0.128	0.134	0.271	0.095
O	0.209	0.110	0.112	0.087	0.362	0.119	0.175	0.205	0.116	0.134	0.270	0.102
P	0.210	0.112	0.112	0.088	0.359	0.118	0.176	0.200	0.127	0.138	0.271	0.089
Q	0.209	0.112	0.112	0.088	0.361	0.118	0.179	0.210	0.125	0.127	0.266	0.094
R	0.210	0.112	0.112	0.088	0.362	0.117	0.175	0.200	0.130	0.140	0.264	0.092
S	0.209	0.112	0.111	0.089	0.361	0.118	0.183	0.204	0.143	0.142	0.259	0.069
T	0.209	0.110	0.110	0.087	0.364	0.120	0.175	0.199	0.116	0.125	0.275	0.110
U	0.209	0.112	0.112	0.088	0.360	0.119	0.173	0.195	0.122	0.134	0.274	0.102
V	0.208	0.111	0.111	0.086	0.365	0.119	0.175	0.190	0.116	0.123	0.285	0.111
W	0.208	0.111	0.111	0.087	0.365	0.119	0.179	0.191	0.105	0.124	0.288	0.113
X	0.208	0.111	0.110	0.087	0.365	0.119	0.179	0.198	0.110	0.124	0.277	0.111
Y	0.208	0.111	0.111	0.087	0.364	0.120	0.168	0.188	0.114	0.128	0.287	0.114
Z	0.208	0.110	0.111	0.087	0.366	0.118	0.174	0.188	0.119	0.131	0.277	0.111

The benchmark forecasts are obtained as follows: VI = Benchmark forecast obtained by aggregating voting intentions, ML = Maximum likelihood model, Bayes 1 = Bayesian model using prior $N(0, B)$ with $B = 1.5I_{L^*}$, Bayes 2 = Bayesian model using prior $N(0, B)$ with $B = 100I_{L^*}$.

Table 13: Forecasts of the election outcome for the 2010 test data.

	VVD	PVDA	PVV	CDA	Other	No Vote	VVD	PVDA	PVV	CDA	Other	No Vote
Actual party shares 2010	0.155	0.148	0.116	0.103	0.233	0.246	0.155	0.148	0.116	0.103	0.233	0.246
Party shares in data set	0.184	0.171	0.112	0.129	0.280	0.124	0.184	0.171	0.112	0.129	0.280	0.124
VI	0.111	0.108	0.138	0.141	0.390	0.111	0.111	0.108	0.138	0.141	0.390	0.111
ML	0.107	0.100	0.148	0.149	0.388	0.109	0.107	0.100	0.148	0.149	0.388	0.109
Bayes 1	0.108	0.104	0.149	0.149	0.378	0.113	0.108	0.104	0.149	0.149	0.378	0.113
Bayes 2	0.111	0.103	0.150	0.149	0.375	0.112	0.111	0.103	0.150	0.149	0.375	0.112
	Based on joint distribution						**Based on conditional distribution**					
A	0.116	0.111	0.159	0.151	0.361	0.102	0.111	0.112	0.161	0.136	0.387	0.093
B	0.116	0.111	0.159	0.150	0.361	0.102	0.129	0.121	0.180	0.155	0.346	0.070
C	0.122	0.115	0.158	0.154	0.347	0.105	0.149	0.146	0.194	0.166	0.282	0.063
D	0.123	0.115	0.158	0.154	0.344	0.106	0.157	0.152	0.176	0.163	0.289	0.063
E	0.120	0.116	0.157	0.153	0.344	0.110	0.163	0.151	0.172	0.165	0.285	0.063
F	0.121	0.116	0.158	0.154	0.346	0.106	0.160	0.144	0.178	0.164	0.287	0.066
G	0.119	0.117	0.159	0.154	0.346	0.106	0.161	0.141	0.176	0.167	0.290	0.065
H	0.122	0.116	0.158	0.154	0.346	0.104	0.165	0.153	0.178	0.161	0.281	0.063
I	0.118	0.109	0.157	0.152	0.353	0.110	0.149	0.137	0.169	0.157	0.317	0.072
J	0.123	0.115	0.158	0.154	0.347	0.104	0.170	0.155	0.177	0.161	0.283	0.055
K	0.118	0.112	0.155	0.151	0.354	0.110	0.143	0.149	0.163	0.155	0.318	0.072
L	0.120	0.117	0.159	0.154	0.345	0.106	0.162	0.147	0.174	0.160	0.288	0.069
M	0.119	0.112	0.156	0.153	0.351	0.110	0.139	0.145	0.164	0.152	0.325	0.076
N	0.119	0.113	0.156	0.153	0.350	0.110	0.142	0.139	0.164	0.159	0.323	0.073
O	0.118	0.112	0.156	0.152	0.351	0.111	0.137	0.141	0.165	0.151	0.330	0.077
P	0.119	0.113	0.157	0.152	0.350	0.110	0.137	0.152	0.170	0.158	0.313	0.070
Q	0.119	0.112	0.157	0.153	0.350	0.110	0.143	0.145	0.165	0.157	0.319	0.072
R	0.120	0.113	0.158	0.152	0.352	0.105	0.138	0.143	0.167	0.152	0.315	0.084
S	0.118	0.113	0.157	0.152	0.352	0.108	0.160	0.159	0.179	0.162	0.287	0.054

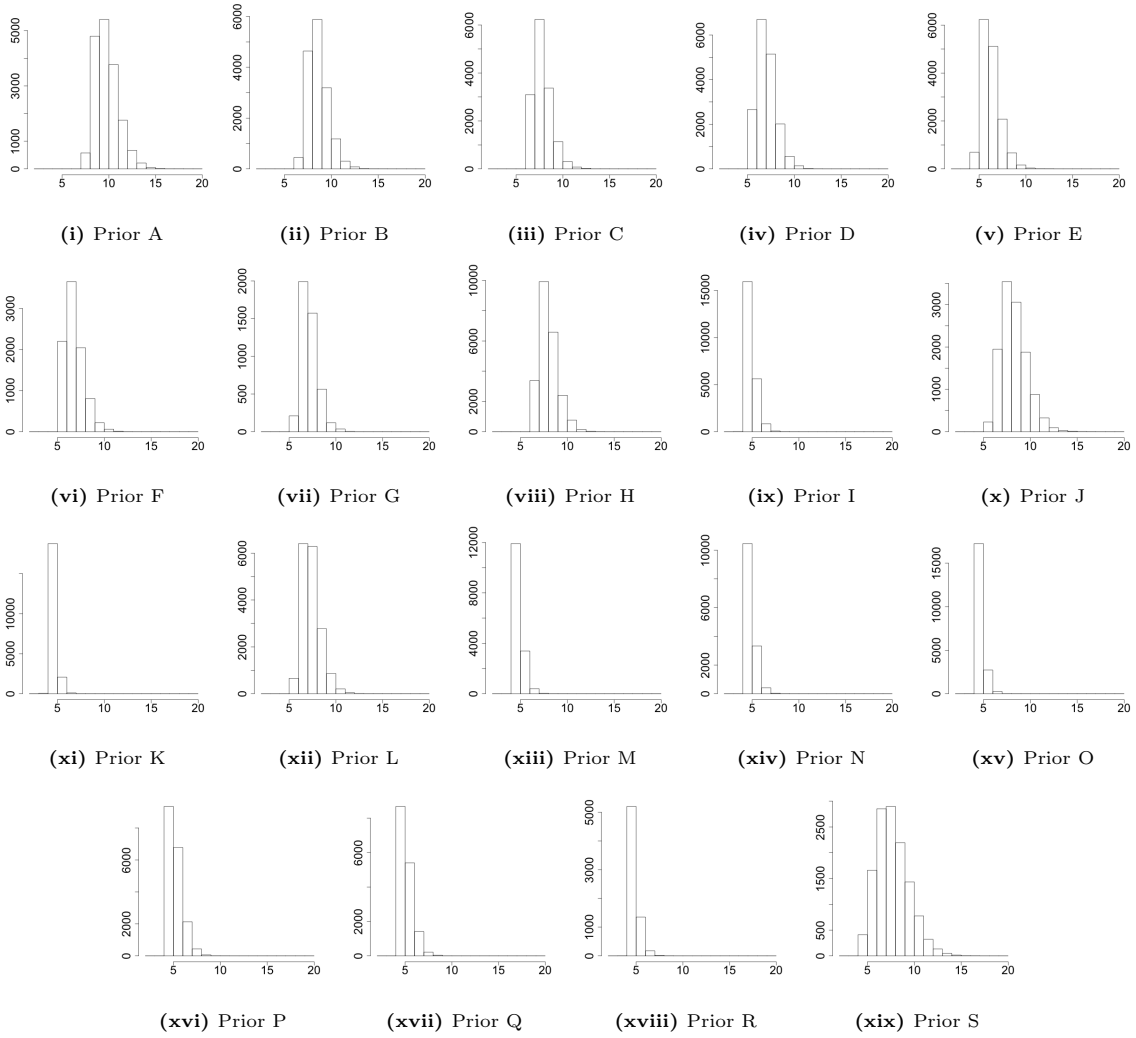
The benchmark forecasts are obtained as follows: VI = Benchmark forecast obtained by aggregating voting intentions, ML = Maximum likelihood model, Bayes 1 = Bayesian model using prior $N(0, B)$ with $B = 1.5I_{L^*}$, Bayes 2 = Bayesian model using prior $N(0, B)$ with $B = 100I_{L^*}$.

Table 14: Forecasts of the election outcome for the 2012 test data.

	VVD	PVDA	PVV	CDA	Other	No Vote	VVD	PVDA	PVV	CDA	Other	No Vote
Actual party shares 2010	0.198	0.185	0.075	0.063	0.224	0.254	0.198	0.185	0.075	0.063	0.224	0.254
Party shares in data set	0.221	0.220	0.061	0.087	0.286	0.124	0.221	0.220	0.061	0.087	0.286	0.124
VI	0.208	0.121	0.099	0.088	0.379	0.106	0.208	0.121	0.099	0.088	0.379	0.106
ML	0.213	0.115	0.099	0.075	0.393	0.105	0.213	0.115	0.099	0.075	0.393	0.105
Bayes 1	0.211	0.117	0.103	0.078	0.384	0.108	0.211	0.117	0.103	0.078	0.384	0.108
Bayes 2	0.210	0.119	0.102	0.079	0.382	0.109	0.210	0.119	0.102	0.079	0.382	0.109
	Based on joint distribution						**Based on conditional distribution**					
A	0.211	0.120	0.110	0.079	0.378	0.103	0.179	0.186	0.114	0.085	0.357	0.079
B	0.211	0.121	0.110	0.079	0.377	0.102	0.182	0.188	0.124	0.092	0.343	0.072
C	0.209	0.125	0.115	0.087	0.356	0.107	0.183	0.197	0.139	0.136	0.281	0.065
D	0.208	0.125	0.115	0.089	0.357	0.107	0.170	0.210	0.147	0.148	0.261	0.064
E	0.212	0.123	0.114	0.090	0.353	0.108	0.184	0.202	0.147	0.150	0.257	0.060
F	0.209	0.124	0.114	0.089	0.357	0.107	0.182	0.202	0.139	0.139	0.276	0.063
G	0.212	0.124	0.113	0.088	0.357	0.106	0.187	0.208	0.139	0.138	0.263	0.065
H	0.207	0.122	0.115	0.089	0.358	0.108	0.179	0.209	0.139	0.134	0.277	0.062
I	0.211	0.122	0.108	0.083	0.367	0.109	0.185	0.211	0.120	0.131	0.263	0.091
J	0.213	0.125	0.112	0.089	0.356	0.105	0.189	0.204	0.144	0.139	0.264	0.061
K	0.210	0.124	0.107	0.085	0.366	0.109	0.176	0.204	0.111	0.132	0.275	0.101
L	0.209	0.124	0.114	0.089	0.357	0.107	0.184	0.200	0.144	0.137	0.271	0.064
M	0.213	0.126	0.111	0.086	0.358	0.107	0.176	0.214	0.126	0.128	0.268	0.088
N	0.212	0.126	0.110	0.087	0.359	0.107	0.173	0.206	0.127	0.134	0.272	0.088
O	0.211	0.124	0.111	0.084	0.362	0.108	0.178	0.211	0.115	0.133	0.269	0.094
P	0.213	0.125	0.112	0.086	0.358	0.107	0.177	0.205	0.127	0.137	0.272	0.082
Q	0.213	0.125	0.110	0.086	0.360	0.107	0.180	0.216	0.125	0.127	0.267	0.085
R	0.213	0.125	0.111	0.086	0.360	0.106	0.176	0.208	0.130	0.139	0.264	0.084
S	0.212	0.125	0.109	0.086	0.360	0.108	0.184	0.209	0.143	0.142	0.260	0.063
T	0.211	0.122	0.108	0.083	0.366	0.110	0.177	0.207	0.117	0.124	0.275	0.100
U	0.212	0.125	0.110	0.085	0.360	0.108	0.175	0.202	0.121	0.134	0.274	0.094
V	0.210	0.122	0.108	0.083	0.368	0.109	0.177	0.197	0.117	0.122	0.285	0.101
W	0.210	0.123	0.107	0.083	0.367	0.109	0.182	0.200	0.105	0.122	0.288	0.102
X	0.210	0.123	0.107	0.083	0.367	0.110	0.182	0.206	0.110	0.123	0.278	0.101
Y	0.211	0.122	0.107	0.083	0.367	0.110	0.171	0.196	0.115	0.127	0.288	0.103
Z	0.210	0.122	0.107	0.083	0.369	0.108	0.177	0.196	0.119	0.130	0.277	0.101

The benchmark forecasts are obtained as follows: VI = Benchmark forecast obtained by aggregating voting intentions, ML = Maximum likelihood model, Bayes 1 = Bayesian model using prior $N(0, B)$ with $B = 1.5I_{L^*}$, Bayes 2 = Bayesian model using prior $N(0, B)$ with $B = 100I_{L^*}$.

D Results: Histograms of the number of clusters K



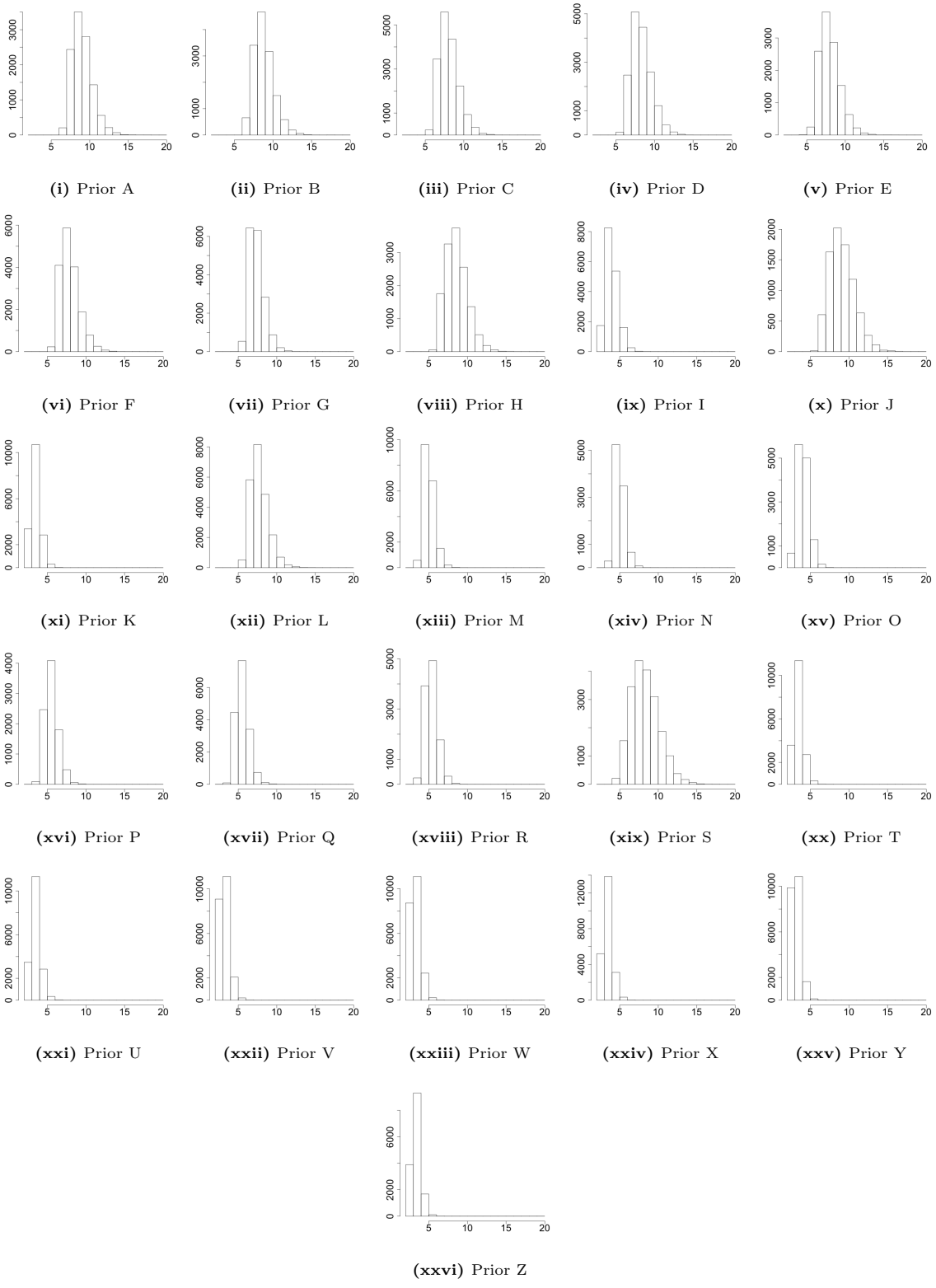
(i) Prior A **(ii)** Prior B **(iii)** Prior C **(iv)** Prior D **(v)** Prior E

(vi) Prior F **(vii)** Prior G **(viii)** Prior H **(ix)** Prior I **(x)** Prior J

(xi) Prior K **(xii)** Prior L **(xiii)** Prior M **(xiv)** Prior N **(xv)** Prior O

(xvi) Prior P **(xvii)** Prior Q **(xviii)** Prior R **(xix)** Prior S

Figure 4: Histograms of the number of clusters K in the posterior draws for the 2010 election under prior specification A-S.



(i) Prior A **(ii)** Prior B **(iii)** Prior C **(iv)** Prior D **(v)** Prior E

(vi) Prior F **(vii)** Prior G **(viii)** Prior H **(ix)** Prior I **(x)** Prior J

(xi) Prior K **(xii)** Prior L **(xiii)** Prior M **(xiv)** Prior N **(xv)** Prior O

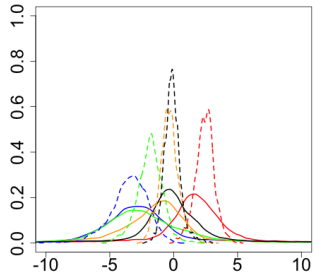
(xvi) Prior P **(xvii)** Prior Q **(xviii)** Prior R **(xix)** Prior S **(xx)** Prior T

(xxi) Prior U **(xxii)** Prior V **(xxiii)** Prior W **(xxiv)** Prior X **(xxv)** Prior Y

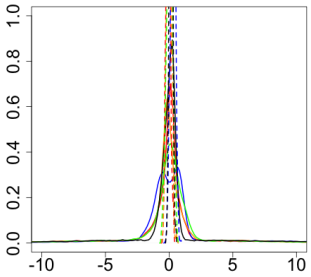
(xxvi) Prior Z

Figure 5: Histograms of the number of clusters K in the posterior draws for the 2012 election under prior specification A-Z.

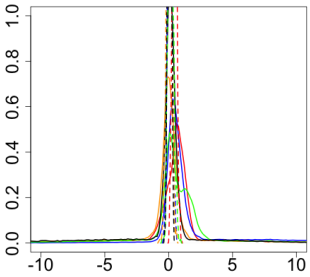
E Results: Posterior Density Plots



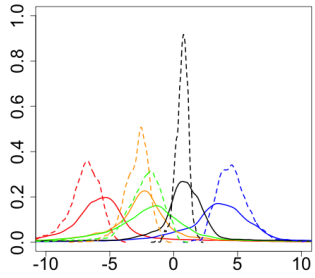
(i) Intercept



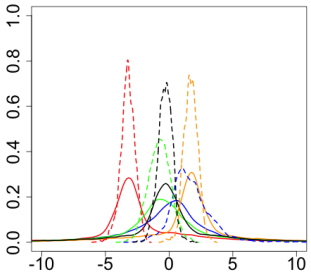
(ii) Logarithm of Age



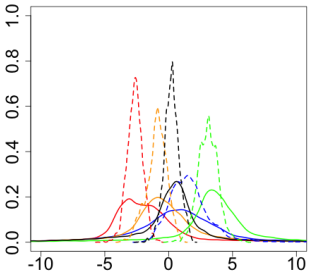
(iii) Logarithm of Income



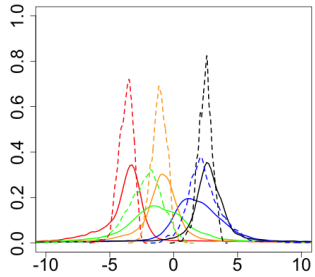
(iv) Party Choice in 2010 = PvdA



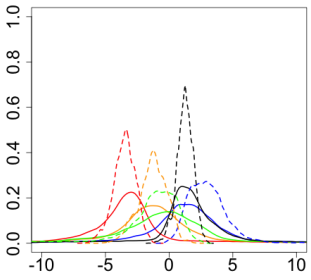
(v) Party Choice in 2010 = PVV



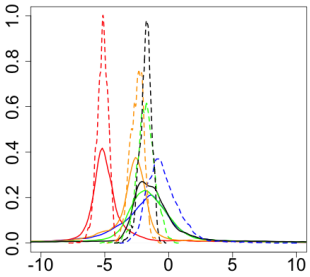
(vi) Party Choice in 2010 = CDA



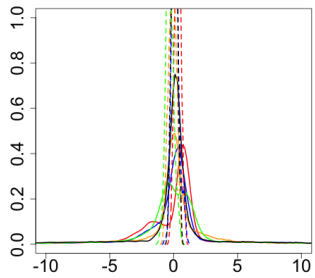
(vii) Party Choice in 2010 = Other



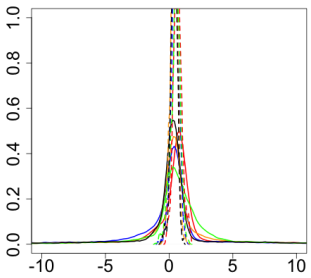
(viii) Party Choice in 2010 = PNTS/DNK



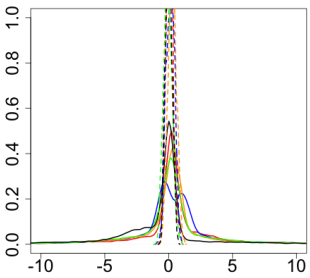
(ix) Party Choice in 2010 = No Vote



(x) Female

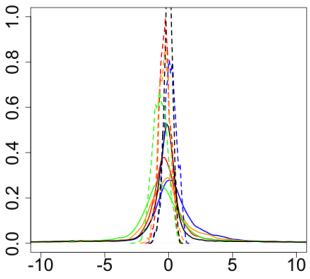


(xi) Home owner

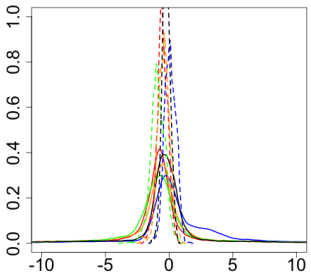


(xii) Married

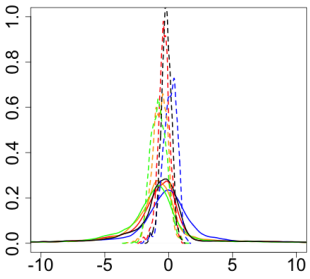
Figure 6: Posterior density plots of all β coefficients under DPM model X and benchmark model Bayes 2. A solid line indicates draws of model X, while the dashed line corresponds to draws from Bayes 2. The different colours denote different party choices, where red = VVD, blue = PvdA, orange = PVV, green = CDA and black = Other.



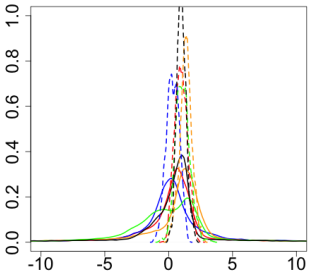
(xiii) Very Urban



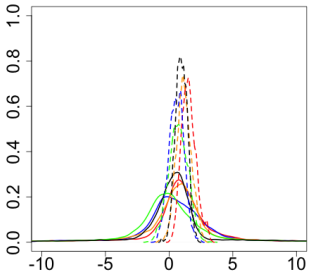
(xiv) Moderate/Slight Urban



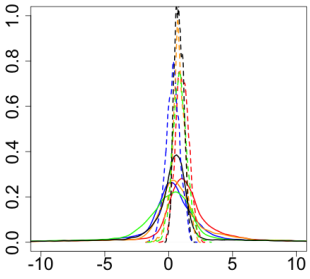
(xv) Not Urban



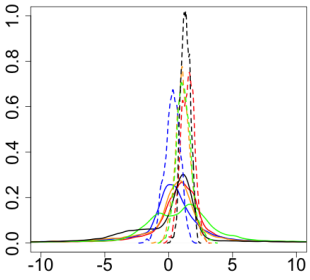
(xvi) Education = Vmbo



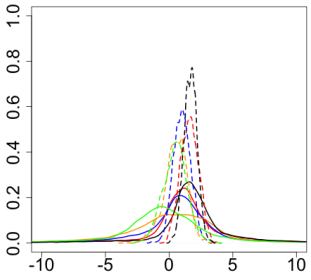
(xvii) Education = Havo/VWO



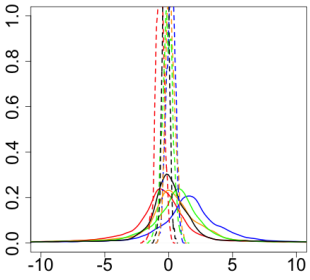
(xviii) Education = MBO



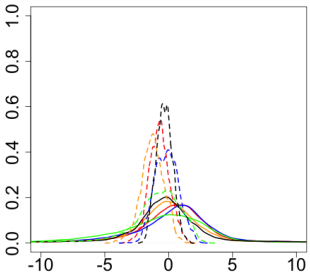
(xix) Education = HBO



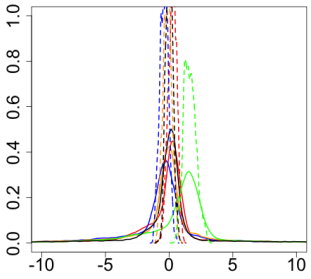
(xx) Education = WO



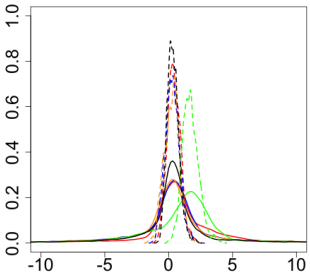
(xxi) Employed



(xxii) Job Seeker

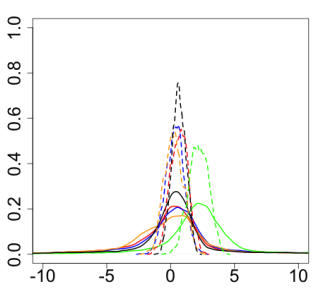


(xxiii) Religion = Roman Catholic

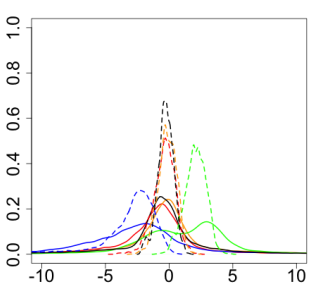


(xxiv) Religion = Protestant

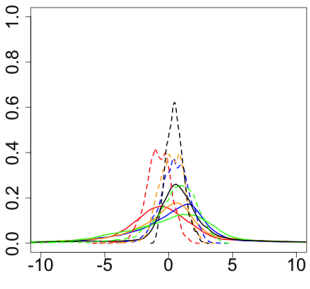
Figure 6: Posterior density plots of all β coefficients under DPM model X and benchmark model Bayes 2. A solid line indicates draws of model X, while the dashed line corresponds to draws from Bayes 2. The different colours denote different party choices, where red = VVD, blue = PvdA, orange = PVV, green = CDA and black = Other.



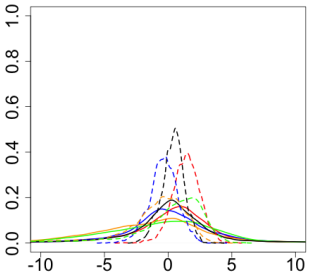
(xxv) Religion = Dutch Reformed



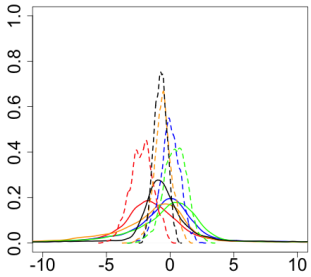
(xxvi) Religion = Reformed Churches



(xxvii) Religion = Other Christian Religions



(xxviii) Religion = Islam



(xxix) Religion = Other

Figure 6: Posterior density plots of all β coefficients under DPM model X and benchmark model Bayes 2. A solid line indicates draws of model X, while the dashed line corresponds to draws from Bayes 2. The different colours denote different party choices, where red = VVD, blue = PvdA, orange = PVV, green = CDA and black = Other.

References

Abramowitz, A. I. (2000). Bill and als excellent adventure: Forecasting the 1996 presidential election. In *Before the vote: Forecasting American national elections*. Sage Publications, Inc.

Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science & Politics*, 41(04):691–695.

Abramowitz, A. I. (2016). Will time for change mean time for trump? *PS: Political Science & Politics*, 49(4):659–660.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355.

Blumenthal, M. (2014). Polls, forecasts, and aggregators. *PS: Political Science & Politics*, 47(02):297–300.

Broderick, T. (2015). Bayesian nonparametrics. Max Planck Institute for Intelligent Systems Tübingen Lecture.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Campbell, J. E. (2008). The trial-heat forecast of the 2008 presidential vote: Performance and value considerations in an open-seat election. *PS: Political Science & Politics*, 41(04):697–701.

Campbell, J. E. (2016a). Forecasting the 2016 american national elections. *PS: Political Science & Politics*, 49(4):649–654.

Campbell, J. E. (2016b). The trial-heat and seats-in-trouble forecasts of the 2016 presidential and congressional elections. *PS: Political Science & Politics*, 49(4):664–668.

Campbell, J. E. and Lewis-Beck, M. S. (2008). Us presidential election forecasting: An introduction. *International Journal of Forecasting*, 24(2):189–192.

Campbell, J. E., Norpoth, H., Abramowitz, A. I., Lewis-Beck, M. S., Tien, C., Erikson, R. S., Wlezien, C., Lockerbie, B., Holbrook, T. M., Jerôme, B., et al. (2017). A recap of the 2016 election forecasts. *PS: Political Science & Politics*, 50(2):331–338.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Erikson, R. S. and Wlezien, C. (1999). Presidential polls as a time series: the case of 1996. *Public Opinion Quarterly*, pages 163–177.

Erikson, R. S. and Wlezien, C. (2008). Leading economic indicators, the polls, and the presidential vote. *PS: Political Science & Politics*, 41(04):703–707.

Erikson, R. S. and Wlezien, C. (2016). Forecasting the presidential vote with leading economic indicators and the polls. *PS: Political Science & Politics*, 49(4):669–672.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Frühwirth-Schnatter, S. and Frühwirth, R. (2010). Data augmentation and mcmc for binary and multinomial logit models. In *Statistical Modelling and Regression Structures*, pages 111–132. Springer.

Frühwirth-Schnatter, S. and Frühwirth, R. (2012). Bayesian inference in the multinomial logit model. *Austrian Journal of Statistics*, 41(1):27–43.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments.

Geweke, J. and Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics*, 138(1):252–290.

Görür, D. and Rasmussen, C. E. (2010). Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664.

Graefe, A. (2015). German election forecasting: Comparing and combining methods for 2013. *German Politics*, 24(2):195–204.

Graefe, A., Armstrong, J. S., Jones, R. J., and Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, 30(1):43–54.

Graefe, A., Jones Jr, R. J., Armstrong, J. S., and Cuzán, A. G. (2016). The pollyvote forecast for the 2016 american presidential election. *PS: Political Science & Politics*, 49(4):687–690.

Graefe, A., Küchenhoff, H., Stierle, V., and Riedl, B. (2015). Limitations of ensemble bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3):943–951.

Greenberg, E. (2012). *Introduction to Bayesian econometrics*. Cambridge University Press.

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12(Jun):1923–1953.

Hibbs, D. A. (2008). Implications of the bread and peacemodel for the 2008 us presidential election. *Public Choice*, 137(1-2):1.

Hibbs, D. A. (2012). Obama's reelection prospects under bread and peace voting in the 2012 us presidential election. *PS: Political Science & Politics*, 45(04):635–639.

Hibbs, D. A. (2013). The bread and peace model: 2012 presidential election postmortem. *PS: Political Science & Politics*, 46(1):41–41.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press.

Holbrook, T. M. (2012). Incumbency, national conditions, and the 2012 presidential election. *PS: Political Science & Politics*, 45(04):640–643.

Holbrook, T. M. (2016). National conditions, trial-heat polls, and the 2016 election. *PS: Political Science & Politics*, 49(4):677–679.

Holmes, C. C., Held, L., et al. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168.

Hummel, P. and Rothschild, D. (2013). Fundamental models for forecasting elections. *ResearchDMR. com/HummelRothschild_FundamentalModel*.

Hummel, P. and Rothschild, D. (2014). Fundamental models for forecasting elections at the state level. *Electoral Studies*, 35:123–139.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283.

Jackman, S. (2014). The predictive power of uniform swing. *PS: Political Science & Politics*, 47(2):317–321.

Jerme, B. and Jerme-Speziari, V. (2016). State-level forecasts for the 2016 us presidential elections: Political economy model predicts hillary clinton victory. *PS: Political Science & Politics*, 49(4):680–686.

Kayser, M. A. and Leininger, A. (2016). A predictive test of voters' economic benchmarking: The 2013 german bundestag election. *German Politics*, 25(1):106–130.

Kennedy, R., Wojcik, S., and Lazer, D. (2017). Improving election prediction internationally. *Science*, 355(6324):515–520.

Lewis-Beck, M. S. and Dassonneville, R. (2015a). Comparative election forecasting: Further insights from synthetic models. *Electoral Studies*, 39:275–283.

Lewis-Beck, M. S. and Dassonneville, R. (2015b). Forecasting elections in europe: Synthetic models. *Research & Politics*, 2(1):2053168014565128.

Lewis-Beck, M. S. and Rice, T. W. (1984). Forecasting presidential elections: A comparison of naive models. *Political Behavior*, 6(1):9–21.

Lewis-Beck, M. S. and Tien, C. (2008). The job of president and the jobs model forecast: Obama for'08? *PS: Political Science & Politics*, 41(04):687–690.

Lewis-Beck, M. S. and Tien, C. (2016). The political economy model: 2016 us election forecasts. *PS: Political Science & Politics*, 49(4):661–663.

Linzer, D. A. (2013). Dynamic bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, 108(501):124–134.

Lockerbie, B. (2016). Economic pessimism and political punishment. *PS: Political Science & Politics*, 49(4):673–676.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). The bugs book. *A Practical Introduction to Bayesian Analysis, Chapman Hall, London*.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1):207–240.

Müller, P. and Quintana, F. A. (2004). Nonparametric bayesian data analysis. *Statistical science*, pages 95–110.

Neal, R. M. (1992). Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Norpoth, H. (2016). Primary model predicts trump victory. *USA TODAY*, 7:16.

Nydick, S. W. (2012). The wishart and inverse wishart distributions. *Electron. J. Statist.*, 6:1–19.

Paap, R., van Nierop, E., Van Heerde, H. J., Wedel, M., Franses, P. H., and Alsem, K. J. (2005). Consideration sets, intentions and the inclusion of don't know in a two-stage model for voter choice. *International Journal of Forecasting*, 21(1):53–71.

Panagopoulos, C. (2009). Polls and elections: preelection poll accuracy in the 2008 general elections. *Presidential Studies Quarterly*, 39(4):896–907.

Pasek, J. (2015). Predicting elections: Considering tools to pool the polls. *Public Opinion Quarterly*, 79(2):594–619.

PollyVote (n.d.). Which method provides the most accurate election forecasts? Retrieved April 25, 2017, from http://pollyvote.com/en/combining-forecasts/which-forecasting-method-provides-the-most-accurate-forecasts/.

Quinn, K. M., Martin, A. D., and Whitford, A. B. (1999). Voter choice in multi-party democracies: a test of competing theories and models. *American Journal of Political Science*, pages 1231–1247.

Rigdon, S. E., Jacobson, S. H., Tam Cho, W. K., Sewell, E. C., and Rigdon, C. J. (2009). A bayesian prediction model for the us presidential election. *American Politics Research*, 37(4):700–724.

Roberts, G. O., Rosenthal, J. S., et al. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367.

Rossi, P. E. (2014). *Bayesian non-and semi-parametric methods and applications*. Bayesian non-and semi-parametric methods and applications.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley & Sons.

Scott, S. L. (2011). Data augmentation, frequentist estimation, and the bayesian analysis of multinomial logit models. *Statistical Papers*, 52(1):87–109.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Shahbaba, B. and Neal, R. (2009). Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, 10(Aug):1829–1850.

Silver, N. (2014). Which method provides the most accurate election forecasts? Retrieved April 7, 2017, from https://fivethirtyeight.com/features/how-fivethirtyeight-calculates-pollster-ratings/.

Silver, N. (2016a). A users guide to fivethirtyeights 2016 general election forecast. Retrieved May 8, 2017, from https://fivethirtyeight.com/features/a-users-guide-to-fivethirtyeights-2016-general-election-forecast/.

Silver, N. (2016b). Who will win the precidency? Retrieved May 8, 2017, from https://projects.fivethirtyeight.com/2016-election-forecast/?ex_cid=2016-election.

Taddy, M. A. and Kottas, A. (2010). A bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics*, 28(3):357–369.

Teh, Y. W. (2010). Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer.

Teh, Y. W. (2013). Bayesian nonparametrics. Max Planck Institute for Intelligent Systems Tübingen Lecture.

Villani, M., Kohn, R., and Giordani, P. (2009). Regression density estimation using smooth adaptive gaussian mixtures. *Journal of Econometrics*, 153(2):155–173.

Wang, S. S.-H. (2015). Origins of presidential poll aggregation: A perspective from 2004 to 2012. *International Journal of Forecasting*, 31(3):898–909.

West, M., Müller, P., and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty, Freeman P. R., Smith A. F. M. (eds.)*, pages 363–386. John Wiley.

Wlezien, C. (2015). The myopic voter? the economy and us presidential elections. *Electoral Studies*, 39:195–204.