Measuring Political Preferences of the U.S. Voting Population

Measuring Political Preferences of the U.S. Voting Population

Ali Nahm

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of Bachelor of
Arts in Computer Science

Measuring Political Preferences of the U.S. Voting Population

by

Ali Nahm

Submitted to the Department of Computer Science
on April 1, 2015, in partial fulfillment of the
requirements for the degree of
Bachelor of Arts in Computer Science

Abstract

Political polarization is a common topic in the news and media, but not much has been done to understand the distribution of the preferences of the U.S. voting population. Political scientists have drawn different conclusions on the current state of political polarization within the U.S. voting population based on survey data and basic spatial voting models. In this work, I present a spatial voting model that analyzes voting data at a more fine-grained level in order to use Bayesian techniques to infer the underlying distribution of political preferences of the population. Further, I verify these results by comparing it to alternative public opinion measurements and measuring the accuracy in completing prediction tasks. This work adds a new perspective to the current discussion within the political science community of the recent trends of political polarization.

Acknowledgments

I would like to first thank my mentor Peter Krafft for all of his guidance and help throughout this research process. Since the first day I stumbled into his office, he has always been patient with me, willing to answer all of my questions, help debug my code, and step through different math ideas together. I would also like to thank my thesis advisor Matt Blackwell for his willingness to jump onboard in the middle of this project and provide valuable feedback on work not directly within his academic field. Further, I would like to thank Sandy Pentland, Krzysztof Gajos, and David Parkes for their insights and encouragement on this project despite their busy schedules. My computer and I are especially grateful to David Parkes for introducing me to the world of cluster computing and enabling me to generate much-needed results faster. I am indebted to all these people mentioned above for inspiring me to continue work in the interdisciplinary field of computational social science through their ground-breaking research and engaging conversations.

In addition to academic mentorship, I could not have written this thesis without the unconditional support of my family and friends. Countless times, I went to them saying I wanted to stop, and, countless times, I have been convinced otherwise. I cannot thank my friends enough for understanding the many times I have abandoned them the past couple of months and still helping me in times of need. True friends are those willing to run code on their computers to help you generate results and help create LaTeX tables late at night, among other things. A true boyfriend does all of that of a friend, in addition to tolerating much more grumpiness. I would like to thank Alex, for being my voice of reason throughout this process and reminding me of the importance of having fun every once in a while. Through bouts of grumpiness and excitement over random plots, you've always been there for me. And finally, I would like to thank the inventor of waffles for creating the greatest recipe because, well, why not.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The political elite in the United States has become increasingly polarized in recent years [44, 19]. This political polarization puts Congress in a gridlock and hinders legislative productivity from effectively serving the public [31]. This standstill in Congress not only affects the political elite, but also the general U.S. population. Scholars believe some of these detrimental effects include a lack of updated laws to reduce inequality given the current standard of living and a full government shutdown in October 2013 closing down basic needs such as the annual influenza program hosted by the Center for Disease Control [31, 10]. Political polarization appears to be prevalent in the U.S. federal government with negative effects on the entire country.

But what about the general U.S. population? Is political polarization prevalent within the U.S. electorate as well? In the year 2014, political scientists estimate there are over 245 million people eligible to vote in the U.S. (from here on referred to as the U.S. voting population) [32]. These people hold the important power in the political process of electing who contributes to decisions for the nation. Many political scientists have developed methods of measuring the distribution of political preferences in the U.S. voting population. I will add to these works with another method based on the idea that subgroups of the voting population follow unique distributions of political preferences.

The media has also developed methods and claims about the voting population. While these claims state the population is very polarized, most of the media focus is on the outspoken minorities of the U.S. voting population with extreme political views, such as the Tea Party Movement [35]. In this work, I consider those with extreme views, as well as those within the "silent majority" that was coined by President Nixon [36]. Nixon proposes that the silent majority consists of the majority of the U.S. voting population with moderate political views that are unknown to the public, whereas the minority of the population with extreme political views are sharing their views through events such as public demonstrations and protests. It is important to understand the political preferences of a variety of voters rather than only focusing on those with extreme views in order to understand the state of political polarization within the entire voting population.

1.1 The Motivating Problem

Political scientists have already been debating the shape of the distribution of political preferences of the U.S. voting population and the resulting conclusions on the state of political polarization. Political polarization is the phenomenon that political preferences are becoming much more extreme relative to each other. Some political scientists claim that the U.S. voting population is becoming more polarized due to a decrease in the turnout of voters in the silent majority [20]. They theorize that because politicians and the political elite have more more extreme views and are more polarized than before, more moderate voters feel more cynical about the government and choose to not participate [36, 22]. McCarty, Poole, and Rosenthal have also considered polarization in a different light, by observing a strong relationship between the increase in political polarization and economic inequality [31]. On the other hand, other political scientists believe voters are not as polarized as what is portrayed by the media [12, 17, 19]. Fiorina hypothesizes that voters who identify strongly within a political party are more divided because of polarization among the political elite rather than the electorate [19].

However, most of the methods in these studies discussed so far only measure political preferences and polarization using survey data, which may not have enough data points to represent each portion of the population well. The state governments U.S. government have released precinct-level voting data of past elections. This voting data provide information about political behavior of the U.S. voting population, which I can utilize to have a new perspective of the political preferences and extent of polarization in voters. Some researchers have already begun to utilize this data to infer political preferences of the U.S. voting population, which I will discuss later on in Section 2.3 of this thesis. However, these works have not used the inferred distributions of political preferences of voters in order to examine political polarization.

1.2 Contributions

My thesis intends to contribute to the intersection of computer science and political science. In the field of computer science, this work aims to apply theoretical statistics models to large datasets of the entire U.S. voting population. Most computer science research focuses on developing faster or more complicated theoretical models rather than fitting models to datasets of real human behavior. This work will demonstrate the importance of calibrating models with real datasets to prove the robustness of a model.

Furthermore, in the field of political science, my thesis investigates an alternative model to quantify political preferences of the U.S. voting population. The new model I will introduce makes use of more fine-grained precinct-level voting results instead of district-level voting results or nationwide surveys which yields a better approximation of regional preferences. The new model also infers preferences of clusters of multiple precincts rather than individual precincts in order to borrow information from precincts with similar distributions of voters to make our estimates more precise. This new model will suggest alternative assumptions and adjustments to models of voting behavior for future work.

1.3 Overview

Over the course of this thesis, I will discuss the new model and methods used to infer the political preferences of the U.S. voting population. Chapter 2 will provide context for my work, through a discussion in Sections 2.1 and 2.2 on the necessary background of the U.S. political and electoral systems and a summary in Section 2.3 on related works that develop quantitative metrics of political preferences. Given this context, Chapter 3 will then describe relevant spatial voting models in Section 3.1 and the specifics of the new model I present in this thesis.

Chapter 4 explains the data mining methods that yield a final dataset of precinct-level voting results throughout the U.S. I then test the fit of the new model to the voting results data by generating sets of parameter values through Bayesian inference techniques that are discussed in Chapter 5. In addition, I will discuss how we can apply the inference results to answer the motivating question and better understand the state of political polarization within the U.S. voting population. Chapter 6 outlines various validation methods used to ensure the accuracy of the new model. Finally, I summarize and suggest improvements for future work in Chapter 7.

Chapter 2

Preliminaries

The purpose of this section is to explain the larger context of this work. The first portion of this section will explain the basics of the U.S. federal election system. This information is important to understand the model and data in this work. I then discuss related works that have developed methods to compute quantitative estimates of political preferences of the political elite and U.S. voting population.

2.1 American Politics for Dummies

U.S. federal election involves many rules and terminology that are necessary to understand the model I will introduce later in chapter 3. The U.S. voting population votes in elections for three main federal positions that are described in Table 2.1. Each position varies in the total number holding the position at any given time, the level of representation, and the duration of a single term. The level of representation of a position signifies the geographic region that is represented by that specific position. The duration is measured in years and can also be thought of as the number of years until another election for the federal position occurs.

For my work, I will be focusing on Congressional elections of candidates vying for a seat in the House of Representatives. In these elections, candidates must win the majority of the popular vote of a Congressional District.

Federal Position	Total in Office	Level of Representation	Frequency of Election
President	1	national	4
Senator	100	state	6*
House Representative	435	Congressional District	2

Table 2.1: Table describing the three main federal elections. The * notes that Senatorial elections are on different cycles such that *at most* only a third of the Senate changes every two years.

2.2 Geographic Regions for Vote Tabulation

The U.S. is broken into a variety of geographic regions with different average land area. Each region yields a different level of granularity of analysis of the behavior of the population residing in that region. For this paper, I focus on three types of regions, ordered from largest to smallest average land area: states, Congressional Districts, and voting precincts. Each of these geographic regions can be seen in Figure 2-1.

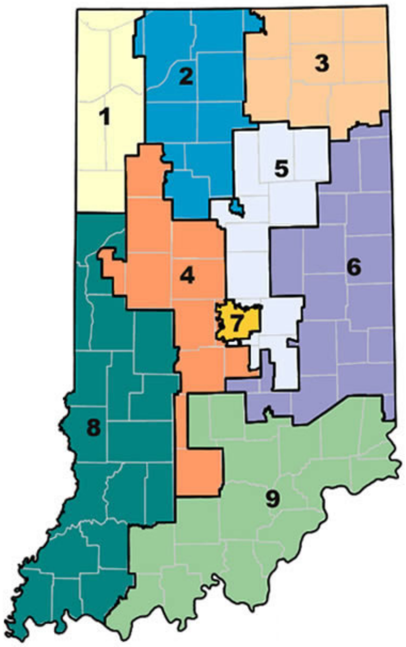

Figure 2-1: Boundaries of the state of Indiana, Congressional Districts (numbered, outlined in black), and precincts (outlined in grey) for the 2001 election cycle [13].

There are fifty states in the U.S., each with a wide variance in land area and population size. The state boundaries were determined upon the formation of each state and are have not changed during the 10 year span of election data we are considering.

Congressional Districts are geographic areas located completely within their assigned U.S. state. The number of Congressional Districts is assigned to each state after the tabulation of each decennial census. Each state is assigned a minimum of one Congressional District, with additional districts assigned roughly proportional to its population, such that there are a total of 435 representatives [1]. Given the amount assignment, each state government determines its own Congressional District boundaries. The new Congressional District boundaries generally go into effect two years after the Census is completed. For instance, after the 2000 Census, the same Congressional District boundary lines are used for the 2002, 2004, 2006, 2008, and 2010 Congressional Elections.

The data specifically used in my model depend on Congressional District boundaries that were set by the results of the 2000 Census. As for July 2001, the average size of a Congressional District based on the 2000 Census apportionment population was 646,952 people [1]. A political process known as gerrymandering can occur when the suggested boundaries of Congressional Districts will provide partisan advantage [2].

Precincts, also known as voting districts (VTDs), are smaller geographic regions within Congressional Districts that are established by state governments in order to easier tabulate elections [3]. Because the geographic areas of precincts are smaller than the areas of Congressional Districts, there also tends to be a lower population in a precinct than a Congressional District. After each election, each precinct must report to its assigned district the resulting vote shares. Each district then aggregates all of the reported precinct-level results to have district-level election results to determine the winner of the Congressional election. By breaking down the problem of counting votes into smaller parts, there is less likelihood of error in reported vote share.

For reference, in the decade of 2000-2010, Texas was allocated 32 Congressional Districts based on the 2000 U.S. Census [1]. The Texas state government then broke the 32 Congressional Districts into 8,400 precincts [1]. On average, this means there are approximately 262 precincts per Congressional District. Thus, we can see that precinct-level voting records would provide more fine-grained analysis of the political ideologies of voters in the overall district than district-level voting records.

2.3 Related Works

Some work has already been done in the field of political science to develop methods of approximating quantitative scores of ideology. These scores are useful summary tools of candidate positions, which is commonly believed to be a subjective concept. Bonica, for instance, writes:

> Ideological measures of political actors and institutions are essential for testing theories about political behavior and institutions and are commonplace in research topics ranging from public opinion, elections, and representation to legislative and judicial behavior and political institutions [9].

In this thesis, I apply the quantitative ideology scores that my model infers to better understand political polarization within the U.S. voter population, which we will discuss later on. Other applications of these quantitative scores could be to measure the accuracy of legislatures representing their constituents or to suggest more fair Congressional District redistricting plans.

2.3.1 Understanding the Political Elite

Most related works approximate quantitative political ideology scores of the political elite, including elected officials and key leaders of private industry. There is more tracked data and information about the political elite, such as actions in office or financial contributions, which allows for better predictions and approximations of their political views.

Poole and Rosenthal approximate an ideological score, called a NOMINATE score, for each elected official in Congress. The NOMINATE scores are approximated using the roll call voting history of the legislators [37]. Bonica, the political scientist cited earlier, approximates ideological scores not only for elected officials, but also for losing candidates and major political donors. In his work, he creates ideal point estimate scores as common-space campaign-finance scores (CFscores), which are based on the assumption that donors will donate to politicians with similar political beliefs [9].

2.3.2 Understanding the Voters

There exist related works that approximate political ideologies of the U.S. voting population. Unlike the previously discussed works that analyze the political elite, there is only one main piece of information about the political behavior of voters, which is his or her vote in an election. To circumvent this issue, many political scientists have turned to survey-based methodologies to better understand the political preferences of the U.S. population [34, 38, 42]. Miller and Stokes disaggregated national surveys to gain district-level opinions on different important political topics [34]. However, the survey sample size was very small of 13 responses per district on average. The small sample size added a large measurement error term to each estimate that makes the estimated value less reliable for inference [16].

More recently, Christopher Tausanovitch and Christopher Warshaw introduce a multi-level regression and post-stratification (MRP) model to approximate ideal points of Congressional Districts [38]. This MRP model is also based on aggregated national survey data, but Tausanovitch and Warshaw attempt to overcome the small sample size issue of Miller and Stokes by incorporating additional demographic and geographic information about Congressional Districts. However, recent work by Stephen Ansolabehere and Eitan Hersh found that 52% of the non-voter respondents of the 2008 CCES survey, one of the surveys used by Tausanovitch and Warshaw, misreport that they voted in the recent election [4]. Given that a portion of respondents misreported their voting behavior, it is plausible that respondents misreported other responses in the 2008 CCES survey, or any other survey for that matter [4].

There also exists literature that has used vote share data and electoral returns data to approximate political ideologies of the U.S. voting population [29, 26, 25, **?**]. However, most of these works use district-level voting results, whereas I use more fine-grained precinct-level voting results. The population of voters of a precinct is a subset of the population of the corresponding district. Thus, precinct-level results summarize the voting behavior of a subset of the voting population of the district. Aggregating estimates of subsets compared to estimating a single whole item yields more variability in the estimate and more detailed analysis of the item. People rarely behave in identical manners, so considering more variability of behaviors is a more realistic analysis of the group of people. To understand how aggregating estimates of subsets is a more detailed analysis, consider the task of estimating the color of a red and white blanket. An estimation of the whole blanket would be the color pink, whereas the combination of estimates of sections of the blanket would be the colors red and white. The latter is a more accurate description of the blanket. The usage of precinct-level voting results allows my model to make inferences of smaller subgroups of the population, and thus more detailed inferences of the population as a whole.

The most similar work to this work is by Levendusky, et al. [29]. Both my work and the work of Levendusky, et al. create models based on vote shares of federal elections in the past decade in order to infer ideal point estimates of smaller populations of the U.S. electorate [29]. However, Levendusky, et al develops a model using district-level vote shares and a latent parameter representing the partisanship of each district, while we develop a model using precinct-level vote shares and a latent parameter representing the the political preferences of individual voters [29]. Levendusky, et al. also addresses the issue of missing data due to elections with uncontested candidates by including an additional term in their model that accounts for additional information such as an incumbency offset and uncontested candidate offset.

Georgia Kernell also uses election returns to infer the political ideologies of districts [26]. Kernell chooses to infer parameters based on a compilation of multiple election returns in districts rather than a single election at a time [26]. I chose to create a separate model for each election year in order to analyze the changes in the distribution of political preferences over time. This decision is especially important when I apply the inferred distributions from my model to observe the trends in political polarization over time in section 5.4.

Chapter 3

Model

The purpose of this chapter is to discuss a newly developed statistical model to represent the political preferences of the U.S. voting population. I begin by providing more context in section 3.1 for the model by describing related works that have developed similar models of voter behavior. In the following section, I describe the specific details of the model. I then re-introduce the model in terms of probability distributions for the purposes of Bayesian inference.

3.1 Spatial Voting Models

I first describe some related works that develop mathematical theories of voting behavior that inform the construction of my model. Anthony Downs was the first to introduce a spatial voting model for rational voting and turnout behavior [14, **?**]. The Downsian model is based in an ideal world where each voter and political candidate has an ideal point on a one-dimensional policy space [21]. Further, Downs assumes that elections select a single candidate out of two by the majority vote of a single constituency [21]. Given these assumptions about the world and elections, the Downsian model states that each voter selects the candidate that yields higher expected utility in an ideal world. A graphical interpretation of the main tenet of the spatial voting model can be seen in Figure 3-1.

Melvin J. Hinich further develop the Downsian spatial voting model to also ac-

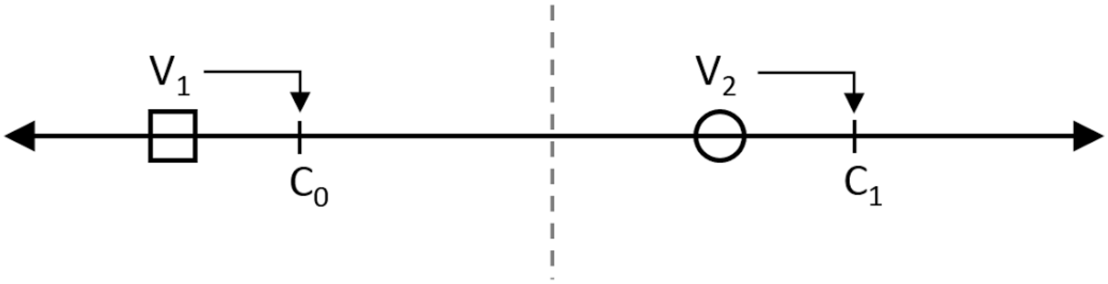

Figure 3-1: Basic line graph plot of the idea of the spatial voting model. The dotted line represents the midpoint of the two candidate positions.

count for the uncertainty of voters [23]. His model assumes that voters consider each candidate as a distribution of political preferences rather than a single point on a one-dimensional scale [23]. This assumption is more realistic of the behavior and beliefs of voters. Intuitively, in the model, voters are more uncertain about voting for a candidate if the candidates present position and past record diverge, even if that candidate may yield a higher expected utility than the alternative.

Other political scientists have improved the Downsian spatial voting model by adjusting for specific attributes of the candidate or environmental factors [15, 18, 27, 28]. James Enelow and Melvin J. Hinich develop a Downsian model that considers additional qualities of candidates unrelated to determining the candidate's position on the one-dimensional policy preference space, such as the personality of the candidate [**?**]. Gerald H. Kramer develops a Downsian model that adjusts the expected vote share of a political party based on economic conditions of the environment of the voter, such as per capita income [27]. He reasons that the political preferences of voters are based on policy outcomes and resulting economic events of the current set of legislatures, in addition to the voters' belief of the ideological ideal points of both candidates. This class of spatial voting models is more robust than the previous two types of model in modeling the behavior of voters in uncontested elections because candidate and environmental information exist in every election. The model I introduce in Section 3.2 does not consider additional factors affecting voting behavior. Factoring in candidate information or additional data about districts to better utilize uncontested election data would be a fruitful topic for future research.

3.2 A Novel Model

My statistical model consists of a generative process for the vote shares of both major candidate in a Congressional election of each U.S. voting precinct. A basic idea of the generative model can be seen in Figure 3-2

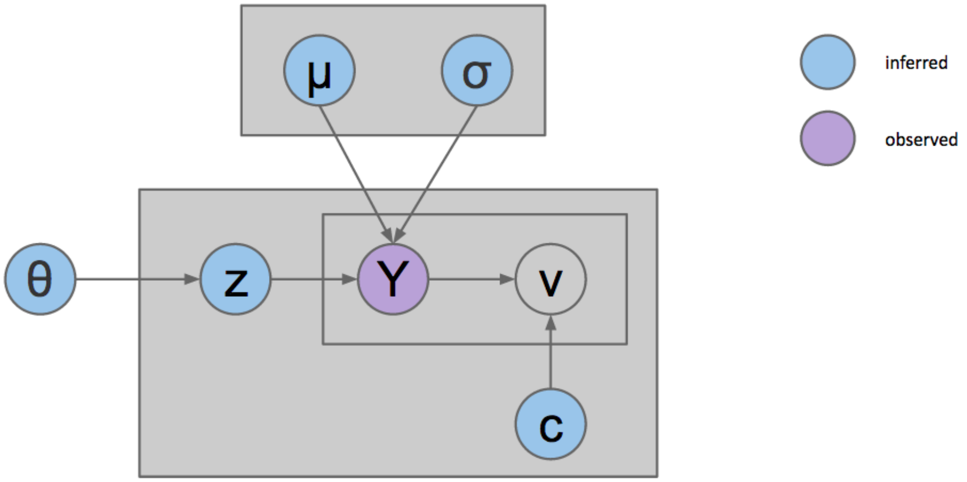

Figure 3-2: Simple diagram of the generative model I introduce in this section.

As discussed earlier, a U.S. voting precinct is the smallest geographic unit used to divide the U.S. voter population for the purposes of election results tabulation. In the model, each precinct (i) with N_i total voters is associated with an election of exactly two candidates, where each candidate is assigned a number of votes from the set of voters within the precinct.

In line with a traditional spatial voting model, I assume that both candidates and voters in a precinct election have positions in the same one-dimensional latent space [14]. I define these positions as the *political preferences* of candidates and voters. I assume that these candidate positions are common knowledge among all participants of the election.

Let c_{0i} and c_{1i} to be the political preferences of the two candidates running in the election in precinct i. Like other Downsian models, I assume that each voter j

of precinct i ($j \in \{1, \ldots, N_i\}$) votes for the candidate closest to his preference in the one-dimensional latent policy space according to Euclidean distance [14]. In other words, the number of votes of candidate c_{0i} in the election of precinct i increases once for each voter j when $|j - c_{0i}| < |j - c_{1i}|$.

Further, I assume that each precinct i has a precinct-specific distribution of the political preferences of its voters. For statistical tractability, I assume that each precinct has one of K distributions of preferences, where K is a positive integer value. The precinct-specific distribution for each precinct is determined by that precinct‘s *cluster assignment*. I consider a group of precincts with the same cluster assignment as a *cluster* of precincts. Each cluster is associated with a Normal distribution of preferences, a cluster distribution, that is defined by a unique mean and standard deviation.

The cluster assignment parameter for each precinct i is a latent variable in this new model. The model instead relies on the mixture proportion parameter ($\vec{\theta}$) that is a K-dimensional vector, where each component (θ_k) represents the probability of a precinct being assigned to cluster k, where $k \in \{1, \ldots, K\}$. By the rules of probability, the components of $\vec{\theta}$ follow the constraint that $\sum_{i=1}^{K} \theta_i = 1$.

I treat the candidates‘ positions as fixed, but I treat the voters‘ positions, the precinct assignments, the cluster means and variances, and the expected proportion of precincts assigned to each particular cluster as unknown. By conditioning on direct estimates of candidates‘ positions and observed vote shares per candidate, I use Bayesian inference to arrive at likely values for the unknown parameters, thus estimating the overall distribution of political preferences of the population the data represent.

3.3 Limitations

The challenge of using election returns data to estimate distributions of preferences is that the number of observations per precinct is limited by the number of candidates in the election with reported vote share values. The main statistical leverage for this model comes from two main assumptions I discuss in this section.

First, I assume that groups of precincts can share the same distribution of voter preferences. This assumption is based on the principles of homophily and geographic relatedness for precincts. Homophily is the idea that people who interact with each other more often tend to share similar characteristics [33]. Past research in sociology has shown that human social networks are quite homophilous [33, 24]. Alison C. Watts has also developed a theoretical result that suggests voters vote as they would with full information about candidates even if they adjust their preferences based on their social network [43]. As a result, it is reasonable to assume that social networks of voters also have high degrees of homophily and similar distributions of preferences. With this assumption, my model is able to assign precincts to the same cluster distribution of voter preferences and aggregate information about precincts within the same cluster.

The second main assumption is that the observed vote share data of precinct-level elections is tied to the underlying precinct-specific distributions of preferences. This allows the model to utilize an aggregation of ideology and partisan preferences of individual voters within precincts. However, some political science research suggests that voters tend to vote for candidates in the same party rather than with more similar ideologies to the own personal beliefs of the voters [7]. Additional work should verify this research, and develop better voting models if this finding proves to be a serious limitation. One potential solution is to use alternative data sources as proxies for the political preferences of the U.S. voting population, such as political party registration.

3.4 Bayesian Inference

This section will describe the inference method used to estimate the model. I chose to use two different methods of inference, a MCMC Metropolis-Hastings algorithm and a Python library optimizer function (Scipy.optimize), which I describe in further detail in the next chapter. I describe the same model discussed in Section 3.2 as the following:

$$z_i \sim \text{Multinomial}(\vec{\Theta}) \tag{3.1}$$

$$Y_{ij} \sim N(\mu_{z_i}, \sigma_{z_i}) \tag{3.2}$$

$$v_{ij} \sim \begin{cases} 0 : \text{if } |Y_{ij} - c_{i0}| < |Y_{ij} - c_{i1}| \\ 1 : \text{otherwise} \end{cases} \tag{3.3}$$

where z_i is the cluster assignment for precinct i, such that $z_i \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, M\}$. $\vec{\theta}$ is the K-dimensional vector of probabilities of a precinct being assigned to each of the K possible clusters. Y_{ij} is the political preference of voter j in precinct i that follows a Normal distribution with μ_{z_i} and σ_{z_i}, the mean and standard deviation respectively of the Normal distribution associated with cluster z_i. Finally, v_{ij} is the index of the political candidate that was selected by voter j in precinct i based on the political preferences of the corresponding voter j and the two candidates of the election, c_{0i} and c_{1i}. The data I use only includes reported vote share of two candidates, so $v_{ij} \in \{0, 1\}$.

In order to apply Bayesian inference methods, I find the posterior distribution of the unknown parameters of the model. Bayes' Rule states the posterior distribution is proportional to the product of the prior distributions of all unknown parameters and the likelihood. Note that $\vec{\mu}$ and $\vec{\sigma}$ are vectors of the parameters μ_{z_i} and σ_{z_i} for each possible cluster assignment. Similarly, $\vec{v_j}$ is the vector formed by all votes v_{ij} of precinct j.

$$\begin{aligned}
P(\vec{\mu}, \vec{\sigma}, \theta|\vec{v}) &= \sum_{\vec{z}} P(\vec{\mu}, \vec{\sigma}, \vec{z}, \theta|\vec{v}) \tag{3.4} \\
&\propto P(\vec{\mu})P(\vec{\sigma}) \sum_{z_1=1}^{K} \sum_{z_2=1}^{K} \cdots \sum_{z_M=1}^{K} \left[\prod_{j=1}^{M} \big(P(\vec{v_j}|x_j, \mu_{x_j}, \sigma_{x_j}) P(x_j|\theta) \big) \right] \tag{3.5} \\
&= P(\vec{\mu})P(\vec{\sigma}) \prod_{j=1}^{M} \left[\sum_{x_j=1}^{K} P(\vec{v_j}|x_j, \mu_{x_j}, \sigma_{x_j}) P(x_j|\theta) \right] \tag{3.6} \\
&= P(\vec{\mu})P(\vec{\sigma})P(\vec{v}|\vec{\mu}, \vec{\sigma}, \theta) \tag{3.7}
\end{aligned}$$

In the set of equations above, the posterior distribution is first rewritten in terms of the assignment vector ($\vec{z}$) in Equation 3.4. Because there is a cluster assignment for each precinct, the assignment vector $\vec{z} \in \{0, 1\}^M$, where M is the total number of precincts in the election. I further simplify the posterior distribution in Equation 3.6 by assuming the distributions of political preferences of each of the precinct are independent to each other. The full derivation of the likelihood and log-likelihood can be found in Appendix A.

In the model, I assume the following weak priors for the unknown parameters. The prior distribution of the mixture proportion vector ($\vec{\theta}$) is the Dirichlet distribution with a K-dimensional concentration parameter, where each component is assigned value 1. The prior distribution of the mean values of each cluster distribution is a Normal distribution with a mean of 0 and a variance of 100. Further, the prior distribution of the variance of each cluster distribution is an Inverse Gamma distribution with scale and shape parameters both set to 1.

Chapter 4

Datasets

The purpose of this chapter is to discuss the datasets that are compiled together as input for the model. The two main datasets used are precinct-level voting results and quantitative estimates of the ideologies of political candidates. I combine these two datasets using additional information about elections and geographic boundaries in the U.S. 2000 Census results and shapefile data.

I describe each dataset, as well as the process of acquiring and cleaning each of them. I then summarize the process to compile the separate datasets into one final dataset for each election year to be inputted into the model. This final dataset provides the names of candidates for each election in each precinct and the vote share per candidate.

4.1 Precinct Data

In this section, I describe the dataset of precinct-level voting results provided by the Harvard Election Data Archive [5]. As discussed, precincts are the finest granularity of the U.S. population with publicly accessible aggregated vote shares. The Harvard Election Data Archive provides the vote shares for the top Republican and top Democrat candidate of every state and federal election in the precinct that occurred between 2000 and 2012.

In this work, I focus on the U.S. Congressional elections within the states Texas and New York in 2006, 2008, and 2010. I did not include state election results because of the variety of state government positions among the states. I chose to use Congressional elections rather than the other two types of federal elections discussed in the previous section because most states have multiple Congressional Districts, and thus more Congressional candidates than Presidential or Senatorial candidates in an election year. Thus, assuming Congressional candidates have different political preferences, the additional candidate preferences should yield more information about voter preferences.

I chose to examine the election years 2006, 2008, and 2010 because they all have the same Congressional District geographic boundaries set by the 2000 U.S. Census results. I specifically chose to use election data of the states Texas and New York, as they are the second and third states with the largest population, respectively. Further, Texas is commonly known to be stereotypical conservative states, and New York a liberal one. In the three election cycles I will be focusing on, 65% of the Texas district elections were won by Republican candidates, and 80% of the New York district elections were won by Democratic candidates [39, 40, 41]. Analysis of both states allows me to test my model to infer distributions of preferences that are more centered around both conservative and liberal views.

4.2 Candidate Data

The other main dataset used for this project was a set of quantitative estimates of the political preferences of candidates within the Database on Ideology, Money in Politics, and Elections (DIME) by Adam Bonica [8]. Campaign-finance scores (CFscores) are one-dimensional quantitative estimates of the political ideology of political candidates, with -2 the most liberal score and +2 the most conservative.

Bonica developed these CFscores by utilizing the political ideology of elected officials to approximate the ideology of losing candidates who received campaign contributions from the same individual contributor [9]. While DW-NOMINATE scores are widely accepted measurements of ideology, Bonica goes one step further to approximate the ideology of unelected candidates [37]. Bonica assumes that individual contributors donate to political candidates with similar political ideologies. With this assumption, Bonica uses publicly available Political Action Committee (PAC) campaign funding data and datasets on the actions of elected officials to approximate an ideal point estimate for the unelected official as well.

The DIME provides a wealth of information on every political candidate in a local, state, and federal election from 1979 to 2012 [8]. The dataset includes an assigned CFscore for each candidate, as well as the name, party, and Congressional District (if applicable).

While the DIME provides data about every candidate in these elections, I only consider the data about the top candidate of the two main political parties in the U.S., namely the Democrats and the Republicans. I then match each candidate to the Democrat and Republican vote shares provided by the precinct-level voting results. In the Texas and New York elections were examining, I verified that these candidates selected from the DIME are also the candidates with the largest Republican and Democratic vote share according to the final election results posted by the New York Times [39, 40, 41].

Granted, some political scientists argue against basing a model on a two-party political system. First, the additional, smaller political parties choosing to participate, or not participate, in the election can skew the voting results for the two main parties. This could cause two-party methods to always overstating the vote share the two main parties will receive [25]. Second, some argue that there is information loss regarding election behavior and vote distribution by only considering two parties in the district election out of all of the potential national political parties [25].

Figure 4-1 plots the distributions of the CFscores of the political candidates in consideration and all political candidates in the Congressional elections in Texas and New York. Notice that the distributions visually are not very different from each other. This suggests that our model is not heavily affected by considering less candidates because the general distribution of preferences has been captured by the two main candidates.

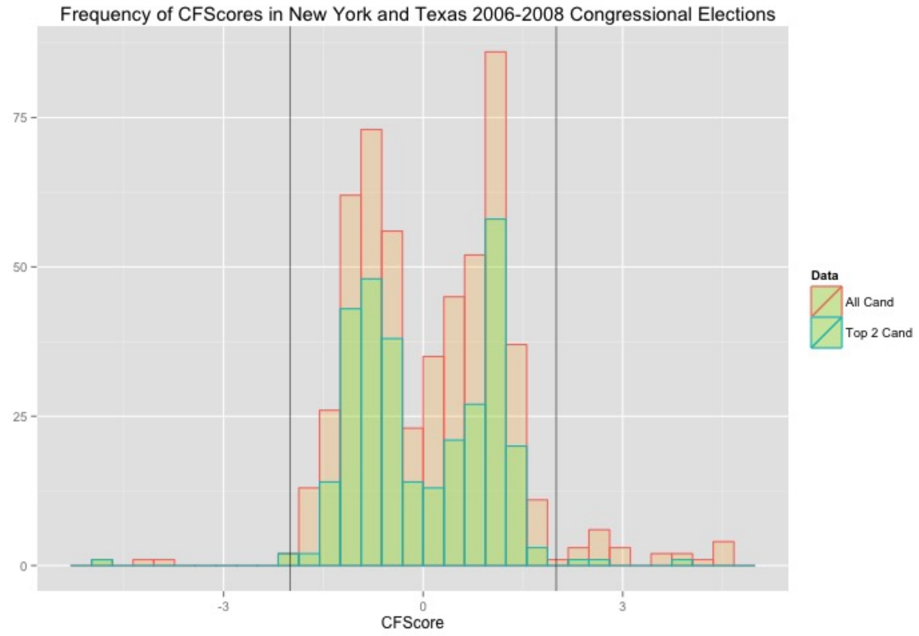

Figure 4-1: Histogram comparing CFScores of all candidate sand the two candidates in each precinct-level election.

4.2.1 Mapping Precincts to Congressional Candidates

In this section, I discuss the datasets used to connect the two datasets described in the earlier sections. So far, I have described two separate datasets, but there is a missing link still needed to assign precinct-level vote shares to each Democrat and Republican candidate in the election for the respective Congressional District.

I use U.S. 2000 Census Data to find the latitude and longitude of the geographic center of each precinct [11]. I also gather shapefiles of the geographic boundaries of Congressional Districts in the 2006, 2008, and 2010 terms [30]. I assigned any precincts to the Congressional District whose center fall within the specific Congressional District boundary lines. This process was done using the Geospatial Data Abstraction Library (GDAL/OGR) package within the Open Source Geospatial (osgeo) Python library.

Note that I assume that a precinct reports to a Congressional District if the geographic center of the precinct is located within the district boundaries. However, I got the same district assignments when mapping precincts to Congressional Districts by the geographic boundary of the precinct rather than the latitude and longitude of the center of each precinct. My method using latitude and longitude of each precinct was a faster computation for the same results.

This merged dataset allows me to analyze the distribution of candidate preferences across every precinct in the state of Texas and New York. One such plot demonstrating the value of this dataset merging political preferences and geographic information can be seen in Figure 4-2.

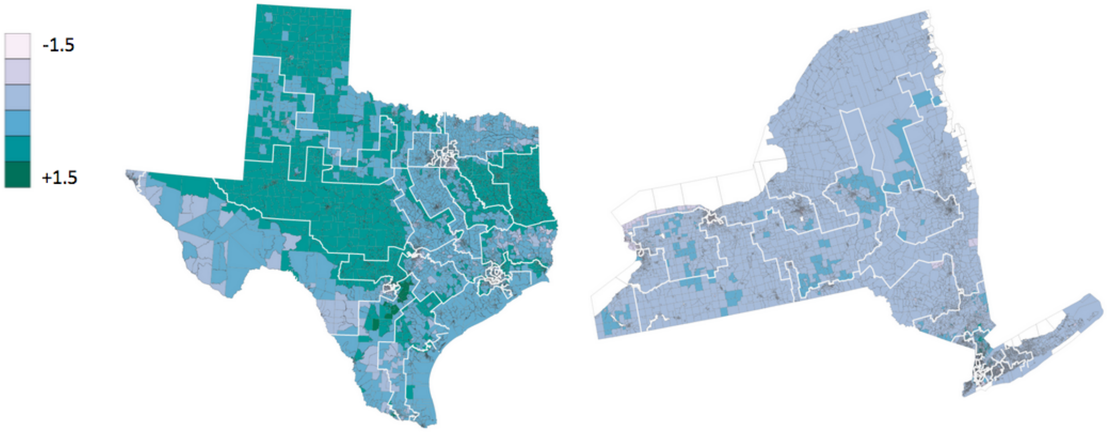

Figure 4-2: Chloropleth map that visualizes the weighted candidate CFScore values given the candidate‘s vote share in each precinct within the state of Texas.

Chapter 5

Experimental Results

The purpose of this section is to discuss the Bayesian inference results of the previously described model. To infer the unknown parameters in the model, I marginalize out the voter positions and use two different methods to infer the rest of the parameters conditional on the model and data described in the previous chapter. In section 5.2, I discuss the general inference practices and results in the context of simulated data. In section 5.3, I apply those same inference methods to actual precinct-level voting data that was introduced in chapter 4. In addition, I describe some general observations on the trends of political polarization within the U.S. electorate. These conclusions create a new perspective to the discussion on polarization that utilizes data about voters in more fine-grained detail than other related works.

5.1 Model Implementation

I ran two different Bayesian inference methods to approximate the underlying parameters of the model given a dataset of precinct-level election returns. The two methods I use to draw inferences from the model are a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) sampling algorithm and a Python function within the Scipy.optimize library that performs unconstrained minimization of multivariate scalar functions [**?**].

I wrote my own version of the Metropolis-Hastings algorithm in order to tailor the algorithm to our specific model and data (See Appendix for source code). This allows for further optimization, which is important because the Metropolis-Hastings algorithm is notoriously known for its long running time. I selected the approximate parameter values from the results of ten chains of the Metropolis-Hastings algorithm, where each chain is initialized randomly, rather than a single chain. As a result, I consider a wider range of approximate parameter values to find the set that yields the largest log-posterior. Further, the final parameter values are selected from all possible parameter values in the chain of parameter updates instead of only the final parameter values of each chain. From all of these possible sets of parameters from the Metropolis-Hastings algorithm, I select the set that yields the largest log-posterior value because it means the data fitted the model under those parameters.

I also inferred the underlying parameter values using an unconstrained minimization function within the Scipy.optimize Python library (Scipy.minimize) [**?**]. A second inference method is helpful to validate the inference results of both methods when the true underlying parameters are unknown. Similar to the Metropolis-Hastings algorithm, I ran the Scipy.minimize function 200 times on each dataset using the Powell method with random initialized starting points for each function call.[1] From the returned values of the 200 Scipy.minimize function calls, I again select the set of inferred parameters that yield largest log-posterior value.

The mixture proportion vector components and all cluster standard deviation parameters have specific bounds, so they had to be transformed to be unbounded parameters that could be inferred by the Scipy.minimize function. For each proportion vector component θ_j, where $j \in [1, \ldots, K]$ and K is the number of clusters, I instead use the substitute a_j such that $\theta_j = \frac{exp(a_j)}{\sum_{x=1}^{K} exp(a_x)}$. This ensures that all components of the mixture proportion vector satisfy the two necessary constraints, $0 \leq \theta_j \leq 1$ and $\sum_j \theta_j = 1$. Similarly, for each standard deviation of cluster j (σ_j), where $j \in [1, \ldots, K]$ and K is the number of clusters, I instead use b_j such that $\sigma_j = exp(b_j)$.

[1] explanation on the Powell method

5.2 Experiments on Simulated Data

I first ran both inference methods discussed previously on simulated data to ensure the functionality of the methods. The simulated data has similar properties to the true data, but predetermined parameters that can be used to verify the results of the inference methods with the predetermined underlying parameter values of the simulated data.

5.2.1 Generating Simulated Data

I generated simulated datasets to be similar to the actual precinct-level data used in order to construct the most realistic inference situation. The general idea of the data creation process of the simulated data is to regenerate the voter preferences and the number of votes per candidate based on true data of precinct elections gathered according to the details of Chapter 4. Specifically, the data generation process is initialized by a set of predetermined underlying parameter values for $\vec{\mu}, \vec{\sigma}$, and $\vec{\theta}$ and actual candidate CFScores and total number of voters for precinct-level elections of the same year and state. For each dataset about each real election, I randomly draw a cluster assignment from a Multinomial distribution that follows the predetermined mixture proportion. Given the cluster assignment, I randomly draw a point for the political preference of each voter that are i.i.d. distributed by a Normal distribution that follows the predetermined cluster parameters. I assigned votes to candidates according to the spatial voting rule described earlier in the context of the model. Namely, a voter will vote for the candidate in the election with more similar political preferences relative to his political preferences than the opposing candidate.

5.2.2 Inference Results Based on Simulated Data

This section describes the inference results of the two methods described in Section 5.1 based on simulated data generated as outlined in the previous section. For this section, I display the results given simulated data based on true data of the 2008 precinct-level Congressional elections in the state of Texas. I also assumed the simulated data is broken into four clusters. The other inference results given simulated data based on true data of other elections yield similarly accurate inference results. Table 5.1 demonstrates that these results from both inference methods are fairly accurate compared to the true underlying parameters.

	True Values	MH Inferences	Optim Inferences
Log-Posterior	-43724.873	-43724.899	-43726.723
θ_0	0.1	0.160	0.089
θ_1	0.2	0.265	0.240
θ_2	0.3	0.185	0.255
θ_3	0.4	0.391	0.417
μ_0	-1.5	-1.952	-1.128
μ_1	-1	-0.929	-0.799
μ_2	0	-0.001	-0.001
μ_3	1	1.022	0.924
σ_0	1	1.288	0.731
σ_1	1	0.920	0.806
σ_2	1	1.017	0.976
σ_3	1	1.035	0.942

Table 5.1: Final inferred parameter values and the corresponding log-posterior value generated by the Metropolis-Hastings algorithm (MH) and the Scipy.optimize function (Optim) methods.

The accuracy of these methods based on simulated data can also be viewed by comparing the posterior distributions with the true distribution of simulated preferences in Figure 5-1. This figure suggests that the inference methods were fairly accurate in inferring the underlying parameters of the distribution and are valid methods to use in our later experiments with real election data.

5.3 Experiments on Actual Data

The purpose of this section is to describe the results of the inference methods given real datasets of true precinct-level election results. As mentioned earlier, I infer results given six different elections, which are the Congressional elections of 2006, 2008, and 2010 within the states of Texas and New York.

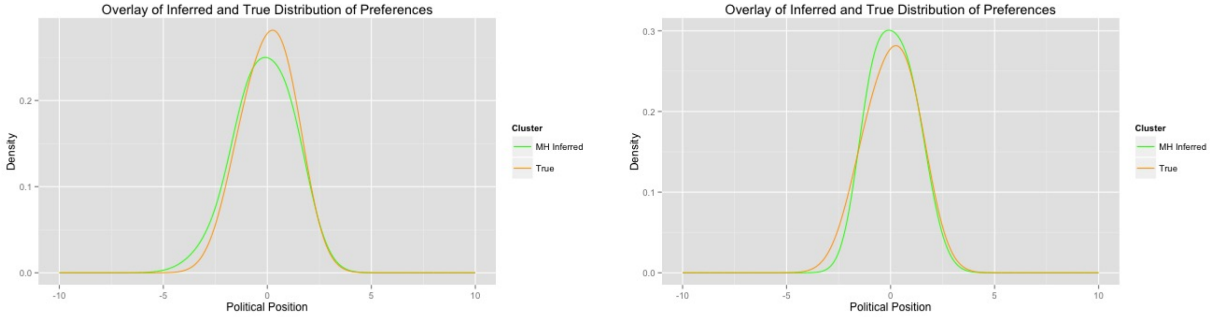

Figure 5-1: Posterior distribution given the set of inferred parameters against the true distribution of the simulated preferences.

5.3.1 Inference Results

I used both the Metropolis-Hastings and Scipy.minimize inference methods, but I chose to only focus on the results of the Metropolis-Hastings algorithm in this section. I did validate the results of both inference methods against each other and found that the log-posterior values were very similar for both inference methods. For many of these results, I also only display the inferences given the model assumes that the cluster number is four, which means that there are four possible clusters that precincts can be assigned to. Table 5.2 contains the final inferred parameter values for each of the six elections we are considering.

I also perform a posterior predictive check to better visualize and comprehend the inferred parameter values. In the posterior predictive check, I create simulated data given the set of inferred parameters and compare the distribution of simulated data to that of the true data. These posterior predictive checks for the Texas and New York Congressional elections in all three election cycles can be found in Figure 5-2 below. Note the inferred mixture proportion incorrectly assumes that all precincts in the state have the same number of voters. Thus, I recompute the mixture proportion by weighting the number of precincts assigned to each cluster by the number of voters in the precinct.

Overall, all of these inferred mixed Gaussian distributions seem to be unimodal, suggesting that the majority of the preferences of voters are still moderate. Further, the inferred posterior distribution of preferences of voters in the state of Texas seems

	Texas			New York		
	2006	2008	2010	2006	2010	2010
Post	-1600719.14	-3003988.275	-1765375.07	-1496355.685	-2500917.648	-1924061.809
θ_0	0.112	0.103	0.124	0.171	0.185	0.142
θ_1	0.292	0.246	0.269	0.205	0.088	0.263
θ_2	0.317	0.34	0.299	0.389	0.469	0.31
θ_3	0.279	0.312	0.308	0.236	0.258	0.285
μ_0	-1.294	-0.68	-4.38	-73.193	-14.632	-51.896
μ_1	-0.268	-0.096	-0.697	-2.277	7.086	-1.094
μ_2	1.128	0.545	1.212	-0.808	26.344	-0.382
μ_3	58.235	2.108	2.433	0.737	51.119	26.581
σ_0	1.348	0.93	4.181	55.43	86.591	32.375
σ_1	1.958	1.206	3.168	3.237	5.475	1.528
σ_2	3.433	1.504	3.188	3.468	116.206	2.68
σ_3	82.609	2.677	2.717	4.05	80.602	88.883

Table 5.2: This table contains the inferred parameter values of each election given a cluster number of four. Note the Post row represents the log-posterior value.

to become more broad, whereas the distribution in the state of New York seems to become more narrow. More discussion of the trends across the years in terms of political polarization will come in Section 5.4.

5.3.2 Inference Assuming K Clusters

Earlier, I was inferring the parameter values of the actual data under the assumption that there are only two possible clusters with precincts. I now discuss the accuracy of inferring parameter values of the actual data for various values of the cluster number (K), which represents the total number of clusters. While K can also be a parameter in our model, I leave it as a tuning parameter that must be set by the researcher based on his beliefs.

Figure 5-3 demonstrates that the log-posterior does increase as the cluster number increases. This suggests that an increase in possible clusters to assign precincts causes my model to better fit the provided data.

The posterior distributions of the inferred parameter values of the Metropolis-Hastings algorithm for every Texas election cycle and cluster number further suggest this idea. In Figure 5-4, note how the posterior distributions are less broad as the cluster number increases.

5.4 Applications of Results: Political Polarization

These experimental results are important to make claims about the political preferences of the U.S. voting population. In this section, I interpret the inference results based on the true data in the context of the discussion on political polarization within the U.S. voting population. Some political scientists hypothesize that the distribution of the political preferences of U.S. electorate is unimodal and comparatively moderate [19]. Yet other evidence suggests increasing polarization in the American population [31]. Given these differing pictures, the results from our method rather than data from the typical survey collection methodology seem desirable to add a third perspective.

These experimental results can be interpreted to answer many other open political science questions regarding the public opinion of the U.S. voting population. For instance, the results could be used to approximate the extent to which constituents are represented by their respective Congressmen. I arbitrarily chose to focus on the topic of political polarization as an example of the contributions in the political science field that the model and inference methods of this thesis yield.

5.4.1 Political Polarization within the Electorate

For this paper, I follow a multidimensional definition of polarization that was described by DiMaggio, Evans and Bryson [12]. The authors define polarization as the extent of preference disagreements over time [12]. This definition suits this work as the results yield distributions of preferences of different clusters of precincts over a six-year time period. Based on their definition, DiMaggio, et al. proceed to describe four separate causes of polarization, and quantitative metrics to understand the extent of each cause [12]. These causes are the principles of dispersion, bimodality, constraint, and consolidation within a population.

The first principle is the idea of dispersion, in which more dispersed opinions in the opinion increase the difficulty for a centrist political consensus, and thus the amount of polarization, to exist in the population [12]. Dispersion can be measured through the variance of the distribution of political preferences. An increase in variance signifies that voters have more extreme conservative or liberal political preferences and less moderate preferences in the middle of the distribution. Given the inference results and a cluster number of K, I compute the variance of the posterior distribution (σ) the following way:

$$\sigma^2 = \sum_{i=1}^{K} (\theta_i \sigma_i)^2 \tag{5.1}$$

Table 5.3 contains all of the variances of the inferred posterior distribution given the inference values of all six elections I analyzed. All of these posterior distributions correspond to the distributions plotted in Figure 5-2 that assumes a cluster number of 4. The variances for the Texas elections in Table 5.3 are generally decreasing, which suggests the principle of dispersion is decreasing in the population of Texas. On the other hand, the variances for the New York elections in Table 5.3 are generally increasing, which suggests the principle of dispersion is increasing in the New York electorate.

Election Cycle	Texas Variance	New York Variance
2006	533.51	93.07
2008	1.06	3659.95
2010	2.60	665.67

Table 5.3: Variances of the inferred posterior distribution given the data corresponding to the election cycle and state, as well as a cluster number of 4.

The second principle is the bimodality principle, which states that the increase in separate opinions of each group leads to a higher chance of social conflict. Bimodality can be measured with the kurtosis of the distribution. Intuitively, kurtosis can be viewed as a measure of the difference in positions of different clusters of points in a distribution. If a distribution is flatter and more bimodal, there is negative kurtosis. If the distribution centers around one peak and more unimodal, there is positive kurtosis. In Figure 5-2, qualitative observations indicate that inferred posterior distributions become more peaked over time. This suggests that the kurtosis is positive and increasing over the years, and that the preferences are not splitting apart.

The constraint principle states that groups with similar attitudes and political preferences resemble voting coalitions over time that vote similarly together and are less likely to vote according to another political group. This leads to more separation between voters, and thus more political polarization. The constraint principle can be measured by comparing the variances of specific clusters of precincts that are grouped due to similar distributions of political preferences of their constituents. If the individual cluster variances are low, the clusters are less likely to have similar opinions and vote similar to voters of other clusters. According to the results in Table **??**, the variances of each individual cluster seem to decrease. This suggests that specific groups of clusters are voting more similar to each other over the years and polarization is increasing in the electorate.

The consolidation principle is the idea that identity and demographic characteristics of voters are correlated to attitudes on social issues, and thus cause conflict between different groups with similar identities. It is difficult to use the experimental results to quantitatively measure the consolidation principle because the model does not consider voters in specific groups based on identity or demographic similarities. The consolidation principle should be further investigated in future work by developing a new model that accounts for additional demographic data.

Analysis of the individual factors related to political polarization outlined by DiMaggio, et al. suggests different stories about the trends of polarizations over the three election cycles [12]. Analysis of the preferences of the voters of Texas in terms of the dispersion and constraint principles suggest the Texas electorate is becoming more polarized, whereas analysis of the same voters in terms of the bimodality principle suggests the Texas electorate is becoming more moderate. Similarly, analysis of the preferences of the voters of New York in terms of the dispersion and bimodality principles suggest the New York electorate is becoming less polarized, whereas analysis in terms of the constraint principle suggests the opposite.

One possible explanation of my results is that while voters within clusters are behaving more similarly than before, the overall distributions of the clusters themselves are shifting. In the state of Texas, these cluster distributions are shifting closer together and having more similar distributions of preferences to each other. In the state of New York, on the other hand, these cluster distributions are shifting further apart. It would be interesting in further research to investigate political or economic events in those states between 2006 to 2010 that may have caused these changes.

5.4.2 The Electorate vs. the Political Elite

Furthermore, I want to understand the extent to which polarization exists in the electorate compared to within the political elite. I investigate this question through qualitative comparisons of the known distributions of political candidate preferences and inferred distribution of voter preferences. To do this, I create a plot that compares a normalized histogram of the candidate positions against the inferred posterior distribution. These plots given each state and election cycle analyzed can be found in Figure 5-5.

In general, the unimodal distributions of preferences of the electorate are centered around the center of the bimodal distribution of the candidates‘ political positions. This suggests that the most common political preference of voters is the moderate view between the average preferences of the two main clusters of candidates.

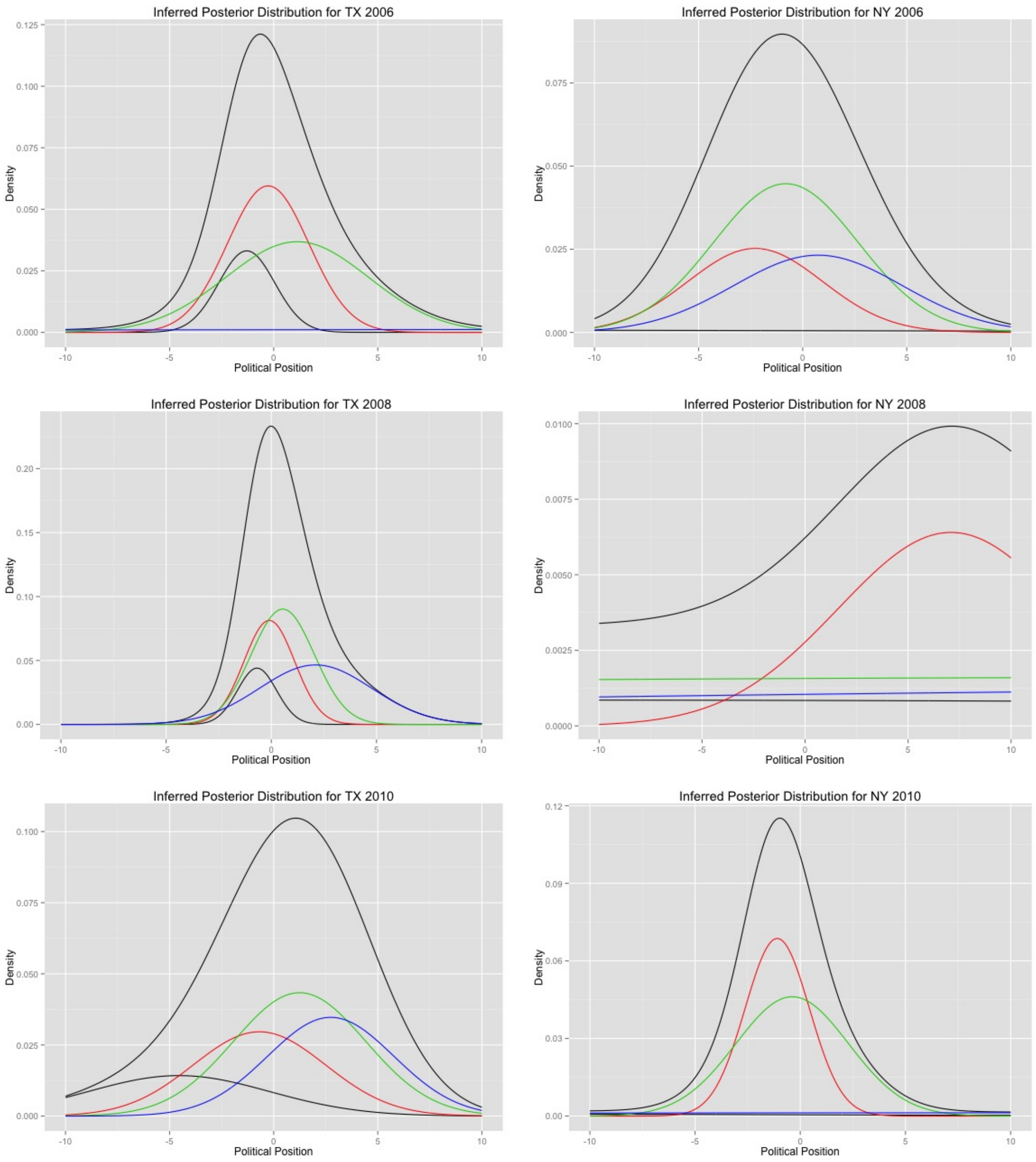

Figure 5-2: Inferred posterior distributions (in black) based on the data of the 2006, 2008 and 2010 U.S. Congressional elections in Texas and New York. Individual cluster distributions for each election are the colored lines.

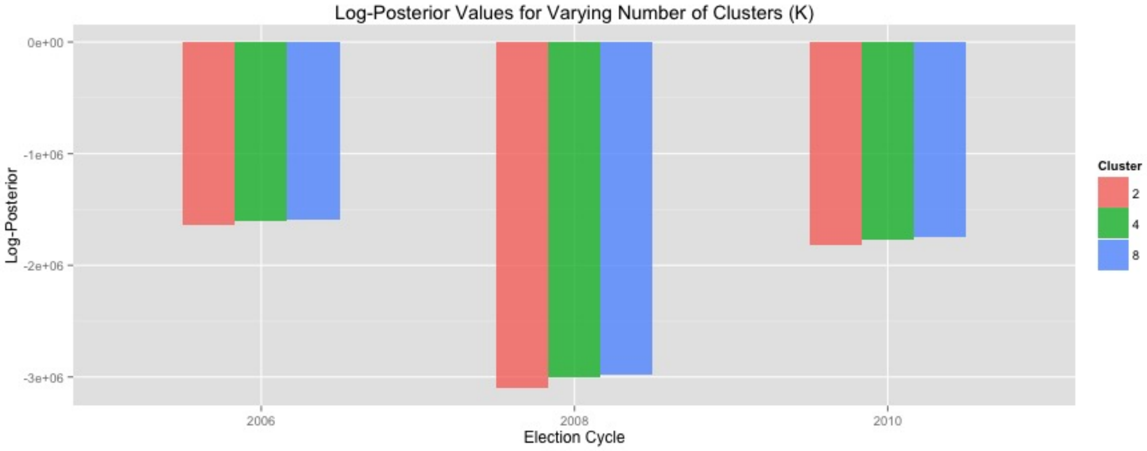

Figure 5-3: Bar plot of the log-posterior values for Texas precinct-level election in the 2006, 2008 and 2010 election cycles with varying number of clusters assumed.

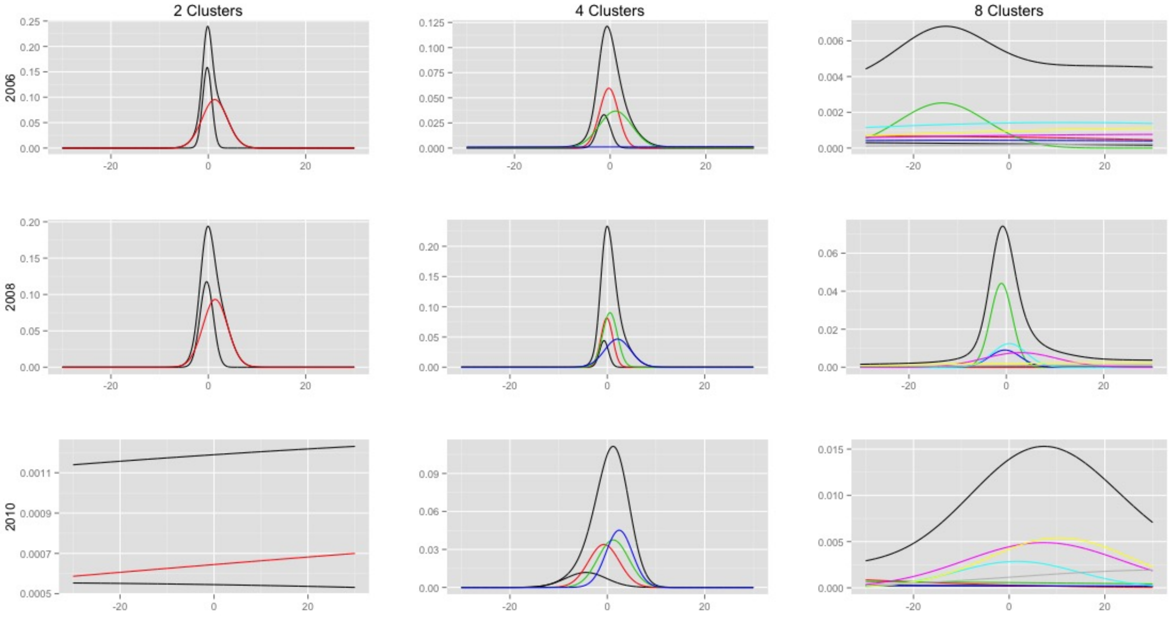

Figure 5-4: Inferred distributions for Texas precinct-level election in the 2006, 2008 and 2010 election cycles with varying number of clusters assumed. In each subplot, the x-axis is the one-dimensional preference space and the y-axis is the density.

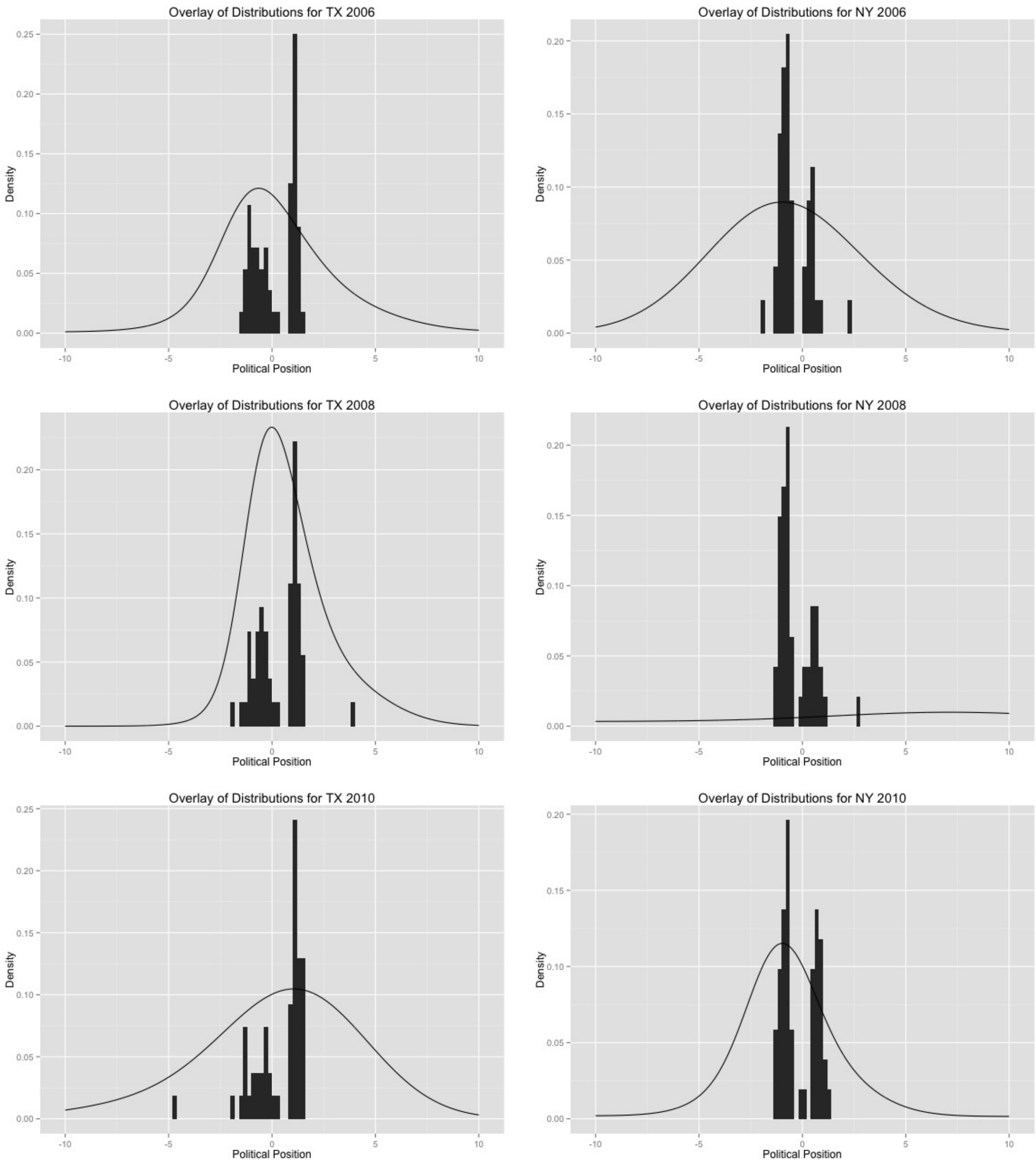

Figure 5-5: Candidate CFScores as a normalized histogram overlaid on the inferred distribution of the political preferences of the U.S. voting population as a smooth curve.

Chapter 6

Validation

The purpose of this section is to discuss the validity of our model and inference results based on real datasets that were discussed in earlier chapters. In Section 6.1, I compare my results with other methods of measuring public opinion on political issues. I then investigate the accuracy of my results to predict the outcomes of unrelated elections and other political events dependent on political preferences of precincts. I also validate my results by comparing the political preferences of voters to their elected representatives in Section 6.3.

6.1 Value Comparison to Related Works

One validation of my results is to compare it to the results of alternative methods that measure the political preferences of the U.S. voting population. In this section, I will compare my results to two alternative sets of results. First, I will compare my results to the results of the Cooperative Congressional Election Study of the public opinion of the U.S. population [6]. Second, I will compare my results to the approximated district-level preferences computed by Chris Tausanovitch and Christopher Warshaw [38].

To compare my results with these, I had to first obtain a single-point estimate of the preference of each Congressional District in Texas and New York from my precinct-level estimates. To do this, I first select an assignment variable for each precinct such that the mean and standard deviation of the cluster associated with that assignment variable maximizes the likelihood of the precinct‘s election results.

$$\text{Let: } \phi = \text{NormCdf}\left(\frac{c_{i,0} + c_{i,1}}{2}, \mu_{z_i}, \sigma_{z_i}\right) \tag{6.1}$$

$$p_i = P(z_i \,|\, \vec{X}_i, \vec{\theta}) \propto P(\vec{X}_i \,|\, z_i)P(z_i \,|\, \vec{\theta}) = \text{BinomPdf}(X_{i,0}, X_{i,0} + X_{i,1}, \phi) \times \theta_{z_i} \tag{6.2}$$

I assume that each precinct has the same single-point estimate of preferences as its assigned cluster. Thus, I can approximate the single-point estimate of the preferences of each district by taking the weighted average of the individual estimates of each precinct within that district given the proportion of voters in the precinct to the district population overall.

6.1.1 Comparison to Survey Results

Analyzing survey data is a common method used by political scientists to understand the public opinion of the U.S. population. I assume that most survey respondents are within the voting age, and thus can easily compare my results about the U.S. voting population with survey results about the general population.

The Cooperative Congressional Election Study (CCES) surveys over 50,000 Americans throughout the country every election year [6]. The CCES also requests each survey respondent to report his Congressional District, while most other national surveys of the public only ask respondents to report their state of residence [6]. The association of district for each respondent helps me compare the more fine-grained results of my model.

I specifically focus on two questions within the CCES. The first asks the survey respondent to determine their ideology given a set of seven discrete possible choices.

> Thinking about politics these days, how would you describe your own political viewpoint?
>
> - Very Liberal
> - Liberal
> - Moderate

- Conservative
- Very Conservative
- Not sure

The second question I compare my results to asks the survey respondent to provide a score of his own ideology on a continuous scale from 0 to 100.

> One way that people talk about politics in the United States is in terms of left, right, and center, or liberal, conservative, and moderate. We would like to know how you view the parties and candidates using these terms. The scale below represents the ideological spectrum from very liberal (0) to very conservative (100). The most centrist American is exactly at the middle (50). Where would you place yourself?

This results of second question are more ideal to compare with my results to because my results are also on a continuous spectrum. However, this question was only asked in the CCES surveys of 2006 and 2008, so I was unable to validate my results of the 2010 election cycle.

I compare the my inferred distributions of state voting populations with the responses to these two questions by district. Figure 6-1 displays these comparisons as a scatter plot for the elections I analyzed within the state of New York. The figure only displays results given data about elections in New York, but the other validation results were similar.

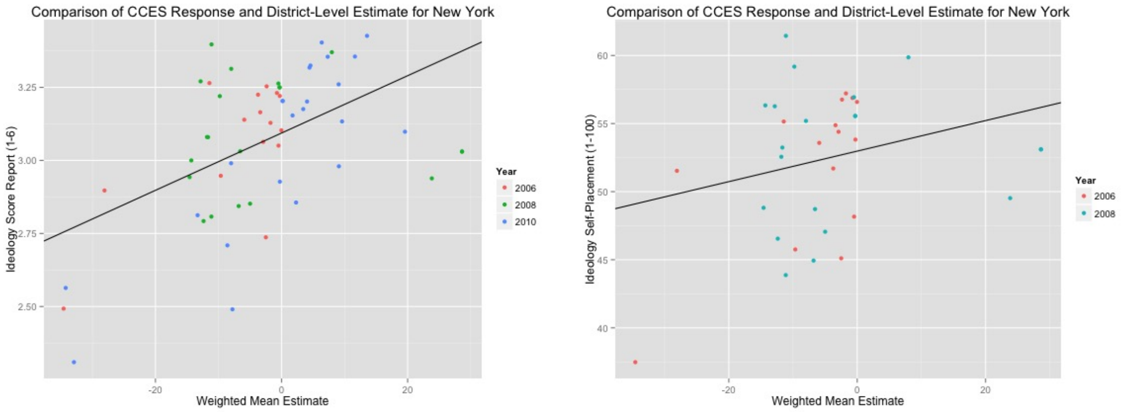

Figure 6-1: Scatter plots comparing weighted mean estimates to CCES survey responses about ideology.

6.1.2 Comparison to MRP Results

In addition to survey results, I thought it would be useful to validate my results against inferred ideological scores based on survey results. I justified that these ideological scores might be more representative of the U.S. population because the inference methods factor in possible sampling bias.

Specifically, I chose to compare my results to the ideological scores of Tausanovitch and Warshaw that estimate of mean policy preferences of Congressional Districts using disaggregation and multi-level regression with post-stratification (MRP) on survey data from the decade 2000 to 2010 [38]. The work of Tausanovitch and Warshaw is one of the more recent works related to understanding preferences and analyzes election years similar to the ones I analyze [38]. Furthermore, their method is based on the idea of merging multiple survey results together in order to have a better sampling of the full U.S. population [38].

To compare my inference results to the results of Tausanovitch and Warshaw, I had to consolidate my district-level results across the three election cycles into one district-level result. This is a more accurate comparison to their results based on five election cycles, which include the three that I analyze. Figure 6-2 suggests that there is a strong relationship between my estimates and their results. In fact, running linear regressions for both sets of variables finds that the positive trends between our inferred results for each state and the MRP results are statistically significant with p-values less than 0.01.

6.2 Prediction Capabilities and Accuracy

The purpose of this section is to discuss the accuracy of my model in various prediction tasks of alternative political outcomes. It is important to analyze the models ability to predict outcomes of elections that are not factored into the model, but are still dependent on precinct ideology scores. If the new model can predict other outcomes well, this implies that my inferred estimates are not overfitted to a single election outcome and are valid measurements of political preferences of precincts for

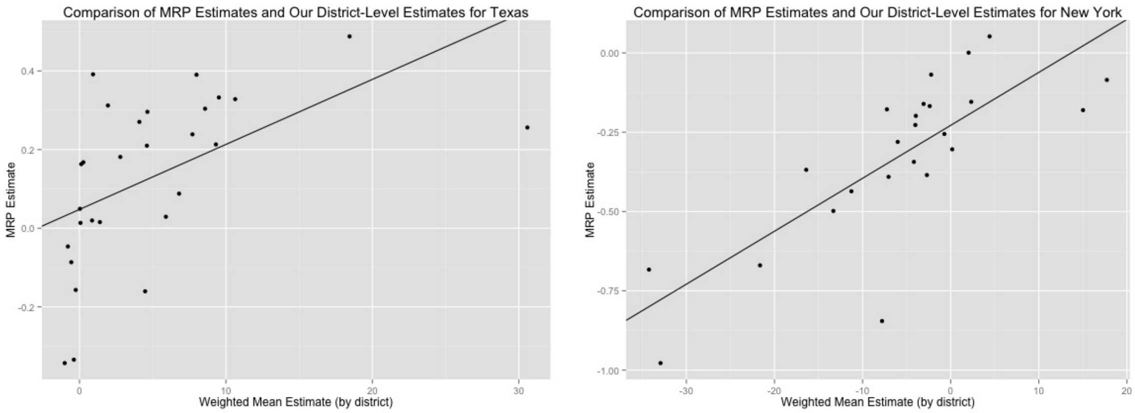

Figure 6-2: Scatter plots comparing weighted mean estimates across all three election cycles to the results of the MRP model.

alternative political outcomes.

6.2.1 Baseline Prediction Tests

Prior to using estimated values for prediction tests, I first use empirical data in some baseline prediction tests. The baseline tests are to ensure that my estimated values are not yielding good predictive performance because the prediction task is too simple. The empirical data I am considering are the vote share per major party candidate, the two candidates ideology scores, the midpoint ideology score, and total number of voters per precinct.

One naive baseline for comparison is to assume the candidates of the same party receive the same number of votes from one election to the next election. For instance, in order to predict the vote share of the Democratic candidate in 2010, the naive baseline assumes the vote share is the same as the vote share of the Democratic candidate from 2008. The results of this naive baseline task against the actual vote shares can be seen in Table 6.1.

6.2.2 Predicting Following Elections

Similar to the naive baseline test outlined in the previous subsection, I measure the accuracy of my predictions of the outcomes of later Congressional elections based on my estimates that took in as input an earlier set of Congressional election vote shares. As discussed earlier, this baseline test given the empirical data is nontrivial, and it would be significant if my estimated values yield more accurate predictions.

Specifically, I used the inferred results of my model given data of one election cycle to predict the outcomes of the following election cycle. For each precinct, I compute the midpoint between the political preferences of the two candidates participating in each precinct election in the following election cycle. I then predict that the more liberal candidate (c_{0i}) for each precinct i of the next election will receive the percentage of votes determined by the cumulative normal distribution given the midpoint value and the cluster parameters inferred from the previous election. Because there are only two candidates per election, the remaining candidate (c_{1i}) for each precinct i is assigned the remaining percentage of votes. I then compute the number of votes for each candidate by multiplying the percentage of votes for the candidate and the total number of voters in the precinct. Finally, I sum the number of votes each candidate receives from all of the precincts in the same Congressional District.

Table 6.1 contains the sum of squared error between the predicted and actual vote share for the more liberal candidate (c_0) for each election for the predicted scores yielded by the baseline and prediction tests outlined.

		2006 → 2008	2008 → 2010
Texas	Baseline	0.010	0.155
	Prediction	0.027	0.229
New York	Baseline	0.046	0.156
	Prediction	0.035	0.058

Table 6.1: Table of the error terms for the baseline and prediction tests for inferred results given the Texas and New York data and a cluster number of 4.

These prediction results tend to outperform the results of the baseline prediction tests for each set of inferences, which suggests that my model yields more informative information about the behavior of voters. A visualization of the comparison between the predicted vote share and the actual vote share of the elections in 2008 can be seen in Figure 6-3. Note that the predicted vote share is based on inferred parameters given the cluster number is 4 and the precinct-level election results data of Texas and New

York in the 2006 election cycle. Similar to the results of the baseline test, the results given the other election data are similar to the ones in Figure 6-3 and can be viewed in Appendix B.

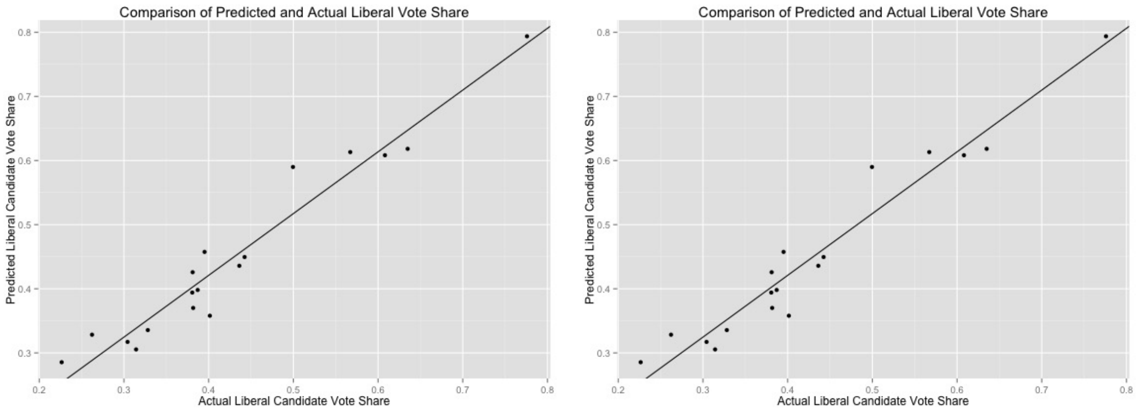

Figure 6-3: Scatter plots comparing prediction results to actual vote share of elections of the 2008 cycle in Texas and New York.

6.3 Comparisons to Political Candidates

Yet another way to validate my results is by comparing them to measurements of other political events or groups that are related to political preferences of voters. The purpose of this section is to validate my results of the political preferences of voters against the political preferences of their supposed representatives. In an ideal government, the elected officials should be representing their constituents and should have similar political preferences.

Moreover, this method is also effective in allowing me to compare my results to the estimates developed by Levendusky, et al.[29]. While Levendusky et al. does have the most similar model to mine, the work analyzes election data from the years 1950 to 1990, whereas I analyze election data from 2006 to 2010. Thus, I thought I could compare my results to those of Levendusky, et al. by comparing the results of using the same validation technique of comparing my results to the political preferences of elected officials.

A scatterplot of the precinct partisanship score against the legislative ideal points can be seen in Figure 6-4. Note that the weighted mean estimates of the 2006 Texas election results were removed because the range of inferred mean values was much larger than the range of the inferred means of the 2008 and 2010 Texas election results. This caused the weighted mean estimates of the 2006 Texas election to have a different range of weighted mean estimates that did not align well with the other results and incorrectly skewed the plot. Both scatter plots in Figure 6-4, as well as the weighted mean values of the 2006 Texas election and the DW-NOMINATE scores, have statistically significant relationships between the two variables with a p-value less than 0.001.

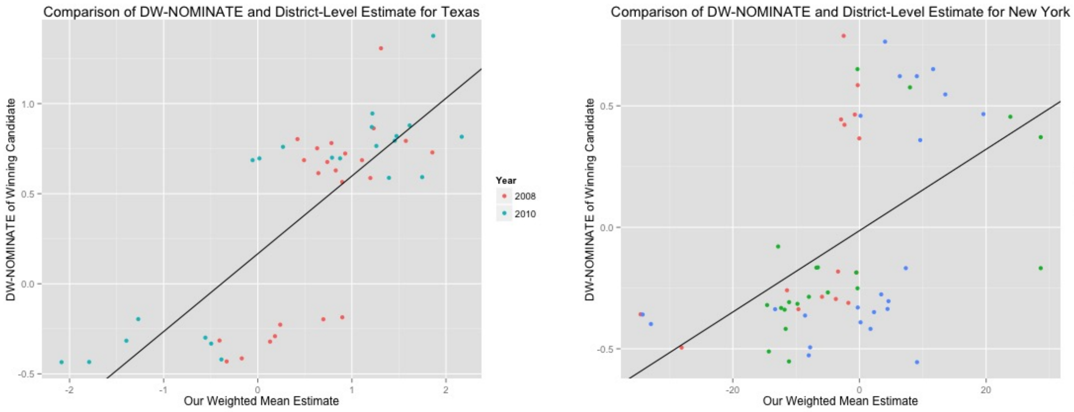

Figure 6-4: Scatter plots comparing district-level inferred results to DW-NOMINATE scores of politicians of the corresponding district.

Levendusky, et al. also found a strong correlation between their results and the quantitative ideology scores of candidates with a similar validation technique [29].

Chapter 7

Concluding Thoughts

In this work, I develop a new method to approximate the political preferences of the U.S. electorate. I use a model-based approach that combines precinct-level voting data and candidate position estimates to infer likely distributions of preferences of the general population.

My method has limitations complementary to those of surveys. Rather than having bias driven by non-response, the biases of my approach come from voter abstention and modeling assumptions. In contrast to surveys, my approach directly estimates the preferences of voters since it is based on actual casted votes. My approach also benefits from a large number of estimates of political preferences at fine-grained spatial resolution, which is usually not feasible from survey-based approaches.

I then apply my model to six different datasets spanning three election cycles, 2006, 2008, and 2010, and two states, Texas and New York. The Metropolis-Hastings algorithm infers posterior distributions that yield an ambiguous answer to the current debate over the state of polarization in the U.S. voting population in terms of the definition of polarization introduced by DiMaggio, et al. [12]. The results do suggest trends of certain principles of polarization changing over the timespan I analyze, but the trends sometimes conflict with each other. My ambiguous results could explain why the debate over the existence of polarization still exists to this day. Political scientists studying this problem have different definitions of polarization. These different definitions and perspectives, such as the different principles of polarization that I analyzed in my results, can yield diverging opinions on polarization by different researchers.

7.1 Future Work

In this work, I have introduced a spatial voting model and applied inference results to better understand the extent of the existence of political polarization within the U.S. voting population. There are many other possible extensions of this work, ranging from improving the model to applying the inference results to other questions about the opinions of the U.S. voting population.

7.1.1 Weaving in Demographics

As of now, this model only works for observed data involving elections with exactly two candidates. I would like to improve the model to also account for data about elections each with only one uncontested candidate. Levendusky, et al. has addressed the issue of missing data due to elections with uncontested candidates by adding an additional term to the mean of the posterior distribution of the log-odds of the Democratic candidate winning [29]. This additional term accounts for information about the precincts election such as including an incumbency offset or an uncontested candidate offset. The model could also be updated to include an offset to account for the number of political parties in the election, which can address the bias of results based on a two-party system that some political scientists voiced concern over.

7.1.2 Congressional Redistricting

Given the ability to infer these distributions of political preferences, another extension of this project is to examine the effects of Congressional redistricting. Redistricting is the process of assigning geographic boundary lines to districts of members of the U.S. House of Representatives. Every ten years, the districts each member of the U.S. House of Representatives represents is redrawn based on the updated U.S. population information from the recent results of the U.S. Census. The makeup of constituents in the district is important for the incumbent, as well as the incumbent's political party, because the constituents ultimately decide who has the vote in policy issues in Congress. Thus, Congressmen often meddle with redistricting through a process called gerrymandering, where voters of similar party alliances tend to be in the same district through a manipulation of geographic boundaries.

Precincts can be considered as the building blocks for congressional districts, so I can use the precinct preference estimates to predict how different redistricting proposals could affect the makeup of Congress. I hope to create more fair redistricting suggestions, as well as more plausible options for legislatures to consider, based on observations by political scientists [**?**].

7.1.3 Ecological Inference

One final extension of this project could be to perform ecological inference on my precinct-level estimates in order to understand individual preferences. Ecological inference is the process of inferring individual behavior from aggregated data. If ecological inference can be properly done, this would make a large contribution to understand the political preferences of a wide variety of individual voters in the U.S. This is useful to many political campaigns and advertising systems that can then better target individuals within the population.

Appendix A

Posterior Distribution Derivation

Part of the posterior distribution was described in Chapter 3.2. In this appendix, I will go into more detail of the math behind the posterior distribution.

As a reminder, in Chapter 3.2, I leave off with the posterior distribution defined in terms of prior distributions and the likelihood.

$$P(\vec{\mu}, \vec{\sigma}, \theta | \vec{v}) = P(\vec{\mu}) P(\vec{\sigma}) P(\vec{v} | \vec{\mu}, \vec{\sigma}, \theta) \tag{A.1}$$

The prior distributions have already been discussed in Chapter 3.2, but I did not include how I compute the likelihood term given a set of observed data and proposed set of unknown parameters. The equation for the likelihood can be found below. Let us define $BinomPdf(x, n, p)$ as the probability distribution function of the Binomial distribution, where x is the number of successes, n is the total number of trials, and p is the probability of success. Further, let us define $\Phi(x, \mu, \sigma)$ as the cumulative distribution function of the Normal distribution, where x is a value, μ is the mean, and σ is the standard deviation. The rest of the variables in the equation below are the same as what was defined in Section 3.4.

$$P(\vec{v}|\vec{\mu},\vec{\sigma},\theta) = \prod_{j=1}^{M}\left[\sum_{x_j=1}^{K} P(\vec{v_j}|x_j,\mu_{x_j},\sigma_{x_j})P(x_j|\theta)\right] \tag{A.2}$$

$$= \prod_{j=1}^{M}\left[\sum_{x_j=1}^{K} BinomPdf(v_{j,0}, v_{j,0}+v_{j,1}, \Phi(\frac{c_{j,0}+c_{j,1}}{2},\mu_{x_j},\sigma_{x_j})\theta_{x_j}\right] \tag{A.3}$$

For numerical stability reasons, I determine the log-likelihood. Given the above equation for the likelihood of the parameters given the data, the equation for the log-likelihood is below. Let us define Φ_{j,x_j} to be a function that returns $\Phi(\frac{c_{j,0}+c_{j,1}}{2},\mu_{x_j},\sigma_{x_j})$, where Φ is the function defined earlier.

$$logP(\vec{v}|\vec{\mu},\vec{\sigma},\theta) = \sum_{j=1}^{M} log\left[\sum_{x_j=1}^{K}\left(BinomPdf(v_{j,0}, v_{j,0}+v_{j,1}, \Phi(\frac{c_{j,0}+c_{j,1}}{2},\mu_{x_j},\sigma_{x_j})\right)\theta_{x_j}\right] \tag{A.4}$$

$$= \sum_{j=1}^{M} log\left[\sum_{x_j=1}^{K}\left(\binom{v_{j,0}+v_{j,1}}{v_{j,0}}{\Phi_{j,x_j}}^{v_{j,0}}(1-\Phi_{j,x_j})^{v_{j,1}}\theta_{x_j}\right)\right] \tag{A.5}$$

$$= \sum_{j=1}^{M} log\left[\sum_{x_j=1}^{K}\left(\binom{v_{j,0}+v_{j,1}}{v_{j,0}}e^{v_{j,0}log(\Phi_{j,x_j})}e^{v_{j,1}log(1-\Phi_{j,x_j})}e^{log(\theta_{x_j})}\right)\right] \tag{A.6}$$

$$= \sum_{j=1}^{M} log\left[\binom{v_{j,0}+v_{j,1}}{v_{j,0}}\sum_{x_j=1}^{K}\left(e^{v_{j,0}log(\Phi_{j,x_j})+v_{j,1}log(1-\Phi_{j,x_j})+log(\theta_{x_j})}\right)\right] \tag{A.7}$$

Appendix B

Additional Validation Results

B.1 Prediction Figures

As Table 6.1 indicates, I also ran the prediction tests on the inferred parameters based on election data in 2008 to predict the election results in 2010. The two additional plots depicting these results can be seen below in Figure B-1.

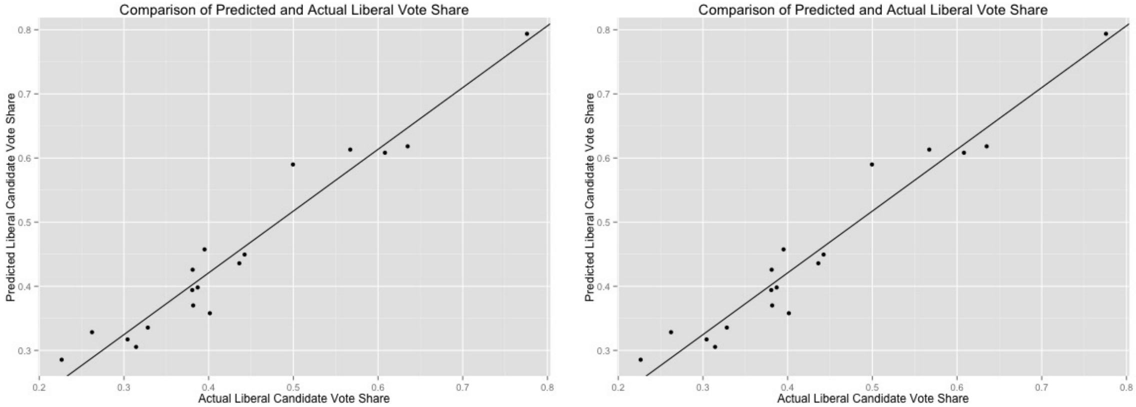

Figure B-1: Scatter plots comparing prediction results to actual vote share of elections of the 2008 cycle in Texas and New York.

Bibliography

[1]

[2]

[3]

[4] S. Ansolabehere and E. Hersh. Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. *Political Analysis*, 2012.

[5] S. Ansolabehere, M. Palmer, and A. Lee. Precinct-Level Election Data, 2014. http://hdl.handle.net/1902.1/21919.

[6] S. Ansolabehere and S. Pettigrew. Cumulative CCES Common Content (2006-2012).

[7] L.M. Bartels. Beyond the Running Tally: Partisan Bias in Political Perceptions. *Political Behavior*, 24(2):117–150, 2002.

[8] A. Bonica. Database on Ideology, Money in Politics, and Elections: Public version 1.0, 2013. http://data.stanford.edu/dime.

[9] A. Bonica. Mapping the Ideological Marketplace. *American Journal of Political Science*, 58(2):367–386, 2014.

[10] C.T. Brass. Shutdown of the Federal Government: Causes, Effects, and Process. *Congressional Research Service*, pages 14–18, 2013.

[11] United States Census Bureau. TIGERweb State-Based Data Files: Voting Districts - Census 2010, 2010. http://tigerweb.geo.census.gov/tigerwebmain/TIGERweb_vtd_census2010.html.

[12] P. DiMaggio, J. Evans, and B. Bryson. Have American's Social Attitudes Become More Polarized? *American Journal of Sociology*, pages 690–755, 1996.

[13] Indiana Election Division. 2001 Indiana Congressional Districts - Repealed January 3, 2013, 2006. http://www.in.gov/sos/elections/3007.htm.

[14] A. Downs. *An Economic Theory of Democracy.* Harper and Row, 1957.

[15] J. Enelow and M. J. Hinich. Nonspatial Candidate Characteristics and Electoral Competition. *The Journal of Politics*, 44(1):115–130, 1982.

[16] R.S. Erikson. Constituency Opinion and Congressional Behavior: A Reexamination of the Miller-Stokes Representation Data. *American Journal of Political Science*, 22(3):511–535, 1978.

[17] J.H. Evans. Have Americans' Attitudes Become More Polarized? An Update*. *Social Science Quarterly*, 84(1):71–90, 2003.

[18] S.L. Feld and B. Grofman. Incumbency Advantage, Voter Loyalty, and the Benefit of the Doubt. *Journal of Theoretical Politics*, 3(2):115–137, 1991.

[19] M. P. Fiorina and S. J. Abrams. Political Polarization in the American Public. *Annual Review of Political Science*, 11:563–588, 2008.

[20] D. Green, B. Palmquist, and E. Schickler. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. 2002. **read once more?**

[21] B. Groman. Downs and Two-Party Convergence. *Annual Review of Political Science*, 7:25–46, 2004.

[22] M.H. Hetherington. Review article: Putting Polarization in Perspective. *British Journal of Political Science*, 39(2):413–448, 2009. **read once more?**

[23] M.J. Hinich. Equilibrium in Spatial Voting: The Median Voter Result is an Artifact. *Journal of Economic Theory*, 16:208–219, 1977.

[24] M.O. Jackson. An Overview of Social Networks and Economic Applications. *The Handbook of Social Economics*, 1:511–585, 2010.

[25] J. Katz and G. King. A Statistical Model for Multiparty Electoral Data. *American Political Science Review*, 93(1):15–32, 1999.

[26] G. Kernell. Giving Order to Districts: Estimating Voter Distributions with National Election Returns. *Political Analysis*, 17(3):215–235, 2009.

[27] G. H. Kramer. Short-Term Fluctuations in U.S. Voting Behavior, 1896-1964. *The American Political Science Review*, 65(1):131–143, 1971.

[28] S.J. Lepper. Voting Behavior and Aggregate Policy Targets. *Public Choice*, 18(1):67–81, 1974.

[29] M.S. Levendusky, J.C. Pope, and S.D. Jackman. Measuring District-level Partisanship with Implications for the Analysis of US Elections. *The Journal of Politics*, 70(3):736–753, 2008.

[30] J. B. Lewis, B. DeVine, L.Pitcher, and K. C. Martis. *Digital Boundary Definitions of United States Congressional Districts, 1789-2012*, 2013. http://cdmaps.polisci.ucla.edu.

[31] N. McCarty, K.T. Poole, and H. Rosenthal. *Polarized America.* The MIT Press, 2006.

[32] M. P. McDonald. 2014 November General Election. *United States Election Project.*

[33] M. McPherson, L. Smith-Lovin, and J.M.Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, pages 415–444, 2001.

[34] W.E. Miller and D.E. Stokes. Constituency Influence in Congress. *American Political Science Review*, 57(1):45–56, 1963.

[35] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas. Vocal Minority Versus Silent Majority: Discovering the Opinions of the Long Tail. *IEEE Third International Conference on Social Computing*, 2011.

[36] J.S. Nye, P. Zelikow, and D. C.King. *Why People Don't Trust Government*, chapter 6, pages 155–178. Harvard University Press, 1997.

[37] K.T. Poole and H.L. Rosenthal. A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*, 29(2):357–384, 1985.

[38] C. Tausanovitch and C. Warshaw. Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities. *The Journal of Politics*, 75(2):330–342, 2013.

[39] New York Times. Election 2006, 2006. http://www.nytimes.com/ref/elections/2006/House.html.

[40] New York Times. Election Results 2008, 2008. http://elections.nytimes.com/2008/results/house/map.html.

[41] New York Times. Election 2010, 2010. http://elections.nytimes.com/2010/results/house.

[42] C. Warshaw and J. Rodden. How Should We Measure District-Level Public Opinion on Individual Issues? *The Journal of Politics*, 74(1):203–219, 2012.

[43] A. Watts. The Influence of Social Networks and Homophily on Correct Voting. *Network Science*, 2(1):90–106, 2014.

[44] Y. Zhang, A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha. Community Structure in Congressional Cosponsorship Networks. *Physica A*, 2008.