Essays on Latent Variable Models and Roll Call Scaling

by

Kevin A. McAlister

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Political Science)
in The University of Michigan
2020

Doctoral Committee:

Professor Walter R. Mebane, Co-Chair
Professor Kevin M. Quinn, Co-Chair
Assistant Professor Christopher J. Fariss
Professor Scott E. Page

Kevin A. McAlister

kamcal@umich.edu

ORCID iD: 0000-0002-4797-0384

For my wife, Dana. Your love and support has given me the energy to complete this process.

ACKNOWLEDGEMENTS

This dissertation symbolizes the end of a long and difficult journey. There is no such thing as a self-made man, and I am no exception. My success is a product of the time and support of many people throughout my young career.

First, I want to thank the University of Michigan political science department for providing me with all the tools I needed to succeed in graduate school and beyond. I thank numerous members of the administrative staff in the department that dealt with all of the bureaucratic work needed for my advancement through the program. Special thanks to Lisa Disch and Robert Mickey for ensuring that I had all the resources I needed to continue my studies at UM.

Second, I want to thank the numerous colleagues and friends that helped me to reach this point. Thanks to the many people that have listened to my work and provided feedback throughout this process. Thanks to my collaborators Erin Cikanek, Hwayong Shin, Diogo Ferrari, and Patrick Wu for your support and invaluable work in developing a number of wonderful projects. Thanks to Jacob Montgomery for supporting my interest in latent variable models and helping me to develop my current research interests.

Third, I want to thank my committee. Each and every member has served as an important mentor throughout the beginnings of my academic career. Thanks to Scott Page, who taught me more about careful and conscientious teaching of difficult

topics than any other person. Thanks to Christopher Fariss, who taught me about the ins and outs of academia and provided thoughtful comments on all of my project ideas, good and bad. Thanks to Kevin Quinn, whose guidance has been invaluable and has always ensured that my work meaningfully interacts with Bayesians, old and new.

A special thanks is warranted for Walter Mebane. From my first day in Ann Arbor, Walter has been the most supportive mentor and friend one could hope for during the Ph.D. process. In my five years of working with Walter, I have learned more about statistics, political science, effective teaching, and academia than I could have ever imagined coming into the program. Walter gave me a chance when few others would and I am eternally grateful for his support over the last half-decade. Five sentences of acknowledgment is nowhere near enough to express how thankful I am for the time and effort he spent ensuring my academic success.

Lastly, I want to thank my friends and family that have supported me through the grueling graduate school process. Thanks to my parents, Tony and Vickie, for your love and support. Thanks to my siblings, Allison and Taylor, for listening to me drone on and on about my work. Thanks to Bryant and Christine for always encouraging me to follow my dreams (and hopefully, one day, Artemis can read this and see that I actually do kind of know what I'm talking about). Thanks to Kiela for the daily encouragement and helping me realize that my work is often better than I think. Finally, thanks to my wonderful wife, Dana - I wouldn't be at this point if not for you.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES viii

LIST OF TABLES x

ABSTRACT xii

CHAPTER

I. Introduction 1

1.1 Overview 1
1.2 Summary of Contributions and Impact 4
1.2.1 Chapter 2 4
1.2.2 Chapter 3 7
1.2.3 Chapter 4 10
1.3 Extensions and Future Work 12
1.4 Bibliography 14

II. Disagreement and Dimensionality: A Varying Dimensions Approach to Roll Call Scaling in the U.S. Congress 18

2.1 Introduction 18
2.2 A Spatial Model of Roll Call Voting 21
2.3 A Roll Call Scaling Model with Varying Dimensions 25
2.3.1 A Spatial Model of Voting with Varying Dimensions 25
2.3.2 Estimating the Roll Call Scaling Model with Varying Dimensions 27
2.3.3 Estimating the Binary Matrix 28
2.3.4 A Beta Process IRT Model 30
2.4 Multidimensionality in the U.S. Congress Over Time 32

2.4.1 Interpretation of Ideal Points 37
2.5 U.S. Legislative Voting and Multidimensionality 42
2.5.1 A Theory of Pivotal Voters 45
2.5.2 A Theory on Party Control 50
2.6 Conclusion 55
2.7 Bibliography 58
A The BPIRT Model and Estimation 62
A.1 Beta Processes 62
A.2 BPIRT Full Model Specification 63
A.3 MCMC For BPIRT 64
A.4 Identification of Structural Parameters in BPIRT . . 67
A.5 Assessing Convergence for the BPIRT Algorithm . . 68
A.6 Methods for Achieving Faster Convergence 69
A.7 Simulation Exercises 70
B Multidimensionality in the U.S. Congress Over Time 74
B.1 IBP and Dimensions Observed in the U.S. House . . 75
B.2 More Summaries of Multidimensional Voting in the U.S. House 77
B.3 More Summaries for the 107th U.S. House 80
C U.S. Legislative Voting and Multidimensionality 82
C.1 A Theory of Pivotal Voters 82
C.2 A Theory on Party Control 84

III. A Bayesian Nonparametric Approach to Estimating Group Dynamics in Roll Call Scaling 89

3.1 Introduction 89
3.2 A Model for Roll Call Analysis 93
3.3 A Model for Roll Call Analysis with Group Influences 95
3.4 Clustered Beta Process IRT 100
3.4.1 Uncovering Unknown Latent Groups 102
3.4.2 Latent Groups within Ideal Point Estimation 104
3.4.3 Estimation of the C-BPIRT Model for Binary Outcomes 109
3.5 Application to Roll Call Scaling in the U.S. Congress 113
3.5.1 107th U.S. Congress 114
3.5.2 88th U.S. Congress 129
3.6 Conclusion 137
3.7 Bibliography 139
A Gibbs Sampling for C-BPIRT 143

IV. Interval Estimation on the Marginal Likelihood 147

4.1 Introduction 147
4.2 Computing the Marginal Likelihood 151

4.2.1 Candidate's Estimator 152
4.2.2 Naive Monte Carlo Estimator 153
4.2.3 Importance Sampling Estimators 154
4.2.4 Bridge Sampling Estimator 156
4.2.5 Laplace-Metropolis Estimator 158
4.2.6 Bounds on the Marginal Likelihood via Variational Approaches 159
4.3 Estimators, Bounds, and Divergence 161
4.4 An Interval Estimate for the Marginal Likelihood 166
4.4.1 Estimation and Performance of JSW Bounds 171
4.5 Refinements to the JSW Interval 175
4.5.1 Kappa Bounds on the Marginal Likelihood 177
4.5.2 Estimation of Kappa Bounds 180
4.5.3 Performance of Kappa Bounds 188
4.6 Application of Kappa Bounds to Ordered Factor Analysis Models191
4.6.1 Kappa Bounds Applied to Simulated Data 191
4.6.2 Kappa Bounds Applied to Racial Resentment in the U.S. 201
4.7 Conclusion 206
4.8 Bibliography 209
A Proof for Lemma 4.4.1 214
B Proof for Theorem 4.5.1 215
C Proof for Lemma 4.5.2 221

LIST OF FIGURES

Figure

2.1 Dimensionality and Corresponding 95% HPD Intervals Estimated by BPIRT for the 1st - 115th Sessions of the U.S. House and the U.S. Senate. 32

2.2 Proportion of Multidimensional Votes and Proportion of Variance Explained by the 1^{st} Dimension Estimated by BPIRT in the 1st - 115th Sessions of the U.S. House 34

2.3 Geometric Mean Probability of Correct Classification for Unidimensional and Multidimensional Votes in the 1st - 115th Sessions of the U.S. House 36

2.4 Correlation between the 1st Dimension of Ideal Points Estimated by BPIRT, the Ideal Points from WNOMINATE-1D, and the Proportion of Majority Party Voting 38

2.5 Ideal Points and Dimensions Estimated by BPIRT for the 107th Session of the U.S. House (2001 - 2003) 39

2.6 Votes and Cutlines for Department of Defense Authorization Act for Fiscal Year 2003 Vote in the 107th Session of the U.S. House (Roll Call No. 655) 41

2.7 Probability of Cloture Vote Switch by Ideal Point Quartile 47

2.8 Results of Final Passage Votes in the 83rd - 115th Sessions of the U.S. House 52

2.9 Probability of Party Rolls as a Function of Distance between the Party Median and the Floor Median 53

2.10 Relationship Between Number of Votes, Estimated Number of Dimensions, and Prior Number of Dimensions Implied by IBP for each Session of the U.S. House 75

2.11 Number of Unidimensional and Multidimensional Votes Analyzed in Each Session of the U.S. House 77

2.12 Aggregate GMP and Proportion Correctly Classified Votes for the 1st - 115th Sessions of the U.S. House 78

2.13 Highest PVE Dimensions and Number of Dimensions For Each Vote in the 107th U.S. House 81

3.1 Graphical Representation of Beta Process IRT Model 95

3.2 Graphical Representation of C-BPIRT 110
3.3 Ideal point estimates from NOMINATE and C-BPIRT for the 107th U.S. House. 116
3.4 Map of U.S. House districts and group labels for the 107^{th} Session of the U.S. House. 118
3.5 Comparison of first dimension ideal point estimates from NOMINATE, BPIRT, and C-BPIRT for the 107th U.S. House 120
3.6 Comparison of squared residuals for 239 one dimensional votes in the 107^{th} U.S. House. 122
3.7 Ideal point estimates and roll call votes associated with the Shays Amendment to the Bipartisan Campaign Reform Act presented during the 107^{th} session of the U.S. House. 125
3.8 Posterior predictive probability distribution that a new legislator casts a "Yea" vote to approve the Shays amendment to the Bipartisan Campaign Reform Act in the 107^{th} session of the U.S. House given only a C-BPIRT group label. 127
3.9 Ideal Points and Dimensions estimated by C-BPIRT for the 88^{th} session of the U.S. House. 131
3.10 Map of U.S. House districts for the 88^{th} Session of the U.S. House. . 132
3.11 Two Dimensional NOMINATE Ideal Points for the 88^{th} U.S. House. 133
4.1 Density plots of 500 simulated bridge sampling estimates compared to the true marginal likelihood. 164
4.2 Bridge sampling estimator and JSW intervals for 500 randomly generated linear regression data sets with varying numbers of covariates and amounts of posterior skew. 174
4.3 Shrinkage provided by Kappa Bounds compared to JSW Bounds. . 179
4.4 Bridge sampling estimator, JSW intervals, and Kappa intervals for 500 randomly generated linear regression data sets with varying numbers of covariates and amounts of posterior skew. 189
4.5 Comparison of marginal likelihood estimates for ordered factor analysis models. 198
4.6 Comparison of marginal likelihood estimates given by kappa bounds and bridge sampling for an ordered factor analysis model applied to questions related to racial resentment in the 1986 ANES. 205

LIST OF TABLES

Table

2.1 Logistic Regression Results for Krehbiel's Cloture Vote Switching Example . . . 48

2.2 Logistic Regression Results for Cox and McCubbins' Final Passage Vote Example . . . 57

2.3 Number of Dimensions Estimated From Simulated Data With Various Numbers of Voters, Votes, and Known Dimensionalities Using BPIRT . . . 71

2.4 Proportion of Elements in **R** Correctly Classified From Simulated Data With Various Numbers of Voters, Votes, and Known Dimensionalities Using BPIRT . . . 73

2.5 Correlation between Dimensions Estimated for the 107th U.S. House 80

2.6 Summary statistics for cloture vote switches in the 89^{th} - 115^{th} Sessions of the U.S. Senate . . . 87

2.7 Summary statistics for final passage votes in the 83^{rd} - 115^{th} Sessions of the U.S. House . . . 88

3.1 Votes on the Shays Amendment to the Bipartisan Campaign Reform Act presented during the 107^{th} session of the U.S. House . . . 124

3.2 Expected number of Yes and No votes on Shays amendment to the Bipartisan Campaign Reform Act in the 107^{th} session of the U.S. House given only a C-BPIRT group label. . . . 127

3.3 Proportion of Votes Correctly Classified and Geometric Mean Probability of Correct Classification for all roll call votes in the 107^{th} U.S. House. . . . 128

3.4 Probability of same vote cast on a roll call vote for each C-BPIRT group. . . . 135

4.1 Factor loading constraints matrix for each model compared on simulated data. . . . 196

4.2 Probability each ordered factor analysis model best fits the data. . . 198

4.3 Proportion of times over 500 data replicates that each ordered factor analysis model best fits the data. . . . 200

4.4 Loadings Matrix Constraint Specification for Models of Racial Resentment (RR), Individualism (I), and Traditional Racism (TR) Survey Batteries from 1986 ANES. 204

ABSTRACT

This dissertation comprises three essays on latent variable models and Bayesian statistical methods for the study of American legislative institutions and the more general problems of measurement and model comparison. In the first paper, I explore the dimensionality of latent variables in the context of roll call scaling. The dimensionality of ideal points is an aspect of roll call scaling which has received significant attention due to its impact on both substantive and spatial interpretations of estimates. I find that previous evidence for unidimensional ideal points is a product of the Scree procedure. I propose a new varying dimensions model of legislative voting and a corresponding Bayesian nonparametric estimation procedure (BPIRT) that allows for probabilistic inference on the number of dimensions. Using this approach, I show that there is strong evidence for multidimensional ideal points in the U.S. Congress and that using only a single dimension misses much of the disagreement that occurs within parties. I reexamine theories of U.S. legislative voting and find that empirical evidence for these models is conditional on unidimensionality.

In the second paper, I expand on the varying dimensions model of legislative voting and explore the role of group dependencies in legislative voting. Assumptions about independence of observations in the scaling model ignore the possibility that members of the voting body have shared incentives to vote as a group and lead to problems in estimating ideal points and corresponding latent dimensions. I propose a new ideal point model, clustered beta process IRT (C-BPIRT), that explicitly allows for group contributions in the underlying spatial model of voting. I derive a corresponding empirical model that uses flexible Bayesian nonparametric priors to estimate group effects in ideal points and the corresponding dimensionality of the ideal points. I apply this model to the 107th U.S. House (2001 - 2003) and the 88th U.S. House (1963 - 1965) and show how modeling group dynamics improves the estimation and interpretation of ideal points. Similarly, I show that existing methods of ideal point estimation produce results that are substantively misaligned with historical studies of the U.S. Congress.

In the third and final paper, I dive into the more general problem of Bayesian model comparison and marginal likelihood computation. Various methods of computing the marginal likelihood exist, such as importance sampling or variational methods, but they frequently provide inaccurate results. I demonstrate that point estimates for the marginal likelihood achieved using importance sampling are inaccurate in settings where the joint posterior is skewed. I propose a light extension to the variational method that treats the marginal likelihood as a random variable and create a set of intervals on the marginal likelihood which do not share the same inaccuracies. I show that these new intervals, called kappa bounds, provide a computationally efficient and accurate way to estimate the marginal likelihood under arbitrarily complex Bayesian model specifications. I show the superiority of kappa bounds estimates of the marginal likelihood through a series of simulated and real-world data examples, including comparing measurement models that estimate latent variables from ordered discrete survey data.

CHAPTER I

Introduction

1.1 Overview

Measurement models and latent variables models, more generally, are tools that are now common in political science research. While interest in measuring unobservable concepts and understanding relationships between these concepts and real-world outcomes has always existed in political science, recent methodological and computational advancements have led to new and exciting applications of these models to study social phenomena. These new tools provide researchers with a means of measuring difficult to observe concepts based on events, ratings, or other pieces of observable information that are assumed to be a result of the underlying unobservable latent trait (Fariss et al., 2020).[1]

Latent variable models are built on the idea that a set of observable outcomes are manifestations of the underlying latent trait - the latent variable model uses a set of probabilistic assumptions to model the unobservable latent trait as a function of the observed outcomes. These empirical models have been applied across a number of areas within political science: political ideology (Barbera, 2015; Bond and Messing,

[1] A few of the many new innovations in this rich area are presented in Imai et al. (2016); Jackman (2000, 2001); Martin and Quinn (2002); Carpenter et al. (2016)

2015; Martin and Quinn, 2002; Caughey and Warshaw, 2015; Konig et al., 2013; Pan and Xu., 2018; Treier and Jackman, 2008; Windett et al., 2015), political attitudes, knowledge, and preferences (Blaydes and Linzer, 2008; Jesse, 2017; Stegmueller, 2013), human rights abuses (Fariss, 2014; Schnakenberg and Fariss, 2014).

In the study of American political institutions, the most famous application of latent variable models is roll call scaling. Roll call scaling is an attempt to measure the ideology and preferences of members of a voting body via observable outcomes on votes (Poole and Rosenthal, 1984, 1987, 1997; Clinton et al., 2004). In the U.S. legislature, the matrix of binary ”Yea” and ”Nay” votes across a number of roll calls are scaled according to a latent variable model where different legislators are assumed to share a common latent variable that dictates their voting behavior. While there are many instances where researchers treat this uncovered latent variable as ideology, a more correct and nuanced interpretation of the estimated latent variable is as an *ideal point* which encompass a legislator's most preferred legislative outcome within the uncovered policy space. With light changes, this procedure can be used to estimate ideal points in any setting where a set of voters cast a number of distinct votes.

Ideal points estimated by the NOMINATE methodology (Poole and Rosenthal, 1987), the most common strategy for estimating ideal points, have been used extensively in the political science literature - a list attempting to cite every instance of usage would likely be as long as this entire dissertation. The prevalence of this methodology is both a testament to its ease of use and the relatively simple interpretation of the resulting ideal points: legislators with similar voting records have similar ideal points while dissimilar legislators have dissimilar ideal points. The NOMINATE methodology is also brilliant in its construction that links the empirical scaling model to an underlying formal model of utility maximizing voters. This clear formal theoretic underpinning has made NOMINATE a tool that has been useful for decades and has led to many of the important theories and findings related to the U.S. legislative body.

Ideal points estimated by NOMINATE paint a pessimistic picture of U.S. legislative voting: members vote along party lines on a single dimension. Under these scores, members of Congress are not complex voters that consider a variety of pressures when making decisions; legislators do not consider their "concentric circles of constituency" (Fenno, 1978); legislators do not have pet issues (Sulkin, 2005) - legislators are simply agenda-setting party loyalists that have few individual preferences and desires when taking office. When NOMINATE scores are subsequently used to capture legislative ideology in other models, empirical models inherently bake this simplistic view of legislative voting into the resulting findings.[2] NOMINATE scores have been used to show that Congress is completely party driven and has led to much of the polarization that is seen in U.S. politics today (McCarty et al., 2016).

Yet, scholars of the U.S. Congress continue to find that legislators are complex voters that are responding to many pressures when casting legislative votes (Roberts et al., 2016). This begs the question - why do in-depth studies of U.S. legislative voting find complexity when NOMINATE scores do not? In this dissertation, I contend that this theoretical mismatch is due to a number of simplifying assumptions and estimation problems in the NOMINATE procedure which distort inferences made from the roll call scaling model and induce the party-driven legislator findings; when roll call scaling models are allowed to discover complexity, empirical estimates consistently discover complexity. By correcting for the statistical and formal theoretic problems present in the NOMINATE methodology, ideal points can be estimated that match the complexity of the findings from decades of scholarship on the U.S. Congress. In turn, new scores can be used to provide a better understanding of the political structures that have led the U.S. political system to where it is today.

While the focus of much of this dissertation is on roll call scaling, the problems out-

[2] See Cox and McCubbins (2005) and Krehbiel (1992) for two of the most highly cited examples of this empirical work.

lined and proposed solutions are applicable to a wide class of latent variable models in political science and beyond. Dimensionality is a concern in many different areas of latent variable estimation (i.e. survey research, human rights measurement, and interest group behavior just to name a few) and unidimensional assumptions are prevalent even when there is little empirical evidence in support. Similarly, dependence across observations and group-based incentive structures are prevalent in many areas where latent variable models are used, such as educational testing assessment, online behavior and prediction, and modeling of genetic and genomic phenomena. Lastly, model comparison for latent variable models is a well-known problem that has generally led scholars to either use fit approximations or forgo the model comparison step, all together. This dissertation presents solutions for a number of these problems and provides simple empirical models that can be adapted to any situation where latent variable models are used.

1.2 Summary of Contributions and Impact

1.2.1 Chapter 2

In Chapter 2, I address the dimensionality of latent variables and the application of the unidimensional model to the roll call scaling problem. NOMINATE scores have been used to claim that roll call voting in the U.S. Congress is best described by a one dimensional policy space (Poole and Rosenthal, 1997; McCarty et al., 2016). The so-called "unidimensional conjecture" is a key result in the study of the U.S. Congress and provides support for a number of theories related to party power and issue voting (Aldrich, 1995; Rohde, 2010; Krehbiel, 1992; Cox and McCubbins, 2005; Sulkin, 2005). Scholars note, however, that when votes are examined by issue area rather than as an aggregated unit, there is evidence for different liberal-conservative orderings of legislators conditional on the type of policy being considered (Roberts et al., 2016).

While this point may seem purely methodological, the dimensionality of the policy space has significant substantive implications - a one dimensional policy space guarantees predictable outcomes near the median of the legislature (Krehbiel, 1992) while a multidimensional policy space grants no such guarantee (Schofield, 1978). A similar conclusion regards the role of party loyalty. Lee (2009) shows that the first dimension of NOMINATE ideal points corresponds more closely to the party loyalty of a legislator rather than any notion of ideology, so a unidimensional policy space implies that a legislator's allegiance to her party is the only thing that matters when making vote decisions; a multidimensional policy space implies a more complex calculation when making vote choices. Given the substantive implications of the dimensionality, it is important to rigorously nail down the structure of the latent space. However, current roll call scaling techniques are unequipped to rigorously address this concern.

I address this problem by proposing a formal model of legislative voting that allows voters to select from a bundle of dimensions on each vote. Where the standard spatial model that underlies NOMINATE requires that all voters use the same dimensions of the policy space on all votes, I allow voters to select a vote-specific set of dimensions from an overall set of dimensions that then influence the calculus of voting. As pointed out by Lee (2009), Aldrich et al. (2014), and Roberts et al. (2016), policy issues that create divisions within parties are frequent but fleeting; rarely do the same sets of divisions arise across all votes in a session of the U.S. Congress. I derive a new empirical model for roll call votes, beta process IRT (BPIRT), which is an empirical analogue to the varying-dimensions model of legislative voting. Through the usage of an Indian Buffet Process prior on the vote dimensions, I am able to both estimate the total number of dimensions in the policy space and identify which dimensions each vote requires to best model the observed roll call data. Simulations show that this procedure yields the correct answer when enough data is present and successfully defaults to the unidimensional model of voting when the algorithm is not provided with enough information, ruling out the possibility that BPIRT is simply discovering spurious dimensions.

Using the BPIRT model on the roll call voting records for all members of the U.S. Congress over the entire history of the U.S. legislature reveals a number of important findings that should alter how scholars use ideal point estimates in future work. First, I find no evidence that the unidimensional model of voting is an appropriate model for U.S. legislative voting behavior throughout U.S. history. Across sessions that have enough voters and votes, I find that there is strong evidence of multifaceted and complex voting patterns that transcend simple party allegiance. While a legislator's party is the biggest determinant of their vote on any single issue, BPIRT finds many votes, even in modern sessions of the U.S. Congress, that require dimensions beyond the first to explain voting behavior.

Building upon this result, I reexamine two theories of U.S. legislative voting behavior, the pivotal voter model (Krehbiel, 1992) and the party catel model (Cox and McCubbins, 2005), and explore their robustness to multidimensional ideal points using the scores estimated by BPIRT. For votes that were classified as unidimensional by BPIRT, both results hold and the original outcomes were replicated. However, for multidimensional votes, these models performed no better than a simple coin flip. This implies that nearly 70% of all votes by the U.S. Congress do not fit these highly-cited and generally revered theories of U.S. legislative voting behavior.

The implications of these findings are wide-reaching and generally damning for the unidimensional model of legislative voting. The results from BPIRT show that using unidimensional ideal point estimates will frequently yield incorrect inferences about the way that the U.S. Congress makes decisions. This same incorrectness applies to models that have made conclusions about other aspects of the U.S. political system using these results; conclusions about how parties work, how the legislative agenda is set, etc. must be readdressed to account for potential multidimensionality in legislative voting. While the simple unidimensional model of voting may lead to convenient results, this does not mean that they are correct. More generally, the BPIRT model demonstrates that current methods of determining dimensionality are often insufficient for truly uncovering the conplexity of a latent variable. Across disciplines, it is important to test the dimensionality of the latent space against other possible models.

1.2.2 Chapter 3

In Chapter 3, I address the problem of heterogeneous correlation among observations in latent variable models. The NOMINATE methodology (and latent variable modeling techniques, more broadly) assumes that a legislator only uses her ideal point and a bill specific weight to calculate the utility that she would get for casting a "Yea" or "Nay" vote. This particular assumption is puzzling since there is a large body of work that demonstrates that legislators make decisions in groups (Shepsle, 1978). In particular, parties are well known to distort the relationship between legislative preferences and vote outcomes (Aldrich, 1995; Rohde, 2010). Aldrich et al. (2014) show that scaling each party separately in modern sessions of the U.S. Congress reveals multidimensionality in legislative voting. This outcome is expected since party loyalty is the first dimension extracted from NOMINATE (Lee, 2009). If party preference is important, then it should be included in the underlying model and, if it is not, then it will lead to dependence among errors and, in turn, biased estimates of the ideal points. Thus, groups must be included in the underlying calculus of voting if there is any hope to accurately model legislative voting behavior. Again, current roll call scaling models are unequipped to handle correlation in errors due to group voting incentives.

To address this problem, I propose an extension of the varying dimensions model of legislative voting that allows for voters to make decisions in groups - a legislator's choice utility is a combination of both group preferences and individual preferences. This construction of utility is more in line with the work on party structures presented by Aldrich (1995) and Rohde (2010). Under this formal model of voting, I derive a clustered version of the beta process IRT model, C-BPIRT, that estimates ideal points, dimensions of legislative voting, the bundle of dimensions for each vote, and a group label for each legislator. Unlike hierarchical latent variable models, this model requires no *a priori* assumptions about each legislator's group (i.e. the model groups members by party if the data shows that there is correlation in errors among members of the same party). The model uses a novel Dirichlet process prior to assign members to groups where errors are independent conditional on group label. Much like the regression solution presented by Ferrari (2020), C-BPIRT uncovers meaningful groups under the latent variable framework.

I use the C-BPIRT model to explore ideal points in two sessions of the U.S. House: the 107^{th} U.S. House (2001 - 2003) and the 88^{th} session of the U.S. House (1963 - 1965). C-BPIRT ideal points demonstrate that there is a direct relationship between dimensions and groups in the roll call scaling model: dimensions uncovered by C-BPIRT tend to model group conflict. This fits with the findings of Lee (2009) that the first dimension of NOMINATE scores corresponds most closely to party loyalty; since interparty competition is the most prevalent division in the U.S. Congress, it makes sense that this would be the first conflict dimension extracted via roll call scaling. However, I show that the first dimension misses much of the interesting intraparty conflict. For example, in the 107^{th} U.S. House, I find that NOMINATE poorly explains votes on the McCain-Feingold Bipartisan Campign Finance Reform Act because it does not allow the New England Republicans to appear as their own distinct voting group. C-BPIRT uncovers this group and allows their ideal points to be closer to Democrats on issues of finance reform. C-BPIRT also estimates ideal points that are more historically consistent - in the 88^{th} U.S. House, the first dimension of NOMINATE says that New England Republicans and Southern Democrats are essentially the same on all but civil rights issues while C-BPIRT reflects the reality that they coalesced on very few votes and, in fact, could not be more different.

The implications of these findings are again wide-reaching. C-BPIRT ideal points improve upon the scores produced by NOMINATE and BPIRT by more accurately reflecting the calculus of voting in the U.S. Congress. The underlying formal model fits the many decades of theory about the role of parties in legislative voting and should be preferred to the simple individual utility maximization model underlying current roll call scaling procedures. C-BPIRT also improves on other approaches that attempt to address these correlations by either clustering based on party alone or simply forgoing the continuous scale latent variable by combining the continuous scale latent variable with a discrete scale latent variable that prevents projection of the group membership to its own dimension. Third, C-BPIRT provides a new tool for quantitative explorations of the development of the U.S. legislature and its group voting structure - the goal of the model is to estimate groups and their locations rather than individual ideal points which are later grouped according to researcher preference.

More generally, the C-BPIRT model is the first latent variable model that jointly models groups and individual latent variable locations without user assumptions about group structure. This model has applications across fields in any situation where correlation among observations may occur due to an omitted variable - educational research where test outcomes depend on state funding, survey research where responses are conditional on a hierarchical latent construct, etc. This broad and flexible model represents a state-of-the-art approach to estimating latent variables under heterogeneous dependence structures in data.

1.2.3 Chapter 4

In Chapter 4, I address the broader problem of latent variable model selection. Though the NOMINATE model is a statistical model, there is no clear way to compare different specifications of the roll call scaling model to determine which model best fits the data. While the general usage of the NOMINATE model sees researchers simply assume a one dimensional policy space in which all voters and votes occur, there are numerous constraints and choices that influence the outcome. More obvious in the Bayesian implementation of the NOMINATE procedure (Clinton et al., 2004), researchers can choose to allow different votes to correspond to different dimensions, allow different voters to use different dimensions, change the prior distribution over the latent variable, etc.

Yet, there exists no statistical approach to compare these choices. Researchers are left with two choices: post-hoc predictive fit metrics or cross-validation. Post-hoc predictive fit metrics like the proportion of votes correctly classified and the geometric mean of correct classification are commonly used to determine how well the model fits the data, but they incorrectly account for uncertainty and frequently lead to model selections that underestimate the complexity of voting. Computing the proportion of correctly classified votes using held out votes through cross-validation is another possible fitting strategy. This approach requires that votes are truly independent and identically distributed conditional on the ideal points, but there is significant evidence that they are not (Aldrich et al., 2014; Roberts et al., 2016). Without meaningful measurements of model evidence, there is no way for scholars to accurately compare models and test various theories of legislative voting using NOMINATE ideal points.

The more general problem of model selection with latent variable models is a well-known and long-studied problem. While software allows for users to compute various fit statistics and information criteria that go along with common factor analysis or

IRT models, these values are rarely reported. This is due to a number of reasons: 1) model fit metrics do not provide a consistent answer, 2) when the values do agree, they often point to the incorrect model, and 3) the fit statistics cannot distinguish between finely tuned differences in models like altered constraints. Bayesian methods of model selection, particularly marginal likelihood computation, are one area where scholars have found a consistent metric for comparing models (Lopes and West, 2004) - the marginal likelihood for each candidate model is computed and the models are either averaged or the model with the highest marginal likelihood is selected. While this approach shows good results in some cases, the marginal likelihood for latent variable models is difficult to compute and approximations to the marginal likelihood come with no measure of accuracy.

To address this problem, I propose a new interval estimator of the marginal likelihood. While various methods of computing the marginal likelihood exist, such as importance sampling or variational methods, they frequently provide inaccurate results. I demonstrate that point estimates for the marginal likelihood achieved using importance sampling are inaccurate in settings where the joint posterior is skewed, like many latent variable models. I propose a light extension to the variational method that treats the marginal likelihood as a random variable and create a set of intervals on the marginal likelihood which do not share the same inaccuracies. I show that these new intervals, called kappa bounds, provide a computationally efficient and accurate way to estimate the marginal likelihood under arbitrarily complex Bayesian model specifications. I show the superiority of kappa bounds estimates of the marginal likelihood through a series of simulated and real-world data examples, including comparing measurement models that estimate latent variables from ordered discrete survey data.

The method of model selection presented in this section provides a meaningful way for scholars to compare different latent variable model specifications when assessing multiple theories. Too often in the political science, scholars present results from a factor analysis or IRT model that aligns with their theories without providing any comparison to competing theories. This is, in part, due to a lack of available and easily implementable model comparison metrics for these models. The methods outlined in this chapter have applications across disciplines and can be used in many settings where other model comparison metrics are not considered feasible due to its relatively low computational cost. This chapter should serve as a starting point for discussions in the social sciences about model comparison and applying the same level of statistical rigor that is desired of regression models and other standard applied statistical models to latent variable models.

1.3 Extensions and Future Work

This dissertation presents a number of improvements to methods of roll call scaling and latent variable estimation. This work serves as the starting point for a larger exploration of how latent variable models can be used more effectively for modeling social phenomena. In particular, I seek to improve these methods to better model and understand the structure of U.S. legislative voting in the past, present, and future.

The BPIRT model uncovers the dimensions that dictate U.S. Congressional roll call voting. Currently, these dimensions are given names by looking at the content of the bills that are associated with each vote. Though the relationship between bill topic and dimension is not perfect, bills that consider similar topics tend to require similar dimensions in the BPIRT policy space. This implies that joint modeling of bill topics and legislative vote outcomes will provide more information about the meaning of dimensions and provide more information for the scaling procedure to estimate legislators' ideal points. Beyond bill topics, there are other pieces of "metadata" that can give more information about the vote and, in turn, improve estimation of the ideal points. These pieces include committee path through the legislature, sponsorship information, and legislative bill episodes that contain many subsequent votes on the same bill. This information can be introduced into the roll call scaling model through a hierarchical version of the Indian Buffet Process to include information about which bills should cluster; just as errors across legislators are correlated due to party, errors across votes should also be correlated due to common vote topics. Future work will see this relationship included in the model of roll call voting.

A further extension in this vein involves more general estimation of multiple related latent spaces. Like ideal points, bill topics are a latent variable. The ideal point space and the bill topic space contain shared information - knowing which voters were close to the indecision point on a specific vote provides information about which topics the bill contained and where on each issue dimension each bill existed. Just as simultaneous regression models can be run to model shared information about each outcome, simultaneous latent variable models can model the joint information between latent spaces. This model, which is based on joining BPIRT to a text-based topic model, would jointly scale these two pieces of information in hopes of creating a link that would allow translation from one space to another. This model would then be capable of credibly estimating status quo and alternative positions. The nuts and bolts of this model are similar to those presented in this dissertation, so the methods used to address questions of dimensionality and correlation would directly apply to this future model and should serve as a starting point for the work on simultaneous latent variable models.

A final extension to these models involves dynamic modeling of ideal points, policy dimensions, and legislative voting groups. C-BPIRT currently operates under the assumption that each session of the U.S. Congress is a distinct entity; new information about a legislator's preferences starts at the beginning of each sessions and ends at the session's conclusion. Just like DW-NOMINATE, a dynamic variant of the C-BPIRT model would see estimation of groups and dimensions over time. Modeling dimensions over time can be achieved by finding votes that are similar across sessions of the U.S. Congress and inducing a common dimension using the Indian Buffet Process prior. With greater computational power, all sessions of the U.S. Congress can be jointly scaled to see which issue dimensions exist in a given time period and how those dimensions have changed over time. On the other end, allowing groups to dynamically "walk" through the latent space over time would provide information about different legislative coalitions have evolved. Beyond an examination of how the groups have changed, linkages between groups can be established to show how cleavages form and disappear, over time. This model would have significant implications for the study of American political development and assist in understanding how legislative coalitions operate in the complex and ever-evolving U.S. legislature. This analysis would also provide valuable new information on how legislative compromise occurs. All in all, the C-BPIRT model presents a powerful new tool for understanding how U.S. political institutions work and evolve.

Bibliography

Aldrich, John H (1995). *Why parties?: The origin and transformation of political parties in America*. University of Chicago Press.

Aldrich, John H , Jacob M Montgomery, and David B Sparks (2014). Polarization and ideology: Partisan sources of low dimensionality in scaled roll call analyses. *Political Analysis 22*(4), 435–456.

Barbera, Pablo (2015). Birds of the same feather tweet together. Bayesian ideal point estimation using twitter data. *Political Analysis 23*(1), 76–91.

Blaydes, Lisa and Drew A. Linzer (2008, July). The political economy of women's support for fundamentalist islam. *World Politics 60*, 576–609.

Bond, Robert M. and Solomon Messing (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on facebook. *American Political Science Review 109*(1), 62–78.

Carpenter, Bob , Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2016). Stan: A probabilistic programming language. *Journal of Statistical Software 20*.

Caughey, Devin and Christopher Warshaw (2015). Dynamic estimation of latent opinion using a hierarchical group-level irt model. *Political Analysis 23*, 197–211.

Clinton, Joshua , Simon Jackman, and Douglas Rivers (2004). The statistical analysis of roll call data. *American Political Science Review 98*(2), 355–370.

Cox, Gary W and Mathew D McCubbins (2005). *Setting the agenda: Responsible party government in the US House of Representatives*. Cambridge University Press.

Fariss, Christopher J. (2014). Respect for human rights has improved over time: Modeling the changing standard of accountability. *American Political Science Review 108*(2), 297–318.

Fariss, Christopher J. , Michael R. Kenwick, and Kevin Reuning (2020). Measurement models. In L. Curini and R. J. Franzese Jr. (Eds.), *SAGE Handbook of Research Methods is Political Science and International Relations*. SAGE Press.

Fenno, Richard F (1978). *Home style: House members in their districts.* Pearson College Division.

Ferrari, Diogo (2020). Modeling context-dependent latent effect heterogeneity. *Political Analysis 28*(1), 20–46.

Imai, Kouske , James Lo, and Jonathan Olmsted (2016). Fast estimation of ideal points with massive data. *American Political Science Review 110*(4), 631–656.

Jackman, Simon (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain monte carlo. *American Journal of Political Science 44*(2), 375–404.

Jackman, Simon (2001). Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis 9*(3), 227–241.

Jesse, Stephen A. (2017). Don't know responses, personality and the measurement of political knowledge. *Political Science Research and Methods 5*(4), 711–731.

Konig, Thomas , Mortiz Marbach, and Mortiz Osnabrugge (2013). Estimating party positions across countries and time - a dynamic latent variable model for manifestos data. *Political Analysis 21*(4), 468–491.

Krehbiel, Keith (1992). *Information and legislative organization.* University of Michigan Press.

Lee, Frances E (2009). *Beyond ideology: Politics, principles, and partisanship in the US Senate.* University of Chicago Press.

Lopes, Hedibert Freitas and Mike West (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 41–67.

Martin, Andrew D. and Kevin M. Quinn (2002). Dynamic ideal point estimation via Markov chain monte carlo for the U.s. supreme court, 1953–1999. *Political Analysis 10*(2), 134–53.

McCarty, Nolan , Keith T Poole, and Howard Rosenthal (2016). *Polarized America: The dance of ideology and unequal riches.* mit Press.

Pan, Jennifer and Yiqing Xu. (2018). China's ideological spectrum. *Journal of Politics 80*(1), 254–273.

Poole, Keith T and Howard Rosenthal (1984). The polarization of American politics. *The Journal of Politics 46*(4), 1061–1079.

Poole, Keith T and Howard Rosenthal (1987). Analysis of congressional coalition patterns: A unidimensional spatial model. *Legislative Studies Quarterly*, 55–75.

Poole, Keith T and Howard Rosenthal (1997). *Congress: A political-economic history of roll call voting.* Oxford University Press on Demand.

Roberts, Jason M , Steven S Smith, and Stephen R Haptonstahl (2016). The dimensionality of congressional voting reconsidered. *American Politics Research 44*(5), 794–815.

Rohde, David W (2010). *Parties and leaders in the postreform House.* University of Chicago Press.

Schnakenberg, Keith E. and Christopher J. Fariss (2014). Dynamic patterns of human rights practices. *Political Science Research and Methods 2*(1), 1–31.

Schofield, Norman (1978). Instability of simple dynamic games. *The Review of Economic Studies 45*(3), 575–594.

Shepsle, Kenneth A (1978). *The giant jigsaw puzzle: Democratic committee assignments in the modern House.* University of Chicago Press Chicago.

Stegmueller, Daniel (2013). Modeling dynamic preferences: A Bayesian robust dynamic latent ordered probit model. *Political Analysis 21*, 314–333.

Sulkin, Tracy (2005). *Issue politics in Congress.* Cambridge University Press.

Treier, Shawn and Simon Jackman (2008). Democracy as a latent variable. *American Journal of Political Science 52*(1), 201–217.

Windett, Jason H. , Jeffrey R. Harden, and Matthew E. K. Hall (2015). Estimating dynamic ideal points for state supreme courts. *Political Analysis 23*(3), 461–469.

CHAPTER II

Disagreement and Dimensionality: A Varying Dimensions Approach to Roll Call Scaling in the U.S. Congress

2.1 Introduction

Studies of legislative behavior focus on the relationships between legislative preferences, institutional structure, and legislative outcomes. A common method used to understand these relationships uses scaling models that uncover the ideal points of legislators. While there are many approaches to uncovering the ideal points of legislators, by far the most common approach uses the outcomes from the various roll call votes that are cast by members of Congress. Roll call scaling techniques such as NOMINATE (Poole and Rosenthal, 1997) and its Bayesian analogue (Clinton et al., 2004) seek to project roll call data into a low-dimensional policy space that captures the complexities of how members of Congress make vote decisions. The ideal points can then be used to make inferences about the behavior of legislative actors, such as the role of parties (Aldrich and Rohde, 2000; Cox and Poole, 2002; Cox and McCubbins, 2005), influences within and between branches of the U.S. government (Binder, 1999; Krehbiel, 1998), and other features of the legislative institution.[1]

Ideal point models require assumptions that have implications for the interpretation of the estimated quantities. One assumption is the dimensionality of the latent space. Assuming a unidimensional ideal point, legislators behave predictably and rational choice models can provide simple explanations of how legislators make policy proposals and vote choices under the rules of the institution (Krehbiel, 1998; Cox and McCubbins, 2005). On the other hand, multidimensional ideal points create an environment where legislators behave in a more nuanced manner - legislative behavior is conditional on the context of the vote and there are few guarantees of predictable outcomes (McKelvey, 1976; Schofield, 1978; Shepsle, 1978). The choice of dimensionality also has strong substantive implications (Shepsle, 1978; Shepsle and Weingast, 1981; Harbridge, 2015; Lee, 2009).

In most recent studies that leverage unidimensional estimates of ideal points, empirical justification for this assumption is given by referencing the "Unidimensional Congress" arguments of Poole and Rosenthal (2011) - across issue areas within an analyzed period of the U.S. Congress, little improvement to the overall fit of the roll call scaling model can be made by including more than one dimension. However, many works show evidence for multidimensionality in U.S. Congressional roll call voting; there is strong evidence that certain bundles of votes map to different dimensions and less aggregated analysis of roll call votes reveals this heterogeneity (Heckman and Snyder Jr, 1996; Roberts et al., 2016; Smith, 2007; Hurwitz, 2001; Crespin and Rohde, 2010; Norton, 1999; Bateman et al., 2017).

I argue that while some roll call votes in the U.S. Congress are unidimensional, many more are not. I introduce technology that not only demonstrates this but allows

[1] There has been significant work in the area of roll call scaling and ideal point estimation for the U.S. Congress beyond these two models. Lauderdale and Herzog (2016), Tausanovitch and Warshaw (2017), Bonica (2014), Ramey (2016), Jessee and Malhotra (2010), and Tahk (2018) are just a few of the models proposed in recent literature.

the variability in dimensions to be studied. I present a method of roll call scaling that allows for aggregate-level summaries of legislative decision making while also allowing for examinations of multidimensionality at the bill-episode level. Leveraging work from Aldrich et al. (2014) and Roberts et al. (2016), I contend that evidence for the low-dimensional conjecture is due to the statistical tests used to assess inclusion of new dimensions. To address this problem, I present a new spatial model in which a voter makes vote decisions using both the positions of alternatives within the policy space and a vote-specific bundle of dimensions in which those policy positions exist. The corresponding empirical model allows for rigorous statistical inference related to the overall dimensionality of the ideal points and identification of the dimensions of the policy space associated with each vote. Unlike previous approaches, this method accurately estimates the dimensionality of the ideal point space even under high levels of party bloc voting. Similarly, the ideal points and vote-level estimates of dimensionality allow for new tests related to theories of legislative behavior that properly take vote-level dimensionality into account. Given that the vast majority of empirical tests related to legislative decision making use unidimensional NOMINATE scores, the estimates achieved from this new model allow for a finer examination of the role of dimensionality in many important theories of U.S. legislative behavior.

Overall, I make several important contributions to the literature in this article. Methodologically, I present a new spatial model of voting that has an explicit empirical analogue under assumptions about utility structures. This model uses novel advancements in the field of Bayesian nonparametrics related to estimating the infinite latent feature model (Paisley and Carin, 2009; Knowles and Ghahramani, 2011) to address the question of dimensionality in aggregate sets of roll call votes by simultaneously estimating ideal points and dimensionality. Substantively, I analyze the entire history of the U.S. House and U.S. Senate ($1^{st} - 115^{th}$ sessions) and show that there is strong evidence of multidimensional voting through the history of the U.S.

Congress. In line with many of the conclusions by Heckman and Snyder Jr (1996), I find that votes tend to bundle based on topic and these votes share similar multidimensional vote patterns. In turn, this allows for identification of key issues that split members of Congress, both within and across parties. This work produces a new set of ideal points across U.S. Congressional history that should provide a starting point for further work related to topic-level voting in the U.S. legislative body. Finally, I apply the estimates from the new roll call scaling method to two specific theories related to U.S. Congressional voting: the pivotal voter model (Krehbiel, 1998) and the party cartel model (Cox and McCubbins, 2005). I show that much of the empirical evidence that exists for these models changes when dimensionality is properly accounted for in empirical tests of these theories. This analysis is just the starting point for potentially reassessing many other predictions made by models of U.S. legislative voting under conditions of multidimensionality.

2.2 A Spatial Model of Roll Call Voting

For a legislature, assume there are N voting members that cast P votes over the course of time analyzed. For any given vote $j \in (1, P)$, legislator $i \in (1, N)$ must choose between two alternatives: to cast a "Yea" vote for the proposed alternative ($\boldsymbol{\vartheta}_j \in \mathbb{R}^K$) or to cast a "Nay" vote for the proposed alternative ($\boldsymbol{\varphi}_j \in \mathbb{R}^K$). Behavior in this legislature is assumed to be describable in a K-dimensional policy space - all votes that are made by legislator i can be described by the K-dimensional point locations of $\boldsymbol{\vartheta}_j$ and $\boldsymbol{\varphi}_j$ within the space and a K-dimensional ideal point, $\boldsymbol{\omega}_i$, which encapsulates the policy preferences of legislator i.

A legislator must choose whether to vote for $\boldsymbol{\vartheta}_j$ or $\boldsymbol{\varphi}_j$. Using a utility maximization model that assumes quadratic loss in distance from her ideal point, assume that she chooses the alternative which grants the highest utility:

$$\begin{aligned} U_i(\boldsymbol{\vartheta}_j) &= -\|\boldsymbol{\omega}_i - \boldsymbol{\vartheta}_j\|^2 + \eta_{ij} \\ U_i(\boldsymbol{\varphi}_j) &= -\|\boldsymbol{\omega}_i - \boldsymbol{\varphi}_j\|^2 + \nu_{ij} \end{aligned} \tag{2.1}$$

where η_{ij} and ν_{ij} are stochastic elements of the utility functions. This model is completely specified if a known structure is placed on η_{ij} and $\nu_{i,j}$ (Heckman and Snyder Jr, 1996; Poole and Rosenthal, 1997; Clinton et al., 2004).

Let $\boldsymbol{Y}$ be a matrix of roll call votes and $y_{i,j}$ be the vote choice that legislator i makes on proposal j : $y_{ij} = 1$ if legislator i votes "Yea" on vote j and $y_{ij} = 0$ if she casts a "Nay" vote. Given the model construction, the probability that legislator i votes for $\boldsymbol{\vartheta}_j$ can represented as:

$$P(y_{ij} = 1) = F(\boldsymbol{\lambda}_j' \boldsymbol{\omega}_i - \alpha_j) \tag{2.2}$$

where $F(\cdot)$ is the CDF associated with the chosen error structure, $\alpha_j = \frac{\boldsymbol{\vartheta}_j'\boldsymbol{\vartheta}_j - \boldsymbol{\varphi}_j'\boldsymbol{\varphi}_j}{\sigma_j^2}$, and $\boldsymbol{\lambda}_j = \frac{2(\boldsymbol{\vartheta}_j - \boldsymbol{\varphi}_j)}{\sigma_j^2}$.

This construction admits a corresponding statistical model that allows for estimation of the *structural parameters* $\boldsymbol{\alpha}$ and $\boldsymbol{\Lambda}$ and the latent variables, $\boldsymbol{\Omega}$. Assuming the errors are independent and identically distributed, a likelihood function can be derived:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Omega} | \boldsymbol{Y}) = \prod_{i=1}^{N} \prod_{j=1}^{P} F(\boldsymbol{\lambda}_j' \boldsymbol{\omega}_i - \alpha_j)^{y_{ij}} \times \left(1 - F(\boldsymbol{\lambda}_j' \boldsymbol{\omega}_i - \alpha_j)\right)^{1 - y_{ij}} \tag{2.3}$$

Bayesian implementations of this model place priors on all of the structural parameters and estimation proceeds using Markov Chain Monte Carlo methods (Clinton et al., 2004). With minor changes, this model is equivalent to the NOMINATE procedure (Poole and Rosenthal, 1997).

Under this specification, a choice of the number of dimensions, K, leads to fully tractable model that can be estimated. Currently, this choice is made by examining a Scree plot (Cattell, 1966). The spirit of the Scree procedure revolves around estimating the latent variable model under a number of different assumptions for the number of dimensions and plotting the fit metric to find the “elbow” in the plot. Once adding a new dimension no longer adds “enough” value to the fit metric, then no new dimensions are added. Typically, the choice of fit metric revolves around the proportion of votes correctly classified under a model: aggregate proportion reduction in error (Poole and Rosenthal, 2011) or marginal proportion reduction in error (Roberts et al., 2016).

Regardless of the choice of metric, I contend that the Scree test presents many problems for inference. By its very nature, the Scree test is inherently subjective; the choice of how much reduction in error is enough to include a new dimension is subjective and can lead to biases in the number of dimensions chosen. For example, the Scree test cannot detect small improvements in model fit that are due to adding dimensions that only contribute to a few votes. Rather, the Scree test says that the model improvement over the aggregated set of roll call votes is small and the dimension should not be included. This feature of the Scree test is not ideal as there is an entire body of the literature which shows that dimensions in roll call voting appear at the vote-topic level and important dimensions can appear infrequently (Roberts et al., 2016).

Aldrich et al. (2014) and Roberts et al. (2016) point to the high frequency of votes that occur along party lines as a problem for current dimensionality testing procedures. When many votes are explained by party lines, the perceived influence of party can be much higher than what is actually present within the data. Given that there is often correlation between vote choices in specific policy domains and party membership, party bloc voting can appear to account for all of the explainable variation within roll call voting data sets when many dimensions are truly influencing decisions.

This result is corroborated by the finding that scaling within parties reveals many dimensions even when using Scree tests as the decision making criterion (Aldrich et al., 2014). Similarly, multidimensionality is highly apparent when dimensionality is tested within and across topically similar bill-episodes (Roberts et al., 2016). Aldrich et al. (2014) points to the inclusion of zeroes in the matrix of discrimination parameters as a method for properly modeling the covariance between dimensions and, in turn, accurately estimating the dimensionality of the ideal points.

These findings point to a couple of features that a roll call scaling method that accurately uncovers dimensionality should have:

1. Dimensionality should be tested under distributional assumptions. In turn, probabilistic tests of whether or not a dimension provides a non-zero improvement to the model under an assumption about what constitutes random noise can be performed. This ensures the inclusion of dimensions which have a strong influence on a small set of votes.

2. Dimensionality should be tested at the vote level. Each vote should be allowed to draw on a different set of dimensions, if necessary. The aggregated set of vote-level dimensionalities then dictates the dimensionality of the ideal point. However, each set of vote-level dimensions should be subject to overfitting penalties in order to estimate substantively useful parameters that account for both vote-level and aggregate behavior of legislators over the course of time analyzed. This leads to a proper encoding of conditional independence between dimensions at the vote-level and, in turn, an accurate count of the number of dimensions that make up the ideal point space.

A roll call scaling method that meets these conditions should provide an accurate representation of the dimensionality of the data while also reducing the dimensionality of the data to something useful for further examination of aggregate legislative voting behavior.

2.3 A Roll Call Scaling Model with Varying Dimensions

2.3.1 A Spatial Model of Voting with Varying Dimensions

To address the above conditions for accurately estimating dimensionality, I propose a new roll call scaling model with varying dimensions. As before, legislator i must choose to vote for $\boldsymbol{\vartheta}_j$ or $\boldsymbol{\varphi}_j$. She votes for the alternative that maximizes her utility under a quadratic loss function such that:

$$\begin{aligned} U_i(\boldsymbol{\vartheta}_j) &= -\|\boldsymbol{r}_j(\boldsymbol{\omega}_i - \boldsymbol{\vartheta}_j)\|^2 + \eta_{ij} \\ U_i(\boldsymbol{\varphi}_j) &= -\|\boldsymbol{r}_j(\boldsymbol{\omega}_i - \boldsymbol{\varphi}_j)\|^2 + \nu_{ij} \end{aligned} \tag{2.4}$$

Under this specification, the new addition is the binary vector $\boldsymbol{r}_j$. $\boldsymbol{r}_j$ is a vector of length K where $r_{jk} = 1$ if she considers dimension $k \in (1, ..., K)$ in vote j. On the other hand, $r_{jk} = 0$ if she does not use her ideal point on dimension k when making a decision for vote j. Note that $\boldsymbol{r}_j$ is assumed to be globally known to all legislators.

The length of $\boldsymbol{r}_j$ does not need to be set before specifying the model. For example, if only dimensions one and three are needed to dictate the utility function associated with a vote (i.e. $r_{j1} = 1$, $r_{j2} = 0$, $r_{j3} = 1$), then this vector is equivalent to one where $r_{j4} = 0$, $r_{j5} = 0$, and all subsequent elements of the vector are set to zero. Thus, the vector of length three and the corresponding vector of infinite length are equivalent. This characteristic of the binary vector is key to addressing the shortcomings of the standard roll call scaling model.

Placing all of the vote level binary vectors, $\boldsymbol{r}_j$, into a matrix with the number of rows equal to the number of votes and the number of columns equal to the number of possible dimensions creates a binary matrix, $\boldsymbol{R}$, that dictates the mapping of individual votes to ideal point dimensions. $\boldsymbol{R}$ captures the dimensionality of the underlying ideal point space across all votes analyzed. Recall that each $\boldsymbol{r}_j$ can be of infinite size, but only the elements equal to one matter for the underlying utility model. Thus, the dimensionality of the overall space can be modeled as the number of columns in $\boldsymbol{R}$ which have at least one non-zero element. Similarly, the structure of $\boldsymbol{r}_j$ is allowed to vary across votes - each vote can call on a different set of dimensions to construct the parameters of the assumed utility calculations that lead to vote decisions.

As with the standard model, the construction admits a corresponding statistical model. With similar rearrangement, a likelihood function is determined:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Omega}, \boldsymbol{R}|\boldsymbol{Y}) = \prod_{i=1}^{N}\prod_{j=1}^{P} F\left((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'\boldsymbol{\omega}_i) - \alpha_j\right)^{y_{ij}} \times \left(1 - F\left((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'\boldsymbol{\omega}_i - \alpha_j\right)\right)^{1-y_{ij}} \tag{2.5}$$

where $\odot$ is the Hadamard product of two vectors.[2] As with the likelihood function for the standard roll call scaling model, the likelihood is comprised of the structural parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}$ and the ideal points, $\boldsymbol{\Omega}$. $\boldsymbol{R}$ is assumed to modify the loadings, $\boldsymbol{\Lambda}$. Like the standard model, the parameters of the estimated statistical model explicitly link back to the formal theoretic foundation where $\alpha_j = \frac{\boldsymbol{r}_j(\boldsymbol{\vartheta}_j'\boldsymbol{\vartheta}_j - \boldsymbol{\varphi}_j'\boldsymbol{\varphi}_j)}{\sigma_j^2}$ and $\boldsymbol{\lambda}_j = \frac{2\boldsymbol{r}_j(\boldsymbol{\vartheta}_j - \boldsymbol{\varphi}_j)}{\sigma_j^2}$.[3]

The varying dimensions model of voting explicitly adds the two conditions listed previously. First, $\boldsymbol{R}$ constitutes a new quantity that can be estimated. With distributional assumptions, $\boldsymbol{R}$ can be estimated with the other structural parameters of

[2] The standard roll call scaling models is a special case of the varying dimensions version - setting all elements of $\boldsymbol{R}$ equal to one and fixing K to a known finite value replicates the model of Clinton et al. (2004).

[3] Like the Bayesian IRT approach of Clinton et al. (2004) and the NOMINATE model, this model assumes that all voters vote sincerely based on their underlying ideal point. Using a model that ties abstention to strategic behavior, this assumption could be changed.

the ideal point model to determine the number of dimensions needed to effectively model the ideal point space. If this choice penalizes against adding many dimensions, $\boldsymbol{R}$ can dictate a sparse set of dimensions that directly model the aggregate set of roll call votes. Second, each binary vector, $\boldsymbol{r}_j$, contains a mapping of each vote to some subset of the ideal point space. This allows each vote to be modeled in a potentially different set of dimensions. Again, under proper distributional assumptions about $\boldsymbol{R}$, this allows each vote to be modeled as a subset of aggregate dimensions. However, if only one dimension is needed for the collection of votes, $\boldsymbol{R}$ can be reduced to estimate only one dimension. All in all, this model of voting allows for the dimensionality of the latent space to be estimated simultaneous to the structural parameters of the ideal point model.

2.3.2 Estimating the Roll Call Scaling Model with Varying Dimensions

Using equation (2.5) as a starting point, a Bayesian scaling procedure for binary roll call votes with varying dimensions can be defined.[4] Let $\boldsymbol{X}$ be a continuous latent mapping of the observed roll call votes $\boldsymbol{Y}$ such that:

$$x_{ij} \sim \begin{cases} \mathcal{TN}_{-\infty,0}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega_i} - \alpha_j, 1) \text{ if } y_{ij} = 0 \\ \mathcal{TN}_{0,\infty}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{ij} = 1 \\ \mathcal{N}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{ij} \text{ is missing} \end{cases} \tag{2.6}$$

where $\mathcal{TN}_{l,u}(\mu, \sigma^2)$ is the truncated normal distribution truncated between l and u. Assuming a probit structure on the errors and, without loss, an infinite dimension ideal point space, equation (2.6) defines the varying dimension roll call scaling model.

This voting model shares many parameters with the two-parameter item response

[4] The benefits of approaching the roll call scaling problem under the Bayesian paradigm are well documented. A thorough discussion of these benefits are presented by Clinton et al. (2004).

model used in educational testing where $\boldsymbol{\lambda}_j$ is a vector of item discrimination parameters, α_j is the item difficulty parameter, and $\boldsymbol{\omega}_i$ is a vector of ideal points associated with the vote decisions made by a legislator (Poole and Rosenthal, 1997; Londregan, 1999).

2.3.3 Estimating the Binary Matrix

A number of approaches for selecting the appropriate number of dimensions in latent variable models have appeared in the statistics and social sciences literatures (Kim et al., 2018). However, these approaches often are developed for the purpose of finding a reduced dimensionality representation of the covariance matrix and estimation of the structural parameters and item-level differences in dimensionality are not addressed. I use nonparametric priors on the number of dimensions and allow the dimensions to vary stochastically. These priors have strict probabilistic properties that make identification and estimation of structural parameters plausible (Bhattacharya and Dunson, 2011).

One Bayesian nonparametric approach which allows probabilistic modeling of both the overall and vote-level dimensionalities uses the beta process (Paisley and Carin, 2009). For $r_{jk} \in \boldsymbol{R}$, let the prior be:

$$\begin{aligned} P(r_{jk}|\pi_{jk}) &\sim \text{Bern}(r_{jk}; \pi_{jk}) \\ P(\pi_{jk}|a_{jk}, b_{jk}) &\sim \text{Beta}(\pi_{jk}; a_{jk}, b_{jk}) \end{aligned} \tag{2.7}$$

Letting $K \to \infty$ allows all possible dimensions to be potentially present in $\boldsymbol{R}$ and constitutes a beta process.

Without further constraint, the model will always find that the optimal number of features is equal to the number of items - each item is modeled by its own dimension. This outcome is akin to over-fitting in regression and provides a solution that is not useful for summarizing high-dimensional data; the roll call scaling model needs a *sparse* estimate of $\boldsymbol{R}$. Along with other challenges related to fitting an infinite set of dimensions to a finite set of data, estimation under the beta process prior has proven a difficult task in the statistics literature.

A marginalized approach exists that allows for a relatively simple sampling scheme in the infinite limit that prevents against over-fitting by allowing the number of effective dimensions to scale with the number of observations and items (Paisley and Carin, 2009). Here, the beta-Bernoulli process dictating the values of the infinite matrix, $\boldsymbol{R}$, has a prior such that:

$$\begin{gathered} P(\pi_k) \sim \text{Beta}\left(\pi_k; \frac{a}{K}, \frac{b(K-1)}{K}\right) \\ P(r_{jk}|\pi_k) \sim \text{Bern}(r_{jk}; \pi_k) \end{gathered} \tag{2.8}$$

where a and b are hyperparameters and K is arbitrarily large such that:

$$E[\pi_k|K] \approx 0 \tag{2.9}$$

which induces a sparse estimate of the binary matrix.

The beta process prior of this form constitutes a simple stochastic process. Marginalizing over π, this process is the *two parameter Indian Buffet Process* (IBP) (Ghahramani and Griffiths, 2006). IBP has two notable properties for modeling sparse matrices. First, over the entire set of P votes, the number of dimensions sampled follows:

$$\begin{gathered} P(K^+ = h) = \text{Pois}\left(h; \sum_{j=1}^{P} \frac{ab}{b+j-1}\right) \\ E[K^+] \approx \mathcal{O}(\ln(P)) \end{gathered} \tag{2.10}$$

where K^+ is the number of columns of $\boldsymbol{R}$ with at least one element equal to one. This shows that as P increases, the number of dimensions that can potentially appear in the latent space increases. However, the expected number of dimensions is small relative to P and a sparse solution is ensured. On the other hand, if P is small, then the number of dimensions that can potentially appear in the latent space is also small. This property ensures that the number of dimensions estimated is supported by the amount of data present during estimation.

Second, IBP exhibits a "rich get richer" property:

$$P(r_{jk} = 1) \propto \frac{-r_{jk} + \sum\limits_{h=1}^{P} r_{hk}}{P + 1 + r_{jk} - \sum\limits_{h=1}^{P} r_{hk}} \tag{2.11}$$

As a dimension becomes more popular, the probability that it is sampled in other votes increases. In turn, features that are not popular are rarely sampled and have a small chance of being included in the final model specification. Thus, IBP explores the feature space in accordance with P, but still promotes sparsity by allowing popular dimensions to dominate the feature space. This allows the IBP prior to prevent against overfitting. However, in the face of strong statistical evidence, IBP still allows an unpopular feature to emerge.

2.3.4 A Beta Process IRT Model

Under the specification in equation (2.6), $\boldsymbol{r}_j$ induces a *spike and slab* prior on the vector of discrimination parameters, $\boldsymbol{\lambda_j}$. Placing an Indian Buffet Process prior on $\boldsymbol{R}$, the induced prior on $\lambda_{j,k}$ is:

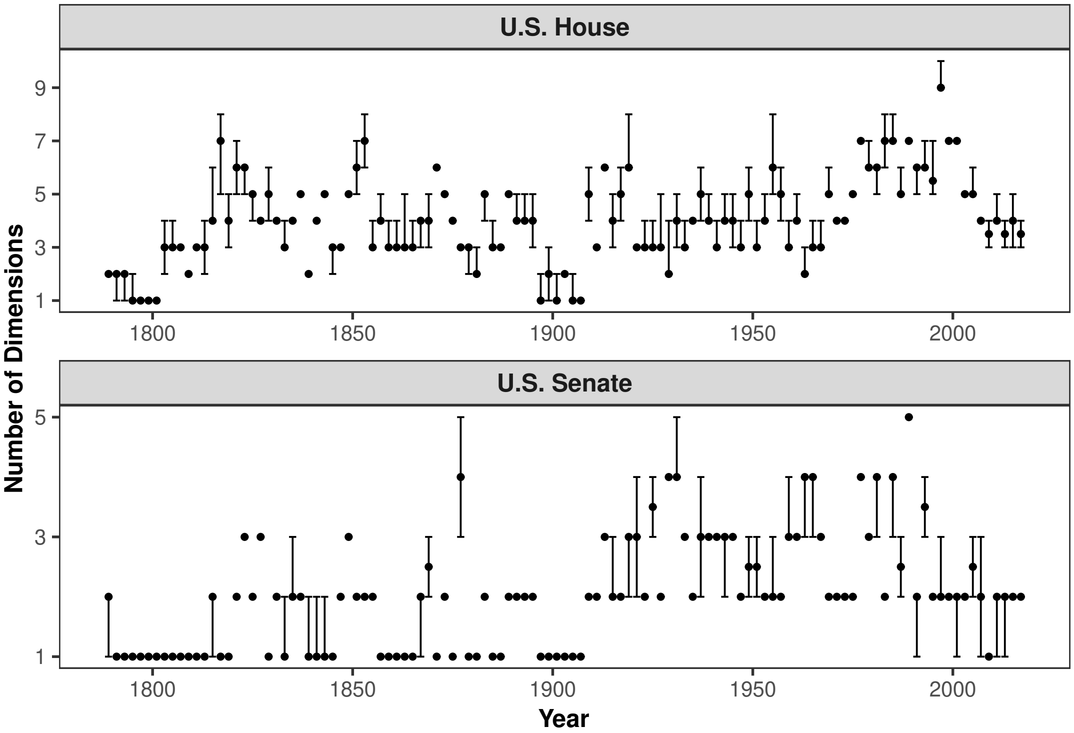
$$P(\lambda_{jk}|r_{jk}) \sim r_{jk}\ P(\lambda_{jk}) + (1 - r_{jk})\delta_0 \tag{2.12}$$

where δ_0 is a point mass PDF at zero and $P(\lambda_{jk})$ is the marginal prior on λ_{jk}. Thus, if $r_{jk} = 1$, λ_{jk} is allowed to take a non-zero value. On the other hand, if $r_{jk} = 0$, it is restricted to be equal to zero. This aspect of the model bridges the infinite number of dimensions implied by IBP and the finite number of dimensions that are actually observed - the first K^+ columns of $\boldsymbol{R}$ contain non-zero elements while the countably infinite set of columns after K^+ include only zeroes. In turn, the corresponding ideal points are simply nuisance parameters. In the context of roll call scaling, the rows of $\boldsymbol{R}$ tell whether or not a vote draws on a specific dimension when estimating the parameters of the underlying utility model.

Using the IBP prior on $\boldsymbol{R}$ allows for a full model definition, which is outlined in Section A.2 of the Appendix. The full model, a beta process item-response theory model (BPIRT), is a close analogue to the *infinite latent feature model* developed by Knowles and Ghahramani (2011). BPIRT has many of the same properties as the standard Bayesian IRT model (Clinton et al., 2004).

Under this specification, estimation of the model using Markov Chain Monte Carlo methods proceeds in a relatively straightforward manner. Technical details related to estimation of the BPIRT model are included in the appendix. The full conditional distributions and sampling methods are outlined in Section A.3. The BPIRT model uniquely identifies estimates for the structural parameters; this property is discussed in Section A.4. Methods for determining good starting values and assessing convergence of the Markov chains are outlined in Section A.5 and Section A.6. Finally, simulations which show how accurately BPIRT uncovers the binary matrix under a known model are shown in Section A.7.

FIGURE 2.1. Dimensionality and Corresponding 95% HPD Intervals Estimated by BPIRT for the 1st - 115th Sessions of the U.S. House and the U.S. Senate.



2.4 Multidimensionality in the U.S. Congress Over Time

The varying dimensions model of voting is well suited to examining the question of multidimensionality in voting and BPIRT provides a rigorous statistical tool for creating estimates of dimensionality in roll call voting. The results related to multidimensionality from BPIRT promise to provide insight into this problem.[5] I begin by assessing how many dimensions are estimated in the $1^{st} - 115^{th}$ $(1789 - 2017)$ sessions for each chamber of the U.S. Congress. Figure 2.1 shows the estimated number of dimensions and corresponding 95% highest posterior density intervals for each session within each chamber of Congress. This plot shows a generally multidimensional legislature. In the case of the U.S. House, sessions near the beginning of U.S. history and some sessions in the late 1800s and early 1900s have credible intervals that include a

[5] Additional figures and discussion for this section can be found in Section B of the appendix.

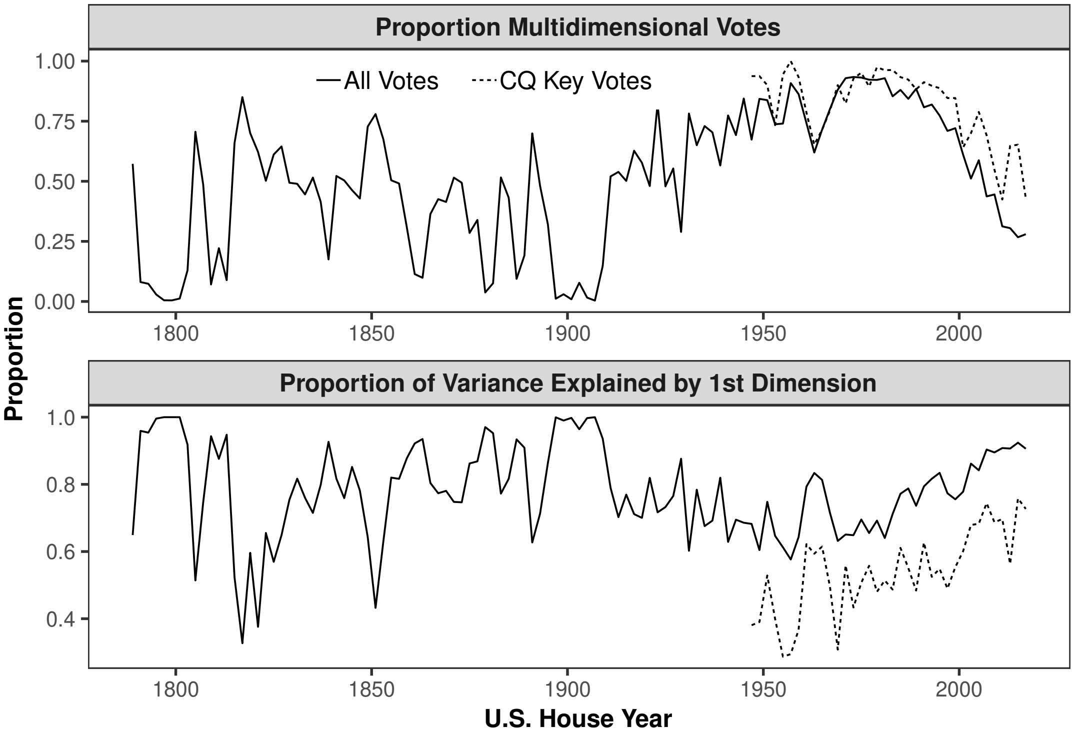
single dimension. However, the vast majority of sessions are estimated to need more than one dimension to best model roll call behavior.

A similar story is seen in the U.S. Senate. While there are more sessions which are estimated to require only one dimension, the majority require at least two dimensions to best model the roll call voting variation. One important caveat for these unidimensional sessions is that the IBP prior which drives BPIRT is limited in its ability to estimate multidimensionality when there are a small number of votes and/or voters. In early sessions of the U.S. House, there were less voting members and there were typically less votes than in more recent sessions of the U.S. House. As shown in simulations, BPIRT underestimates dimensionality in these settings, so there is reason to believe that this is the case for the U.S. House. Similarly, there are always around 100 voting U.S. Senators in a given session which inherently places a cap on the number of dimensions which can be modeled for a session of this chamber. This is not to say that these results should be discarded - rather, it is important to point this aspect of BPIRT out as a weakness for estimating dimensionality in smaller sets of roll call data.

In order to examine multidimensionality at the vote level, a measure of multidimensionality must be established (Roberts et al., 2016; Bateman et al., 2017; Smith, 2007). One method of summarizing the dimensionality of a single vote is to simply use the posterior probability that the vote took on more than one dimension in the binary matrix (MD). However, this metric suffers from minor theoretical deficiencies; it does not explain how much a vote needs each dimension. Thus, I use a second supplementary measure of multidimensionality - the proportion of variance explained (PVE) by each dimension on a specific vote:

$$\text{PVE}_{jk} = \frac{r_{jk}\lambda_{jk}^2}{\sum\limits_{h=1}^{K} r_{jh}\lambda_{jh}^2} \tag{2.13}$$

FIGURE 2.2. Proportion of Multidimensional Votes and Proportion of Variance Explained by the 1^{st} Dimension Estimated by BPIRT in the 1st - 115th Sessions of the U.S. House



Note: Values reported are posterior means.

Using PVE to examine the influence of the first dimension on a vote, PVE takes a value of one when only the first dimensions is needed. As the influence of other dimensions increase, PVE for the first dimensions decreases. Therefore, PVE measures the overall influence of a given dimensions on vote outcomes. MD and PVE are highly correlated, but provide different views of each dimension's necessity in the individual case.

Using the proposed measures of multidimensionality, I examine the role of the first dimension and the set of dimensions beyond the first estimated by BPIRT in explaining variation within U.S. House roll call data sets.[6] One of the many advantages of

[6] For this and the proceeding examinations of multidimensionality in roll call voting, I choose to only present results for the U.S. House. The trends and inferences made from U.S. House data are similar to those that are made from U.S. Senate data. For the sake of brevity, I withhold figures

the BPIRT approach is that these metrics can be examined for any subset of votes within the analyzed sets. An application of the property is examining the difference between the aggregate set of all roll calls and more important "key votes" (Smith, 2007; Roberts et al., 2016).

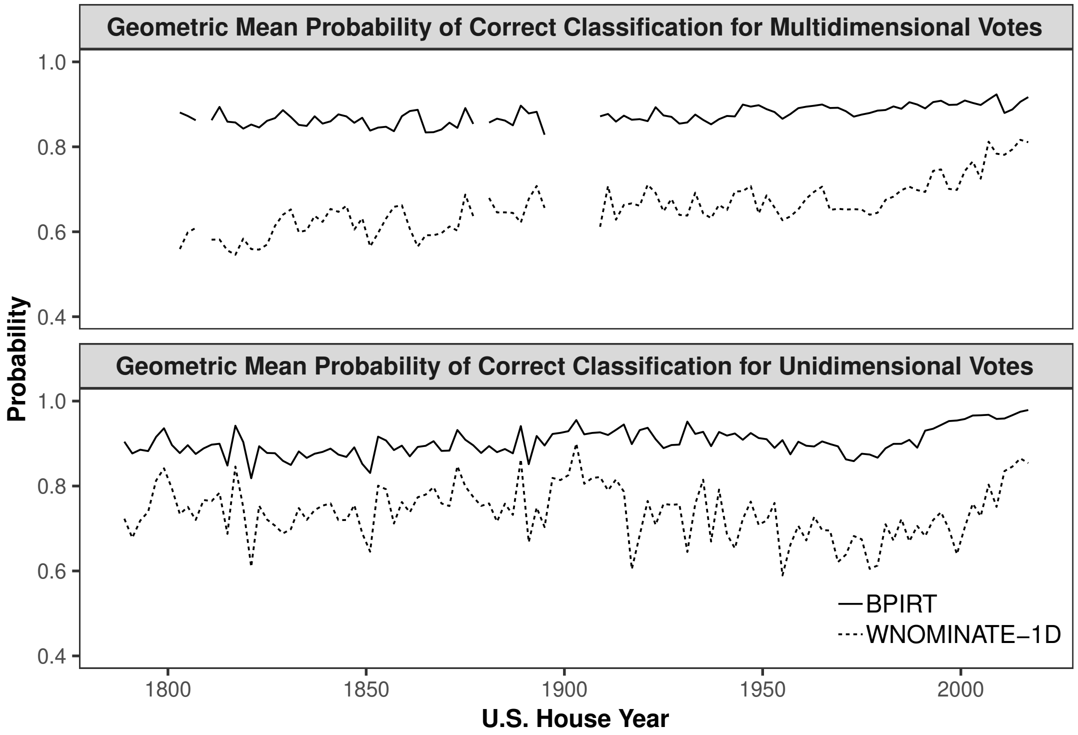
I examine MD and PVE for both the full aggregate sets of roll call votes for the $1^{st} - 115^{th}$ sessions of the U.S. House as well as the set of votes classified as "key votes" by Congressional Quarterly for the $80^{th} - 115^{th}$ sessions. Figure 2.2 shows these quantities over time for the U.S. House. Examining the proportion of multidimensional votes in each session shows highly multidimensional voting within the U.S. House, especially in the 20^{th} and 21^{st} centuries. Even in recent sessions of the U.S. House, which are considered to be extremely party driven and one-dimensional, a significant number of votes require more than one dimension to best explain variation in the roll call votes. On the other hand, the PVE for the first dimension is relatively high throughout time. While voting in recent sessions is certainly explained more heavily by the first dimension than in the mid and late 1900s, the reliance on the first dimension is equal to many non-unidimensional sessions during Reconstruction and the Great Depression.

Looking only at key votes provides support for the theory that important votes are multidimensional and require more than simply using the first dimension of ideal points.[7] While MD shows a modest difference between the aggregate roll call sets and the set of key votes, PVE shows that a significantly lower amount of variance can be explained by the first dimension in key votes. On average, approximately 25% less variance is explained using the first dimension. This finding along with conflicting

and other summaries of the U.S. Senate data in this paper. Results from my analysis of the U.S. Senate can be seen in the replication files included with this paper.

[7] For each session where CQ key votes were examined, there were between 80 and 300 votes that were classified as important votes by Congressional Quarterly.

FIGURE 2.3. Geometric Mean Probability of Correct Classification for Unidimensional and Multidimensional Votes in the 1st - 115th Sessions of the U.S. House



Note: Values reported are posterior means.

results for the aggregate roll call sets provides evidence for the aggregation hypotheses presented by Roberts et al. (2016) and should serve as a starting point for more fine-grained examinations of dimensionality in landmark legislation over time.

Examining vote level dimensionality is not the only way to demonstrate the necessity of dimensions past the first - BPIRT shows marked improvements over other roll call scaling techniques in terms of overall model fit. One method of comparison that rewards correct classification of model outcomes given the ideal points while also penalizing inefficient estimates is the geometric mean probability of correct classification (Carroll et al., 2009). For a given set of votes, the geometric mean probability of correct classification (GMP) is:

$$\text{GMP} = \left(\prod_{i=1}^{N} \prod_{j=1}^{P} P(\hat{y}_{ij} = y_{ij}) \right)^{\frac{1}{N*P}} \tag{2.14}$$

where $\hat{y}_{ij}$ is the predicted vote for a legislator and y_{ij} is the observed vote.

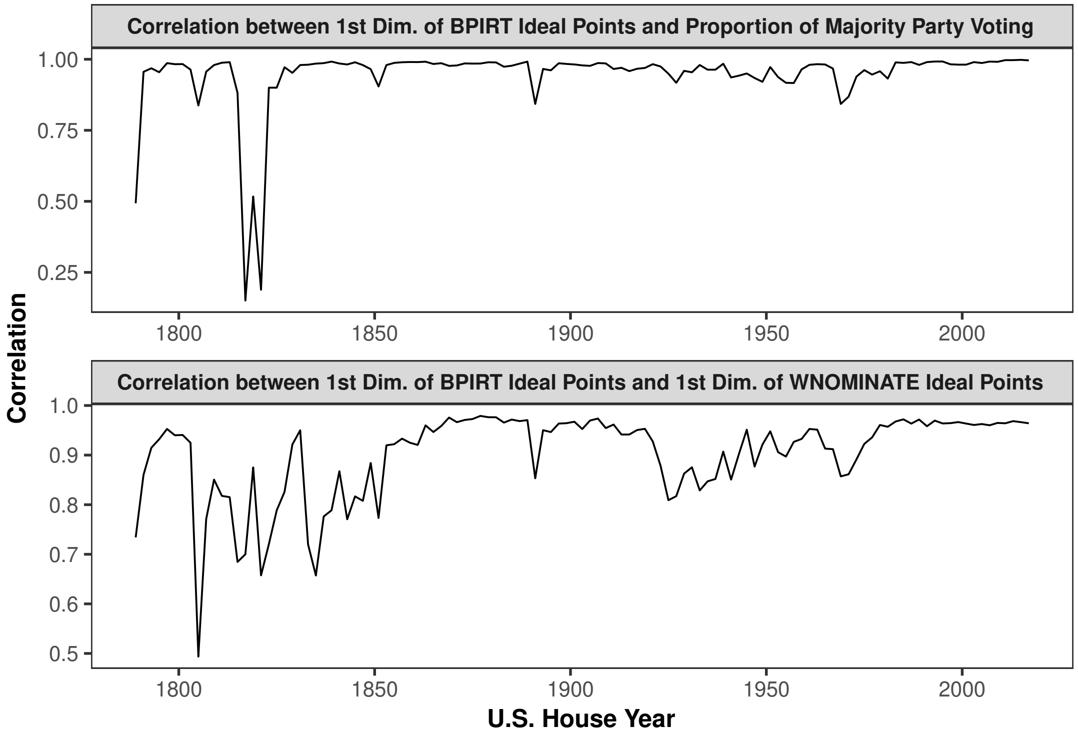
Figure 2.3 shows the GMP for each session of the U.S. House broken out by the dimensionality of the vote implied by BPIRT.[8] Unsurprisingly, there are significant gains made in model fit when examining multidimensional votes. The difference in model fit between WNOMINATE-1D and BPIRT on these votes is quite large. Combined with the knowledge that many votes within each session are multidimensional, this provides strong evidence that unidimensional models are missing out on a large portion of the variation which drives voting in the U.S. House. BPIRT also shows large gains over WNOMINATE-1D when analyzing unidimensional votes. This result is somewhat unintuitive as the underlying formal model for a unidimensional vote is essentially the same for BPIRT and WNOMINATE. However, this result can be attributed to proper placement of zeros in the binary matrix and, in turn, ensuring that each vote corresponds only to the correct subset of potential dimensions of the policy space.

2.4.1 Interpretation of Ideal Points

BPIRT paints a picture of a legislature that behaves in a multidimensional manner; while not all votes require multiple dimensions to explain voting patterns, many votes need something beyond a single dimensional ideal point to best explain voting. A natural question that follows pertains to the meaning of the dimensions - what is represented by the first dimension and what concepts are represented by dimensions

[8] Gaps in the multidimensional plot occur when BPIRT estimates that a session has only one dimension or less than 10 votes required more than one dimension to best model its vote outcomes.

FIGURE 2.4. Correlation between the 1st Dimension of Ideal Points Estimated by BPIRT, the Ideal Points from WNOMINATE-1D, and the Proportion of Majority Party Voting



Note: Values reported are posterior means.

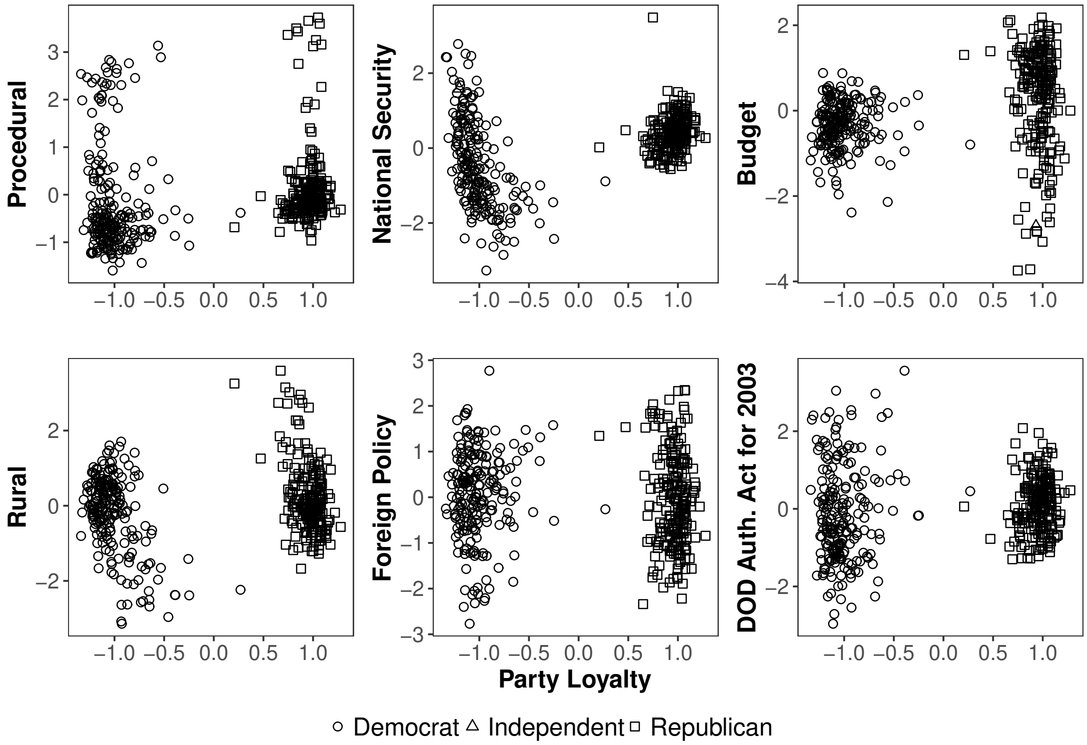
beyond the first?

First, I examine the meaning of the first dimension, over time. In particular, I examine whether the first dimension is simply providing a measure of the individual frequency of party bloc voting (Aldrich et al., 2014; Lee, 2009; Harbridge, 2015). In order to test this hypothesis, I measure how frequently a voting member of the legislature votes with that session's majority party.[9] Given that more than 99% of votes have a non-zero contribution from the first dimension, over time, understanding the meaning of this dimension is key.

Figure 2.4 shows the correlation between the ideal points from the first dimension

[9] Specifically, I determine the majority party vote on a given roll call vote to be the most frequently made choice by the members of the majority party in a given session.

FIGURE 2.5. Ideal Points and Dimensions Estimated by BPIRT for the 107th Session of the U.S. House (2001 - 2003)



Note: The reported ideal points are from the iteration of the MCMC procedure with the highest complete-data likelihood.

for each voting member compared to the proportion of votes for which they cast the same vote as the majority party preference. The relationship between majority party voting and the uncovered ideal points strongly supports the theory that the first dimension of BPIRT ideal points is simply modeling party teamsmanship. Figure 2.4 also shows that BPIRT and WNOMINATE are highly correlated over time. This, in turn, implies that the first dimension of WNOMINATE is largely estimating the same construct with the first dimension.

In order to demonstrate the importance of dimensions beyond the first in explaining legislative behavior that exists outside of party loyalty, I examine one particular session.[10] The 107^{th} session of the U.S. House took place between 2001 and 2003

[10] Roberts et al. (2016) and Bateman et al. (2017) show that dimensions beyond the first are

and contained the September 11^{th} attacks and the ensuing scramble from the U.S. government attempting to respond to domestic and foreign security concerns. These issues created strong divisions within parties and led to a number of outcomes that appeared to favor the pro-war members of the U.S. Congress. It is reasonable to expect that a significant portion of roll call votes in this session require dimensions beyond party loyalty when explaining variation and estimating ideal points.

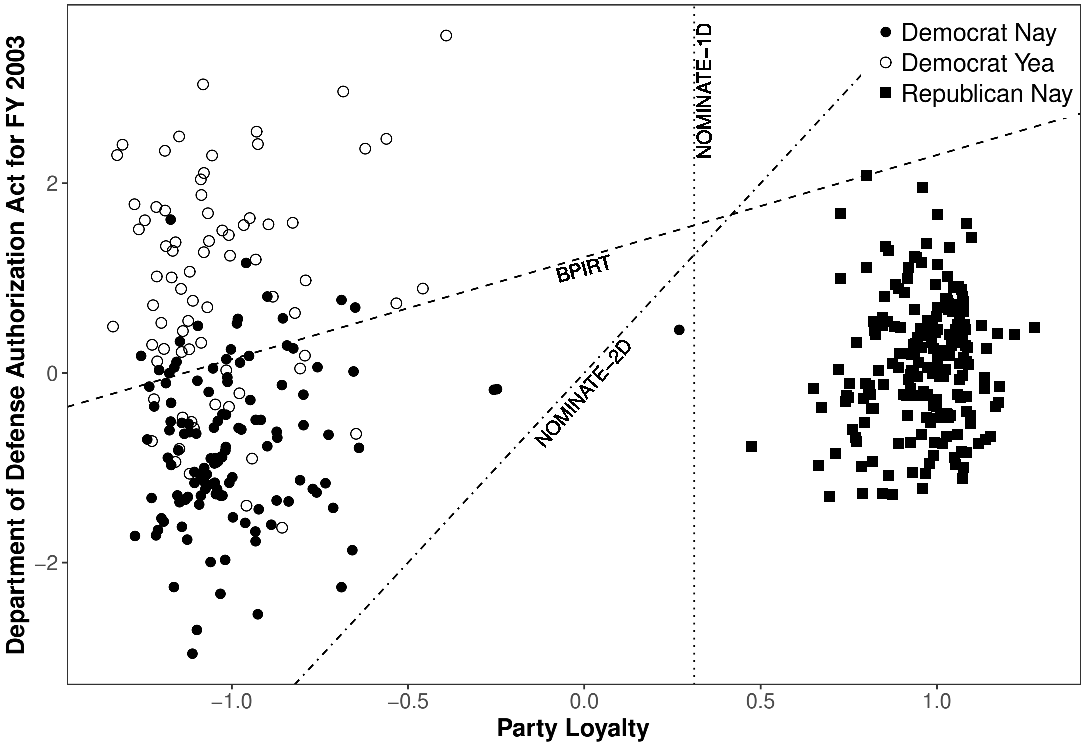
Figure 2.5 shows the seven dimensions of ideal points uncovered by BPIRT for the 107^{th} session of the U.S. House.[11] First, and foremost, the party loyalty dimension is highly apparent and shows a split between Republican and Democratic voting (PVE = .77). Other dimensions are important and explain the other 23% of explainable variance. Some of the dimensions relate closely to specific policy topic areas such as rural/infrastructure issues (PVE = .03) and the government budget (PVE = .05). Another dimension that emerges relates to purely procedural votes, such as ceremonial motions and approving the chamber's journal (PVE = .02). However, the set of dimensions that are most important to this session, beyond party loyalty, relate to the September 11^{th} attacks and the relating security measures. These dimensions include national security (PVE = .07), foreign policy (PVE = .04), and a dimension that relates to funding the war in Afghanistan (PVE = .02). Given the sets of issues that were salient in the 107^{th} session of the U.S. House, this set of dimensions beyond party voting makes sense.

Though all votes do require party loyalty to explain some of the variance, most votes require additional explanations from other sources. One example that is particularly relevant to the 107^{th} U.S. House relates to funding the war in Afghanistan. Votes

colored by the salient issues of their time and the preferences of the agenda setters. This make a general analysis of dimensions beyond the first difficult.

[11] Additional material about how I used vote classifications and bill summaries to determine names for each dimension and about how I determined the structure of dimensions beyond the first are included in Section B.3 of the appendix.

FIGURE 2.6. Votes and Cutlines for Department of Defense Authorization Act for Fiscal Year 2003 Vote in the 107th Session of the U.S. House (Roll Call No. 655)



Note: The reported ideal points and cutlines are from the iteration of the MCMC procedure with the highest complete-data likelihood.

related to funding military action arose after the September 11^{th} attacks. While the Republican party unanimously agreed to motions to increase funding to the Department of Defense for these actions, Democrats were split in these votes. Though the Republicans held the House majority, many Democrats used these votes to signify support for or against the war to their constituents and this created splits in the voting.

The BPIRT estimation procedure selects 14 of these votes and estimates that these votes require a common dimension in addition to the party loyalty dimension. While this dimension only accounts for around 2% of the total variance explained in this session, it models an important heterogeneity in Democrat voting. Figure 2.6 shows the vote outcome by party for one of these votes, which pertained to an amendment to the Department of Defense Authorization Act for 2003 proposed by Loretta Sanchez (D-CA). This figure shows the BPIRT ideal points of the voters in two dimensions: the party loyalty dimension and the DOD dimension. Additionally, I illustrate three separate cutlines which show the line on which a voter would be undecided between a “Yea” or “Nay” vote. When this vote is scaled using WNOMINATE with only one dimension, the cutline indicates perfect within party agreement. This is not the case and is indicative of a second dimension at play. However, WNOMINATE in two-dimensions misses the important cut needed for this vote. On the other hand, BPIRT creates a cutline that splits the Democrats into those that support higher funding and those that oppose spending increases. This example perhaps best demonstrates the differences between BPIRT and other roll call scaling procedures; BPIRT estimates dimensions as a function of clusters of votes that share similar voting patterns and finds dimensions that are necessary for modeling their outcomes. This consistency in topic is a feature unique to BPIRT and provides a tool that can create in-depth inference of the topics that drive legislative voting throughout U.S. history.

2.5 U.S. Legislative Voting and Multidimensionality

BPIRT provides a tool for analyzing roll call votes and understanding the dimensionality of votes as well as the issue sets that drive variation in voting within the U.S. legislative chambers. While BPIRT shows marked improvements over previous approaches to roll call scaling, its benefit can be seen as bridging the aggregate roll call scaling approaches of Poole and Rosenthal (1997) and the issue-specific approaches of Roberts et al. (2016) and Bateman et al. (2017). This gives rise to measures of multidimensionality that are comparable within and across sessions and provides a unique measure that can assess the impacts of multidimensionality on theories of legislative behavior.

I leverage the ideal points estimated by BPIRT and the corresponding measures of dimensionality to explore the relationship between the dimensionality of a vote and the outcomes that are predicted by models of U.S. legislative voting. While there are numerous examples of models that appeal to unidimensionality and test theories leveraging unidimensional NOMINATE scores, I turn my attention to two specific models: the pivotal voter model presented by Krehbiel (1998) and the party cartel model presented by Cox and McCubbins (2005). These two models are widely cited in studies of U.S. legislative behavior and seek to explain the ways in which the organization of legislative voting and parties influence voting outcomes. These two models differ in their explanations of how voting decisions are made, but strongly leverage a unidimensional policy space in the theoretical and empirical examinations of their theories.

Specifically, I seek to understand how robust these theories are to the assumption of unidimensionality. Theoretical outcomes under multidimensionality are well established (Shepsle, 1978; Shepsle and Weingast, 1981), but there are few empirical studies of the impact of multidimensionality on voting in the literature. Unidimensionality can be best described as a stabilizing assumption - when the underlying policy space is unidimensional, the outcome is predictable given assumptions about how legislators behave. In contrast, multidimensional votes are theoretically characterized by outcomes that can take any form. Thus, the goal is to measure the stability of outcomes implied by the pivotal voter and party cartel models when properly taking dimensionality into account.

As discussed previously, multidimensionality comes in many different shapes and sizes. For example, a vote can be multidimensional, but rely very heavily on one single dimension while there is only a small amount of variance explained by another set of dimensions. For this reason, I explore three separate ways in which multidimensionality may relate to vote instability:

1. **No Effect**: As the multidimensionality of a vote increases, there is no discernible change in the stability of voting outcomes.

2. **Continuous Effect**: Vote outcomes become more and more unstable as the multidimensionality of the vote increases; low levels of multidimensionality show more stable outcomes than votes with higher levels of multidimensionality.

3. **Threshold Effect**: Vote outcomes are stable and predictable up to a small amount of multidimensionality. Once this threshold is crossed, vote outcomes fundamentally change (McKelvey, 1976; Schofield, 1978). Even when the amount of multidimensionality is small, there is a marked difference between unidimensional and multidimensional outcomes.

Each of these mechanisms provide a different view of how theories of U.S. legislative voting might be influenced by the assumption of unidimensionality.

Evidence that multidimensionality influences vote outcomes has significant implications for the study of U.S. legislative voting. First, existing theories related to legislative voting must be examined for conditional relationships - if the vote is multidimensional, does the prediction from the theory change? Multidimensionality points to different factors that are necessary for contextualizing the conditions under which a vote are made. Second, the usage of unidimensional ideal points under evidence for multidimensionality leads to potential biases in further results. Given the interpretation of the first dimension examined previously, usage of unidimensional ideal points when multiple dimensions are needed essentially summarizes the level of majority party voting of a member while treating other sources of predictable roll call behavior as noise. Particularly when being used as proxies of preferences to test theories of party control, this can lead to acceptance of theories as a product of an endogenous measure. Finally, under the common assumption of rational voters, evidence that multidimensionality leads to more unpredictable outcomes points to ways in which rational proposers can skew proposals to their advantage (Riker, 1980; Shepsle, 1978; Shepsle and Weingast, 1981; Baron and Ferejohn, 1989). If multidimensional models are appropriate models of legislative voting, then the idea that strategic proposers can use multidimensionality to achieve better outcomes must be accounted for within theories related to the legislative process.

2.5.1 A Theory of Pivotal Voters

Perhaps one of the most well known theories of U.S. legislative behavior, Krehbiel (1998) outlines a theory of pivotal voters in legislative voting. Under this model, a proposal, the status quo, and voters are mapped to a unidimensional, commonly-known policy space. Under the rules of the legislative body, Krehbiel (1998) contends that policies must be proposed outside of the gridlock zone in order to pass the chamber. The gridlock zone is defined by the median voter, the presidential veto pivot, and (when appropriate) a Senate filibuster pivot. These members of the legislature effectively control the proposals which pass and, in turn, rational proposers craft legislation with these constraints in mind. This model of legislative voting is simple and effective, leading to many insights about periods of low and high gridlock within the U.S. legislature.

To examine how multidimensionality influences empirical support for the theory of pivotal politics, I recreate the empirical analysis from Krehbiel (1998, Chapter 5) that examines the behavior of filibuster pivots in successive cloture votes in the U.S. Senate. When a vote to invoke cloture goes before the U.S. Senate, Krehbiel (1998) contends that the filibuster pivot controls whether or not debate on a vote is stopped. An interesting test of this theory arises when a cloture vote occurs multiple times in the same bill-episode. The theory of pivotal voters claims that changes in individual votes from vote to vote are most likely to occur for members close to the filibuster pivot location. Using unidimensional NOMINATE scores to find voters close to the theoretical filibuster pivot, Krehbiel (1998) finds evidence for the pivotal voter model.

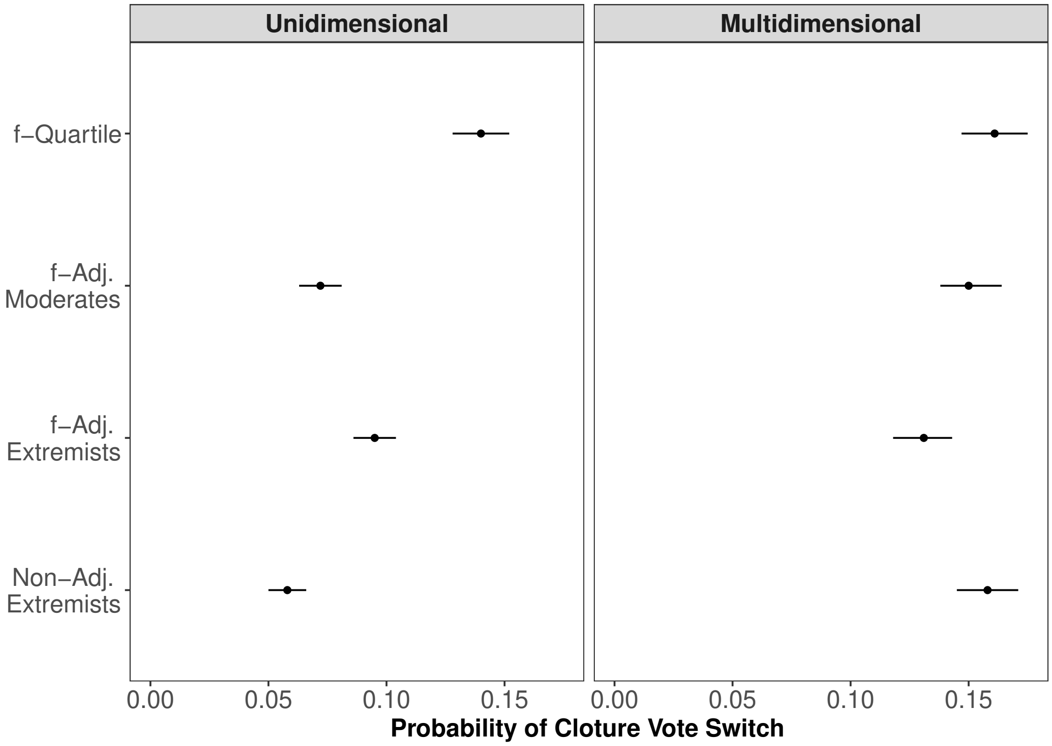
I contend that the empirical evidence shown by Krehbiel (1998) is driven by the unidimensional assumption made when using unidimensional NOMINATE scores; I expect that evidence for the pivotal voter model disappears in pairs of votes which are multidimensional. This conditional view of the pivotal politics model fits well within the original construction - if voters are able to collapse the preference space to a single dimension, then rational proposers can target changes in bills. However, under the multidimensional model, no such targeting can be made.

To examine these competing theories, I recreate this analysis on new data.[12] Using the set of all cloture votes that took place in the $89^{th} - 115^{th}$ sessions of the U.S. Senate, I examine all instances of votes to invoke cloture that occurred at least twice within the same bill-episode. I then grouped these into sequential vote pairs, mimicking the data set of Krehbiel (1998). This led to 477 cloture vote pairs over the course of time analyzed. For each vote pair, I then recorded whether each U.S. Senator switched their vote. This led to $44,710$ vote observations and $3,363$ vote switches. For each individual U.S. Senator, the ideal point associated with the first dimension of BPIRT scores was coded into quartiles: the f-quartile which contains the filibuster pivot, the f-adjacent moderates quartile, the f-adjacent extremists quartile, and the non-adjacent extremists quartile. Similarly, the controls outlined by Krehbiel (1998, Chapter 5) were recorded: President-side vetoes, voting under a unified government, and whether or not the voter was a Democrat.[13]

[12] Additional discussions about how I collected and processed the set of cloture votes and how I estimated the logistic regression model can be seen in Section C.1 of the appendix.

[13] Corresponding switches to the quartile coding were made when the vote met the conditions of a "president-side veto". For a further explanation of this control and the corresponding recoding, see Krehbiel (1998, Chapter 5, p. 106).

FIGURE 2.7. Probability of Cloture Vote Switch by Ideal Point Quartile



Note: Probabilities were calculated setting President Side, Unified Government, and Democrat to zero. Error bars show 95% HPD intervals.

To measure the dimensionality of a pair of cloture votes, I used the PVE for dimensions beyond the party-loyalty dimension in both votes. If both votes were estimated by BPIRT to be unidimensional, then dimensionality was coded as zero. Otherwise, dimensionality was set to be equal to the posterior mean value of PVE. The effect of multidimensionality is measured by testing three models with different underlying mechanisms. The theory of no effect was tested by not including a control for the dimensionality of the vote. Under the theory of continuous effect, multidimensionality was coded as the PVE attributed to dimensions beyond party loyalty. Finally, the threshold effect was tested by including a dichotomous variable for dimensionality coded to be multidimensional if the PVE for other dimensions in the vote pair was greater than .001. 43% of vote pairs were classified as multidimensional under the threshold model.

TABLE 2.1. Logistic Regression Results for Krehbiel's Cloture Vote Switching Example

	Dependent variable:		
	Senator Vote Switch in Cloture Vote Pair		
	No Effect	Continuous	Threshold
f-Quartile	0.40 (0.31,0.50)	0.63 (0.52,0.74)	0.96 (0.80,1.13)
f-Adj. Moderates	0.03 (-0.06,0.14)	0.17 (0.04,0.28)	0.22 (0.03,0.40)
f-Adj. Extremists	0.06 (-0.04,0.16)	0.25 (0.13,0.36)	0.52 (0.36,0.69)
President Side	-1.04 (-1.14,-0.95)	-1.03 (-1.13,-0.92)	-1.01 (-1.12,-0.92)
Unified Government	-0.46 (-0.55,-0.38)	-0.45 (-0.54,-0.36)	-0.42 (-0.51,-0.33)
Democrat	-0.40 (-0.47,-0.33)	-0.39 (-0.46,-0.31)	-0.36 (-0.43,-0.28)
Multidimensional		1.34 (1.07,1.63)	1.10 (0.94,1.26)
f-Quartile × Multidimensional		-2.10 (-2.64,-1.72)	-0.94 (-1.15,-0.73)
f-Adj. Moderates × Multidimensional		-1.03 (-1.47,-0.60)	-0.28 (-0.50,-0.05)
f-Adj. Extremists × Multidimensional		-1.61 (-2.08,-1.14)	-0.74 (-0.97,-0.52)
Intercept	-2.11 (-2.19,-2.02)	-2.29 (-2.39,-2.19)	-2.78 (-2.92,-2.63)
Observations	44,710	44,710	44,710
Log Marginal Likelihood	-11,609.54	-11,577.67	-11,464.47

[1] The comparison group is non-adjacent extremists.

[2] Coefficient values are posterior medians and values in parentheses are 95% HPD intervals.

Following the empirical exercise by Krehbiel (1998), I estimated the probability of a vote switch given the ideal point quartiles, dimensionality of the vote, and the controls.[14] Table 2.1 shows the results of this regression under the three different theories of how dimensionality might influence vote outcomes. Similarly, Figure 2.7 shows the probabilities of vote switching as a function of ideal point quartile and the dimensionality of the vote as defined by the threshold model. The results from the regressions show a number of interesting relationships. First, under the model of no effect, the results from Krehbiel (1998, Chapter 5) are largely replicated in the new data set - assuming that dimensionality has no effect on vote switching, voters in the f-quartile are most likely to switch votes between cloture vote pairs. This relationship is also seen in the other two models when the vote pair is unidimensional, providing strong support that the pivotal voter findings are supported in the strict unidimensional case. However, this relationship disappears in the multidimensional case. When the vote pair is multidimensional, there is no statistical difference between any quartile in the 95% HPD intervals.

Comparing the marginal likelihood across the models provides an assessment of the degree to which controlling for multidimensionality improves model fit. First and foremost, it is clear that the marginal likelihood for both the continuous and threshold models is lower than the model with no effect. Given that the marginal likelihood inherently penalizes against overfitting, this is strong evidence that multidimensionality explains a significant amount of variation within the data. Between these models, the threshold effect has the lowest marginal likelihood indicating that even a small amount of multidimensionality leads to a fundamentally different role of the pivotal voter. This result points to a chaotic result for almost 50% of the votes analyzed; the pivotal voter model provides no information about where vote switches will occur when the vote is multidimensional. This is one example of how properly accounting for multidimensionality in U.S. legislative voting can fundamentally change long held

[14] While a case could be made that session-level random effects are needed, I appeal to the response of Krehbiel (1998) that such controls would be “utterly atheoretic” as there are no session specific elements of the pivotal voter theory.

beliefs about legislative decision making.

2.5.2 A Theory on Party Control

In response to the work of Krehbiel (1998), scholars pointed to parties as another source that influences both the proposals that are made and the decisions made by their respective members. One example of this work is the party cartel theory of agenda control (Cox and McCubbins, 2005). Under this model, rational voters in a unidimensional preference space want to select proposals that are close to their ideal points. However, their desire to be reelected also influences their vote choices and they frequently delegate to the central authority of the party to make vote decisions. This leads to strong party control in vote choice which can lead to votes that are against their individual best actions conditional on their ideal points.

Party cartel theory leads to a number of empirically testable predictions about agenda control in the U.S. Congress. One specific example relates to final passage votes in the U.S. House (Cox and McCubbins, 2005, Chapter 5). For each final passage vote, the result can be classified by the proportion of members from the majority and minority parties voting in support of passage: if less than 50% of the minority party votes in support of passage, the vote is considered a minority roll with similar conditions defining a majority party roll. Under the theory of party cartels, these rolls occur in predictable ways. First, majority rolls should be rare and uniformly distributed conditional on the distance between the ideal point of the chamber median and the ideal point of the median member of the majority party. On the other hand, minority rolls should occur often with the frequency increasing as the distance between the floor median and the minority party median increases. Using unidimensional NOMINATE scores as estimates for the ideal points, Cox and McCubbins (2005, Chapter 5) find empirical evidence for both predictions.

As with Krehbiel (1998), I expect that this empirical evidence is colored by the usage of unidimensional ideal point estimates and the result is again conditional on the dimensionality of a vote. It is reasonable to believe that close alignment between the majority party median and the floor median indicates high levels of party loyalty and, in turn, produces strong agreements within the majority party on votes that can be thoroughly explained by party bloc voting. On the other hand, multidimensionality in a vote points to sources other than party voting and it is reasonable to expect that these votes have less predictable outcomes.[15]

To assess the role of multidimensionality, I examined the set of all final passage votes for the $83^{rd} - 115^{th}$ sessions of the U.S. House.[16] This data was retrieved from the Political Institutions and Public Choice Roll-Call Database (Crespin and Rohde, 2012). I determined if each final passage vote was a majority party roll, minority party roll, or if no roll had occurred. This led to $4,429$ observations of final passage votes with 127 majority party rolls and $1,743$ minority party rolls.[17] I used ideal point estimates from the first dimension of BPIRT for each session and recorded the location of the floor median and the respective party medians for each vote. The absolute difference between these two metrics constitutes the distance between medians.[18]

As before, the three theories of multidimensionality are tested. Under the theory of no effect, the probability of a roll is tested only as a function of distance between the respective medians. The continuous effect was tested by using the PVE for dimensions beyond the first on a given vote, the distance between medians, and a multiplicative

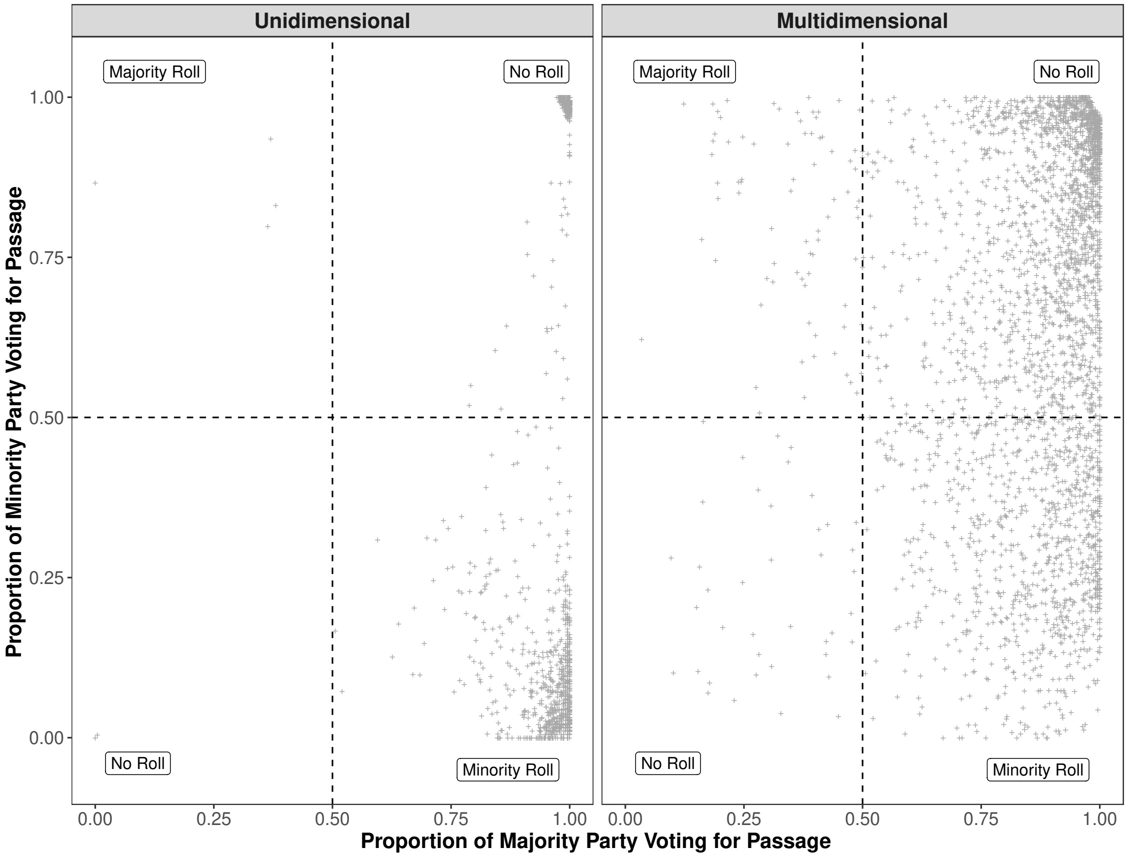
[15] It is worth noting that multidimensionality is mentioned by Cox and McCubbins (2005) and this possibility is explicitly considered, but not examined in depth.

[16] In line with Cox and McCubbins (2005), I restrict this set to votes which required a simple majority for passage.

[17] The rate of passage for final passage votes is around 98%.

[18] Additional discussions about how I collected and processed the set of final passage votes and how I estimated the regression model can be seen in Section C.2 of the appendix.

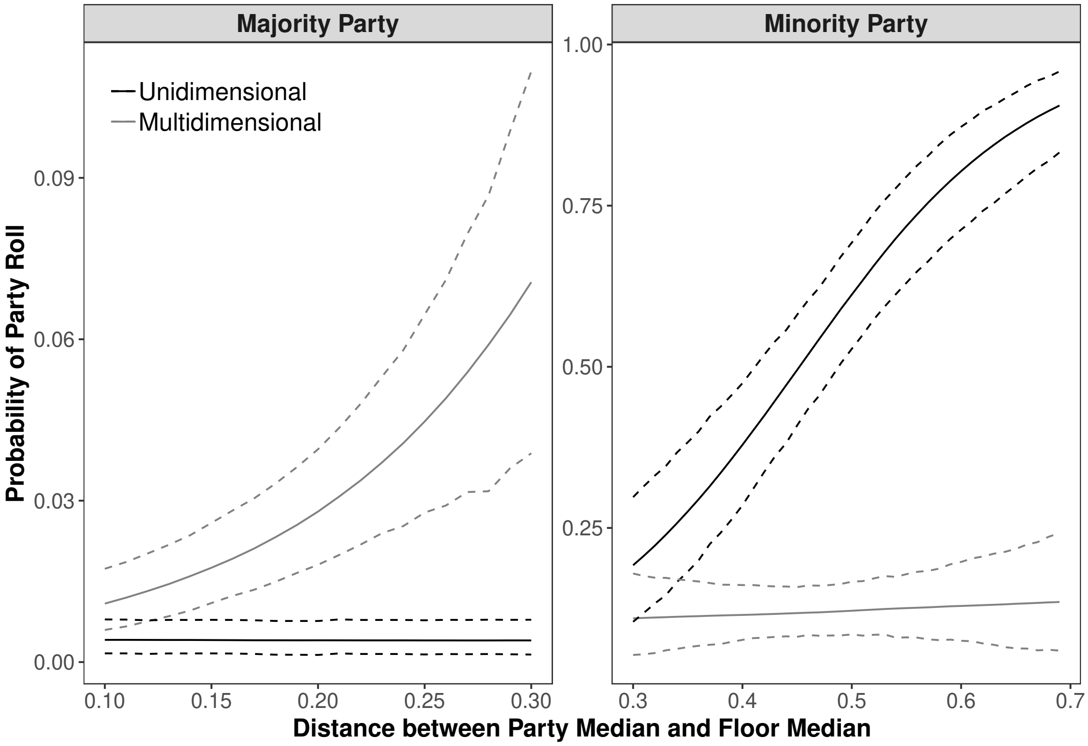
FIGURE 2.8. Results of Final Passage Votes in the 83rd - 115th Sessions of the U.S. House



interaction between the two. Finally, similar to the continuous model, the threshold effect was tested by classifying any vote where the PVE for dimensions beyond the first was greater than .001 as multidimensional. Under the threshold model, 68.8% of final passage votes were classified as multidimensional.

Figure 2.8 shows the set of final passage votes analyzed. Votes are compared based on the proportion of "Yea" votes cast by each party in each case. Votes are then classified as unidimensional or multidimensional using the threshold model. The difference in vote proportions between the two classes of votes is stark. On unidimensional votes, there are two general outcomes: near unanimous support by the entire set of voters or minority rolls with nearly unanimous support from the majority party. Majority party rolls on unidimensional votes are incredibly rare - only 4 out of 1,382 (.002%)

FIGURE 2.9. Probability of Party Rolls as a Function of Distance between the Party Median and the Floor Median



Note: Probabilities were calculated only within the range of observed outcomes for majority and minority party distances. Dotted lines show 95% HPD intervals.

unidimensional final passage votes result in majority party rolls. On the other hand, the results for multidimensional votes are more varied; the rate of minority rolls appears to be equivalent to the rate of votes where the minority party is not rolled and the rate of majority rolls (4%) is much higher than in the unidimensional case.

I used a logistic regression to examine the relationship between party median distance, multidimensionality, and party rolls. In order to best replicate the test from Cox and McCubbins (2005), I chose to include zero-mean session-level random intercepts estimated via a normal random effect. I report the variance of these parameters with the results from these regressions. Diffuse normal spike-and-slab priors with a spike at zero were placed on the regression coefficients.

Table 2.2 shows the results from these regressions. First, examining the model of no effect, the results from Cox and McCubbins (2005) are replicated. When all votes are assumed to be unidimensional, the probability that a final passage vote results in a majority party roll shows no evidence that it is influenced by the distance between medians. Similarly, there is a large positive correlation between the probability of a minority party roll and the distance between the minority party median and the floor median. However, there is strong evidence that controlling for the dimensionality of a vote explains more variation in both sets of party rolls. First, the DIC, a proxy for the marginal likelihood in hierarchical models that penalizes the addition of new parameters, is lower for the models that control for multidimensionality in both majority and minotiry party rolls. While this decrease is moderate in the majority rolls case, the minority rolls case shows a massive decrease in DIC. In each regression, the threshold model shows the smallest DIC, implying that a model that treats votes with even a small amount of variation that can be explained by dimensions beyond the first differently than unidimensional votes provides the best fit to the data.

Figure 2.9 shows the predicted probabilities of majority and minority party rolls as a function of distance between the respective medians and the dimensionality of a vote under the threshold model. These results point to a characterization of negative agenda control that is conditional on the dimensionality of the vote. First, the probability of a majority party roll is low in both unidimensional and multidimensional votes. While certainly lower in the unidimensional case, multidimensional votes show a predicted probability of around .06 in the case of the maximum observed distance between the floor and majority party medians. However, there is a statistically meaningful increase in the probability of majority rolls in the multidimensional case - while negative agenda control from the majority party is apparent, it seems that there are potentially other factors at play (Aldrich and Rohde, 2000).

On the other side, minority party voters benefit from multidimensionality. Under strict unidimensional votes, the minority party roll rate is positively correlated with the distance between medians. However, in multidimensional votes, knowing the distance between the minority party median and the floor median provides no information about the probability of a minority party roll. In other words, the predictions from party cartel theory relating to the minority party only apply to around 30% of final passage votes. This finding does not invalidate the party cartel theory. Rather, it points to multidimensionality creating cross-party support for bills that are not necessarily correlated with the number of times that a member votes with the majority party. Along with theory that strategic proposers use multidimensionality to create passing votes that cater to their own preferences, this opens a new door for research into the role of agenda control in the U.S. Congress under potentially highly multidimensional bills.

2.6 Conclusion

Roll call scaling and the operationalization of the spatial model is critical to the scientific examination and development of theories about how members of the U.S. Congress cast votes. While existing methods produce scores that appear to be best represented in a single dimension, I show that this finding is due to the tests used to determine dimensionality. In turn, I develop a varying dimensional representation of the spatial model and show a corresponding estimation technique that allows for estimation of both the aggregate-level dimensionality of the ideal point space as well as vote-specific sets of dimensions. Using this model, I show that there is little evidence for unidimensional ideal points in the U.S. House and U.S. Senate and that historical voting demonstrates multidimensional patterns. I present a set of ideal points that bridge the gap between the common aggregate methods and the more subject-specific examinations that are present in the roll call scaling literature. Under this new model, I then show that multidimensionality is an important aspect to be considered for further models of U.S. legislative voting that rely on ideal points as summaries of the preferences of voters. All in all, I show that BPIRT is a powerful procedure for determining the dimensionality of latent variables used in social science applications.

While the work in this article is catered to the study of roll call scaling, the models presented here are widely applicable to any study where latent variables are estimated via an item-response theory model. With light changes, Indian Buffet Process priors can be used to test dimensionality in a variety of settings and can be used as validation for claims related to the dimensionality of a latent space. Similarly, there are numerous extensions which can be made to the model presented in this paper. Under the Bayesian nonparametric framework, it is possible to examine how dimensions change over time by providing sufficient changes to the underlying priors. Similarly, different methods of clustering can be used to find interesting similarities in voting between both voters and the topics of the votes, themselves. This model serves a starting point into more complex examinations of what all can be learned from the spatial model of voting and scaling techniques.

TABLE 2.2. Logistic Regression Results for Cox and McCubbins' Final Passage Vote Example

	Dependent variable:		
	Majority Party Roll on Final Passage Vote		
	No Effect	Continuous	Threshold
Distance	0	0	0
	(0,2.31)	(-2.63,0)	(-3.5,0)
Multidimensional		0	0
		(0,0)	(0,0)
Distance ×		10.08	9.75
Multidimensional		(6.68,13.80)	(6.51,13.27)
Intercept	-3.79	-5.55	-5.48
	(-4.32,-3.38)	(-6.35,-4.72)	(-6.28,-4.66)
Variance of	0.72	0.68	0.66
Random Effects	(0.43,1.05)	(0.4,.99)	(0.38,0.97)
Observations	4,429	4,429	4,429
DIC	1131	1096	1077

	Dependent variable:		
	Minority Party Roll on Final Passage Vote		
	No Effect	Continuous	Threshold
Distance	7.78	10.04	9.44
	(4.82,10.45)	(7.17,12.91)	(6.52,12.48)
Multidimensional		3.60	2.00
		(2.16,4.88)	(1.22,2.83)
Distance ×		-15.22	-8.88
Multidimensional		(-18.26,-12.46)	(-10.71,-7.23)
Intercept	-4.15	-4.38	-4.26
	(-5.48,-2.74)	(-5.75,-2.95)	(-5.72,-2.82)
Variance of	0.29	0.24	0.25
Random Effects	(0.2,0.41)	(0.15,0.35)	(0.15,0.37)
Observations	4,429	4,429	4,429
DIC	4994	4264	4158

[1] Coefficient values are posterior medians and values in parentheses are 95% HPD intervals.

[2] Models were estimated with spike-and-slab priors on the coefficients. The spike was placed at zero. Posterior values equal to zero arise when the coefficient is not statistically distinguishable from zero.

[3] Congress-level random effects were estimated for the intercept terms only.

Bibliography

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

Aldrich, John H , Jacob M Montgomery, and David B Sparks (2014). Polarization and ideology: Partisan sources of low dimensionality in scaled roll call analyses. *Political Analysis 22*(4), 435–456.

Aldrich, John H and David W Rohde (2000). *The logic of conditional party government: Revisiting the electoral connection.* PIPC.

Baron, David P and John A Ferejohn (1989). Bargaining in legislatures. *American political science review 83*(4), 1181–1206.

Bateman, David A , Joshua D Clinton, and John S Lapinski (2017). A house divided? roll calls, polarization, and policy differences in the us house, 1877–2011. *American Journal of Political Science 61*(3), 698–714.

Bhattacharya, Anirban and David B Dunson (2011). Sparse bayesian infinite factor models. *Biometrika*, 291–306.

Binder, Sarah A (1999). The dynamics of legislative gridlock, 1947 96. *American Political Science Review 93*(3), 519–533.

Bonica, Adam (2014). Mapping the ideological marketplace. *American Journal of Political Science 58*(2), 367–386.

Carroll, Royce , Jeffrey B Lewis, James Lo, Keith T Poole, and Howard Rosenthal (2009). Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap. *Political Analysis 17*(3), 261–275.

Cattell, Raymond B (1966). The scree test for the number of factors. *Multivariate behavioral research 1*(2), 245–276.

Clinton, Joshua , Simon Jackman, and Douglas Rivers (2004). The statistical analysis of roll call data. *American Political Science Review 98*(2), 355–370.

Cox, Gary W and Mathew D McCubbins (2005). *Setting the agenda: Responsible party government in the US House of Representatives.* Cambridge University Press.

Cox, Gary W and Keith T Poole (2002). On measuring partisanship in roll-call voting: The US House of Representatives, 1877-1999. *American Journal of Political Science*, 477–489.

Crespin, Michael H and David Rohde (2012). Political institutions and public choice roll-call database. retrieved from https://ou.edu/carlalbertcenter/research/pipc-votes/.

Crespin, Michael H and David W Rohde (2010). Dimensions, issues, and bills: Appropriations voting on the House floor. *The Journal of Politics 72*(4), 976–989.

Denwood, Matthew J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software 71*(9), 1–25.

Doshi, Finale , Kurt Miller, Jurgen V Gael, and Yee W Teh (2009). Variational inference for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pp. 137–144.

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

Geweke, John and Guofu Zhou (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies 9*(2), 557–587.

Ghahramani, Zoubin and Thomas L Griffiths (2006). Infinite latent feature models and the Indian buffet process. In *Advances in neural information processing systems*, pp. 475–482.

Harbridge, Laurel (2015). *Is Bipartisanship Dead?: Policy Agreement and Agenda-setting in the House of Representatives.* Cambridge University Press.

Heckman, James J and James M Snyder Jr (1996). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. Technical report, National Bureau of Economic Research.

Hjort, Nils Lid (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 1259–1294.

Hurwitz, Mark S (2001). Distributive and partisan issues in agriculture policy in the 104th House. *American Political Science Review 95*(4), 911–922.

Jessee, Stephen and Neil Malhotra (2010). Are congressional leaders middlepersons or extremists? yes. *Legislative Studies Quarterly 35*(3), 361–392.

Kim, In Song , John Londregan, and Marc Ratkovic (2018). Estimating Spatial Preferences from Votes and Text. *Political Analysis 26*(2), 210–229.

Knowles, David and Zoubin Ghahramani (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 1534–1552.

Krehbiel, Keith (1998). *Pivotal politics: A theory of US lawmaking.* University of Chicago Press.

Lauderdale, Benjamin E and Alexander Herzog (2016). Measuring political positions from legislative speech. *Political Analysis 24*(3), 374–394.

Lee, Frances E (2009). *Beyond ideology: Politics, principles, and partisanship in the US Senate.* University of Chicago Press.

Leeper, Thomas J. (2015). *RPublica: ProPublica API Client.* R package version 0.1.3.

Londregan, John (1999). Estimating legislators' preferred points. *Political Analysis 8*(1), 35–56.

Martin, Andrew D. , Kevin M. Quinn, and Jong Hee Park (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software 42*(9), 22.

McKelvey, Richard D (1976). Intransitivities in multidimensional voting models and some implications for agenda control. *Journal of Economic theory 12*(3), 472–482.

Murray, Jared S , David B Dunson, Lawrence Carin, and Joseph E Lucas (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association 108*(502), 656–665.

Norton, Noelle H (1999). Uncovering the dimensionality of gender voting in Congress. *Legislative Studies Quarterly*, 65–86.

Paisley, John and Lawrence Carin (2009). Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 777–784.

Plummer, Martyn (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Poole, Keith T and Howard Rosenthal (1997). *Congress: A political-economic history of roll call voting.* Oxford University Press on Demand.

Poole, Keith T and Howard Rosenthal (2012). Voteview. *University of California, San Diego. www. voteview. com. Poole, Keith T.*

Poole, Keith T and Howard L Rosenthal (2011). *Ideology and congress*, Volume 1. Transaction Publishers.

Ramey, Adam (2016). Vox populi, vox dei? Crowdsourced ideal point estimation. *The Journal of Politics 78*(1), 281–295.

Riker, William H (1980). Implications from the disequilibrium of majority rule for the study of institutions. *American Political Science Review 74*(2), 432–446.

Rivers, Douglas (2003). Identification of multidimensional spatial voting models. *Typescript. Stanford University*.

Roberts, Jason M , Steven S Smith, and Stephen R Haptonstahl (2016). The dimensionality of congressional voting reconsidered. *American Politics Research 44*(5), 794–815.

Schofield, Norman (1978). Instability of simple dynamic games. *The Review of Economic Studies 45*(3), 575–594.

Shepsle, Kenneth A (1978). *The giant jigsaw puzzle: Democratic committee assignments in the modern House*. University of Chicago Press Chicago.

Shepsle, Kenneth A and Barry R Weingast (1981). Structure-induced equilibrium and legislative choice. *Public choice 37*(3), 503–519.

Smith, Steven S (2007). *Party influence in Congress*. Cambridge University Press.

Tahk, Alexander (2018). Nonparametric ideal-point estimation and inference. *Political Analysis 26*(2), 131–146.

Tausanovitch, Chris and Christopher Warshaw (2017). Estimating candidates' political orientation in a polarized congress. *Political Analysis 25*(2), 167–187.

Tsai, Tsung-han , Jeff Gill, and Jonathan Ripkin (2012). superdiag: R Code for Testing Markov Chain Nonconvergence.

A The BPIRT Model and Estimation

A.1 Beta Processes

A beta process is a random discrete measure that is completely described by a countably infinite set of atoms, where each atom has a finite mass determined from a stick-breaking process (Hjort, 1990). Unlike the well-known Dirichlet process (Ferguson, 1973), the probabilities that an individual unit belongs to a set of potential groups do not have to sum to one. Rather, the masses must only be between zero and one. The beta process is then used as a base measure for a Bernoulli process. In other words, a beta process yields a stochastic process for binary random variables or *feature selection.*

Definition 1. *Let Ω be a measurable space and $\mathbb{B}$ be its σ-algebra. Let H_0 be be a continuous probability measure on $(\Omega, \mathbb{B})$ and α a positive scalar. Assume that Υ can be divided into K disjoint partitions, $(B_1, B_2, ..., B_K)$. The corresponding beta process is generated as:*

$$H(B_k) \sim Beta(\alpha H_0(B_k), \alpha(1 - H_0(B_k))) \tag{2.15}$$

where $Beta(\cdot, \cdot)$ corresponds to the standard two-parameter beta distribution. Allow $K \to \infty$ and $H_0(B_k) \to 0$, then $H \sim BP(\alpha H_0)$.

The beta process can be written in set-function form:

$$H(\nu) = \sum_{k=1}^{\infty} \pi_k \delta_{\nu,k}(\nu) \tag{2.16}$$

where $H(\nu_i) = \pi_i$ and $\delta_{\nu,k}(\nu)$ is an arbitrary measure on ν. In the case of the beta process, π does not serve as a PMF. Rather, π serves as part of a new measure that parameterizes a Bernoulli process:

Definition 2. *Let the column vector, r_j, be infinite and binary with the k^{th} value,*

$r_{j,k}$:

$$r_{i,k} \sim Bern(\pi_k) \tag{2.17}$$

The new measure on the measurable space, Υ, is drawn from a Bernoulli process.

By arranging the samples for a set of infinite vectors as a matrix, we can see that a beta process is a prior over an infinite binary matrix with each row corresponding to a location in the measurable space.

A.2 BPIRT Full Model Specification

Beginning with the likelihood of the data, a full model specification for the BPIRT model can be defined. First, recall that the binary random variable is projected to a latent continuous space through data augmentation (Albert and Chib, 1993) such that:

$$x_{i,j} \sim \begin{cases} \mathcal{TN}_{-\infty,0}(\boldsymbol{\lambda_j}\boldsymbol{\omega_i} - \alpha_j, 1) \text{ if } y_{i,j} = 0 \\ \mathcal{TN}_{0,\infty}(\boldsymbol{\lambda_j}\boldsymbol{\omega_i} - \alpha_j, 1) \text{ if } y_{i,j} = 1 \\ \mathcal{N}(\boldsymbol{\lambda_j}\boldsymbol{\omega_i} - \alpha_j, 1) \text{ if } y_{i,j} \text{ is missing} \end{cases} \tag{2.18}$$

Then, the BPIRT model can be defined as:

$$\begin{aligned} &P(x_{i,j}|-) \sim \mathcal{N}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j, 1) \\ &P(\boldsymbol{\omega}_i) \sim \mathcal{N}_K(\boldsymbol{0}, \boldsymbol{I}_K) \\ &P(\lambda_{j,k}|r_{j,k}) \sim r_{j,k}\mathcal{N}_p(\lambda_{j,k}; 0, \gamma_k^{-1}) + (1 - r_{j,k})\delta_0 \\ &P(r_{j,k}) \sim \text{Bern}(\pi_k) \\ &P(\pi_k) \sim \text{Beta}(a/K, b(K-1)/K) \\ &P(\gamma_k) \sim \text{Gamma}(c, d) \end{aligned} \tag{2.19}$$

In all cases, intentionally vague or improper uniform priors are used on the structural parameters. Similarly, conjugate priors are used for convenience in estimation. While there is debate as to the impact of these choices (Murray et al., 2013), simulation shows that these choices are relatively innocuous given the size of the standard roll call data set.

Under a large, but finite K that approximates an infinite dimensional representation of $\boldsymbol{R}$, the model can be estimated with an explicit beta-Bernoulli prior on the elements of the binary matrix (Paisley and Carin, 2009; Doshi et al., 2009). If $K = \infty$, then we use the Indian Buffet Process prior (Ghahramani and Griffiths, 2006). I choose to use the explicit infinite approach in this article.

A.3 MCMC For BPIRT

Estimation of the BPIRT model uses the following sampling routine (Knowles and Ghahramani, 2011):

1. **Sample Continuous Mappings for the Binary Random Variables, $\boldsymbol{X}$.** For each $i \in (1, ..., N)$ and $j \in (1, ..., P)$, sample $x_{i,j}$ from a truncated normal distribution according to:

$$x_{i,j} \sim \begin{cases} \mathcal{TN}_{-\infty,0}(\boldsymbol{\lambda}_j \boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{i,j} = 0 \\ \mathcal{TN}_{0,\infty}(\boldsymbol{\lambda_j} \boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{i,j} = 1 \\ \mathcal{N}(\boldsymbol{\lambda}_j \boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{i,j} \text{ is missing} \end{cases} \tag{2.20}$$

2. **Sample R and Λ jointly.**

Sampling Currently Observed Features

Define K^+ as the current number of active features. For each $j \in (1, ..., p)$

and $k \in (1, ..., K^+)$ define:

$$\begin{aligned} t_{j,k} &= \frac{P(r_{j,k} = 1|\boldsymbol{X}, -)}{P(r_{j,k} = 0|\boldsymbol{X}, -)} \\ &= \frac{P(\boldsymbol{X}|r_{j,k} = 1, -)}{P(\boldsymbol{X}|r_{j,k} = 0, -)} \frac{P(r_{j,k} = 1)}{P(r_{j,k} = 0)} \end{aligned} \tag{2.21}$$

$$\frac{P(\boldsymbol{X}|r_{j,k} = 1, -)}{P(\boldsymbol{X}|r_{j,k} = 0, -)} = \sqrt{\frac{\gamma_k}{\gamma}} \exp\left(\frac{1}{2}\gamma\mu^2\right) \tag{2.22}$$

$$\frac{P(r_{j,k} = 1)}{P(r_{j,k} = 0)} = \frac{m_{-j,k}}{P - m_{-j,k} + 1} \tag{2.23}$$

where $\gamma = \boldsymbol{\omega}_k'\boldsymbol{\omega}_k + \gamma_k$, $\mu = \frac{1}{\gamma}\boldsymbol{\omega}_k'\hat{\boldsymbol{E}}_j$, $\hat{\boldsymbol{E}}_j = \boldsymbol{x}_j - \boldsymbol{\lambda}_j\boldsymbol{\Omega} + \alpha_j$ setting $\lambda_{j,k} = 0$, and $m_{-j,k} = -r_{j,k} + \sum\limits_{h=1}^{p} r_{h,k}$. Let

$$p_{r=1} = \frac{t_{j,k}}{1 + t_{j,k}}$$

then sample $P(r_{j,k}|-) \sim \text{Bern}(r_{j,k}; p_{r=1})$. If $r_{j,k} = 1$, then sample $P(\lambda_{j,k}|-) \sim \mathcal{N}(\lambda_{j,k}; \mu, \gamma^{-1})$. Otherwise, set $\lambda_{j,k} = 0$.

Sampling New Features

Begin by sampling the two Indian Buffet Process hyperparameters, a and b. Sample a from the full conditional:

$$P(a|-) \sim \text{Gamma}\left(a; K^+, H_P(b)\right) \tag{2.24}$$

where $H_P(b) = \sum\limits_{h=1}^{P} \frac{b}{b+h-1}$. b must be drawn via a Metropolis-Hastings step. Draw a new proposal:

$$P(b^*) \sim \text{Gamma}(b^*, 1, 1) \tag{2.25}$$

Accept b^* with probability $\min(1, r_{b \to b^*})$:

$$r_{b \to b^*} = \frac{(ab^*)^{K^+} \exp[-aH_P(b^*)] \prod_{k=1}^{K^+} \mathbb{B}(m_k, P - m_k + b^*)}{(ab)^{K^+} \exp[-aH_P(b)] \prod_{k=1}^{K^+} \mathbb{B}(m_k, P - m_k + b)} \tag{2.26}$$

where $\mathbb{B}(\cdot, \cdot)$ is the Beta function and $m_k = \sum_{j=1}^{P} r_{j,k}$.

For each $j \in (1, ..., P)$, sample the new number of dimensions to try:

$$P(\kappa_j) \sim \text{Pois}\left(\kappa_j; \frac{ab}{b + P - 1}\right) \tag{2.27}$$

Knowles and Ghahramani (2011) discuss ways to make this proposal explore the feature space in a faster way.

Draw values for each of the new potential dimensions. For $q \in (1, ..., \kappa_j)$:

$$P(\lambda_{j,q}) \sim \mathcal{N}(\lambda_{j,q}; 0, 1) \tag{2.28}$$

which will be referred to collectively as $\boldsymbol{\lambda}_{j,\kappa_j}$.

Using this as the proposal for a Metropolis-Hastings draw, accept the new dimensions with probability $\min(1, r_{\eta \to \eta^*})$:

$$r_{\eta \to \eta^*} = (2\pi)^{(N\kappa_j)/2} |\boldsymbol{M}|^{-N/2} \exp\left[.5 \sum_{i=1}^{N} \boldsymbol{m}_i' \boldsymbol{M} \boldsymbol{m}_i\right] \tag{2.29}$$

where $\boldsymbol{M} = \boldsymbol{\lambda}_{j,\kappa_j}' \boldsymbol{\lambda}_{j,\kappa_j} + \boldsymbol{\mathcal{I}}_{\kappa_j}$, $\boldsymbol{m}_i = \boldsymbol{M}^{-1} \boldsymbol{\lambda}_{j,\kappa_j} \hat{E}_{i,j}$, and $\hat{\boldsymbol{E}} = \boldsymbol{X} - \boldsymbol{\Lambda\Omega} + \boldsymbol{\alpha}$.

If the new proposal is accepted, set $\boldsymbol{\lambda}_{j,(K^+ + 1:K^+ + \kappa_j)}$ to the proposed values. Scheduling this part of the algorithm after updating values of $\boldsymbol{\Lambda}$ improves mixing. Set K^+ to the new number of columns represented in $\boldsymbol{\Lambda}$.

3. **Remove Inactive Features and Normalize $\boldsymbol{\Lambda}$.** For each $k \in (1, ..., K^+)$, if $r_{j,k} = 0 \; \forall \; 1 \leq j \leq p$, remove k from the analysis. Recalculate K^+. Post-process $\boldsymbol{\Lambda}$ to normalize the variance. For each $j \in (1, ..., p)$ and $k \in (1, ..., K^+)$ set $\boldsymbol{\lambda}_{j,k}$:

$$\lambda_{j,k} = \frac{\lambda_{j,k}}{\sqrt{1 + \sum_{h=1}^{K^+} \lambda_{j,h}^2}} \tag{2.30}$$

4. **Sample $\boldsymbol{\Omega}$.** For each $i \in (1, ..., n)$, sample $\boldsymbol{\omega}_i \in \mathbb{R}^{K^+}$ from:

$$P(\boldsymbol{\omega}_i | -) \sim \mathcal{N}_{K^+}(\boldsymbol{\omega}_i; (\boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \boldsymbol{\mathcal{I}}_{K^+})^{-1}\boldsymbol{\Lambda}'\boldsymbol{y}_i, (\boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \boldsymbol{\mathcal{I}}_{K^+})^{-1}) \tag{2.31}$$

5. **Sample Item Level Intercepts, $\boldsymbol{\alpha}$.** For each $j \in (1, ...p)$, sample the item level intercept from:

$$P(\alpha_j | -) \sim \mathcal{N}\left(\bar{\mu}_j, \frac{1}{N^2} \sum_{i=1}^{N} (\mu_{i,j} - \bar{\mu}_j)^2\right) \tag{2.32}$$

   where $\mu_{i,j} = \boldsymbol{\lambda}_j' \boldsymbol{\omega}_i - x_{i,j}$ and $\bar{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} \mu_{i,j}$.

6. **Sample Factor Precisions, γ_k.** For each $k \in (1, ..., K^+)$, sample γ_k from:

$$P(\gamma_k | -) \sim \text{Gamma}\left(\frac{m_k}{2}, \sum_{j=1}^{p} \lambda_{j,k}^2\right) \tag{2.33}$$

   where m_k is the number of sources for which feature $k \in (1, ..., K)$ is active.

A.4 Identification of Structural Parameters in BPIRT

One model consideration which requires further examination relates to identification of the structural parameters. It is well known that ideal points estimated using latent variable models are unidentified without further constraints (Rivers, 2003).

Identification can be induced by placing constraints $K(K+1)$ ideal points or requiring that the matrix of discrimination parameters have a lower-block triangular structure with positive elements on the main diagonal (Geweke and Zhou, 1996).

Using BPIRT, neither of these approaches are viable - since the number of columns in the matrix of discrimination parameters is technically infinite, placing *a priori* constraints is not possible. Fortunately, the sparsity inducing beta process prior on the discrimination parameters ensures identification. First, the number of votes which take on a feature necessarily decreases as the cardinality of the feature increases (Bhattacharya and Dunson, 2011). For example, if $r_{j,1} = 1 \; \forall \; j \; \in \; (1, ..., P)$ and $\sum_{j=1}^{P} r_{j,1} = P$, then $\sum_{j=1}^{P} r_{j,1} < P$. This ensures that $\sum_{j=1}^{P} (1 - r_{j,k}) \geq k \; \forall \; k \; \in \; (1, ..., \infty)$. Second, the spike-and-slab priors on the matrix of discrimination parameters ensures that each element of this matrix has a posterior distribution completely contained on one side of zero (Knowles and Ghahramani, 2011). Together, these two features effectively mimic the requirements for identification presented by Geweke and Zhou (1996) and ensure that all structural parameters estimated using BPIRT are uniquely identified. This is echoed by examining the posterior distributions for the ideal points, which never exhibit bimodality, evidence that the sign-switching problem is not present in MCMC estimation.

A.5 Assessing Convergence for the BPIRT Algorithm

Convergence of the vote level difficulty parameters can be assessed using routine convergence diagnostics included in the `R` package `superdiag` (Tsai et al., 2012). The number of dimensions retrieved from the BPIRT procedure can also be monitored using standard convergence diagnostics. However, due to the discrete nature of this value, it is often more beneficial to assess convergence using visual inspections. Assessing the convergence of the difficulty parameters, ideal points, and elements of the binary matrix proves more challenging due to the varying dimension nature of the set of estimates that make up each set of structural parameters. Convergence for these parameters is not directly assessed. Rather, convergence diagnostics are performed on the mean of the latent distribution passed to the data augmentation step - $(\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j$. Similarly, I monitor the log-likelihood of the data given the implied model at each step and use this to assess convergence of the posterior distribution of the log-likelihood of the data. Under a converged model, the log-likelihood should be 1) unimodal and approximately symmetric and 2) show behavior that appears as a random walk over iterations. Both of these conditions are generally met when allowing the procedure to have a long burn-in. There is relatively low autocorrelation between draws when the stationary distribution is reached, so the parameter space is explored relatively quickly. Similarly, the posterior distributions of interest are normally distributed due to the conjugacy of the model. Thus, the number of draws needed to truthfully represent the posterior distributions are relatively low.

A.6 Methods for Achieving Faster Convergence

As with many MCMC procedures, setting good starting values is a key aspect to achieving fast convergence to the stationary posterior distribution. Using the matrix of binary random variables, I used principal components to put together a reasonable set of starting values. For any values that were missing, I used multiple imputation to quickly fill-in the missing values. I then ran PCA on this full matrix and kept Pois(1) of the scores and loadings as the latent variables and discrimination parameters, respectively. I always started the difficulty parameters at 0 and set 50% of the elements of each column of $\boldsymbol{R}$ to 1. The variance parameters are always started at 1.

One potential problem arises at the beginning of a MCMC chain - if the starting values are particularly bad and the RNG is not giving favorable initial draws from the infinite part of the feature sampler, it is possible for the number of features at the end on an MCMC iteration to move to zero. This is problematic. In order to prevent this behavior, I chose to begin each chain of the MCMC procedure with 100 iterations where the IBP prior 1) did not look at the number of other votes which took on a feature when determining if it would take the feature (i.e. setting Equation 2.23 equal to 1 for all votes) and 2) did not use the infinite search over the feature space. This creates a period where the model simply learns the ideal points over a fixed number of dimensions. Once the ideal points begin to sort, the rest of the model runs smoothly and there is never less than 1 dimension in the analysis.

A.7 Simulation Exercises

In order to understand how BPIRT estimates the binary matrix that encodes each vote's dimensionality and, in turn, the dimensionality of the underlying policy space, I ran simulations on data sets with known parameters and examined how BPIRT uncovers the true underlying latent structure of the data.

The purpose of these simulations is two-fold. First, it is necessary to examine how accurately BPIRT recovers the binary matrix associated with the matrix of discrimination parameters from a known data generating process as a function of the number of voters, number of votes, and the true underlying dimensionality of the latent policy space. Given that the goal of BPIRT is to recover the vectors that dictate which votes correspond to which dimensions of the underlying policy space, it is important to understand when the procedure succeeds in providing an accurate representation of this data. Second, and closely related to the first goal, it is important to understand the properties of the Indian Buffet Process prior across varying numbers of votes, voters, and true underlying dimensionalities. In particular, it is important to examine the ability of the IBP prior to recover the true number of dimensions and to ensure that BPIRT does not uncover *spurious* sets of features that do not correlate with the true underlying feature set. While theory dictates that the IBP will uncover the exact solution when the number of votes and voters is large, roll call data sets are inher-

TABLE 2.3. Number of Dimensions Estimated From Simulated Data With Various Numbers of Voters, Votes, and Known Dimensionalities Using BPIRT

Votes			
Voters	100	450	900
100	1	1	1
250	1	1	1
400	1	1	1

(a) True Dimensionality = 1

Votes			
Voters	100	450	900
100	1	1	1
250	1	3	3
400	1	3	3

(b) True Dimensionality = 3

Votes			
Voters	100	450	900
100	1	1	1
250	1	4	4
400	1	4	5

(c) True Dimensionality = 5

Votes			
Voters	100	450	900
100	1	2	3
250	2	5	6
400	2	5	6

(d) True Dimensionality = 7

Note: Values reported are posterior modes. In almost every case, the posterior for number of dimensions converged to a single value.

ently limited in the number of voters and the number of votes made within a session. Thus, understanding the small and medium sample properties of this nonparametric prior is of interest.

In order to simulate data that has a similar structure to actual roll call data, I used PCA on a set of 928 votes made by 428 members of the 105^{th} session of the U.S. House to estimate a seven-dimensional covariance matrix. This covariance matrix was used to generate simulated roll call data sets with 100/250/400 voters, 100/450/900 votes, and 1/3/5/7 true underlying dimensions in the latent policy space. These simulated data sets were then passed to an implementation of BPIRT and the structural parameters were estimated. Each Markov Chain Monte Carlo routine was run with a burnin of 5000 iterations and 1000 unthinned draws were taken from the stationary posterior distribution over 2 chains. There were no indications of convergence issues in these simulations.

I first examine the relationship between number of voters, number of votes, the true dimensionality of the vote set, and the dimensionality uncovered by BPIRT using the Indian Buffet Process prior. The number of dimensions uncovered by BPIRT for each simulation set can be seen in Table 2.3. On first glance, it is easy to see that the behavior of the IBP prior to uncover the correct number of dimensions expectedly depends on the number of votes. As shown by Ghahramani and Griffiths (2006), the number of features represented by the prior increases in $\mathcal{O}(\log(P))$. This property is apparent in the simulation sets as the ability to estimate the true dimensionality is contingent on having a large number of votes. This relationship is also seen in the number of voters, though not as strongly. This property makes sense, as one would expect that more observations would lead to more accurate estimation of model parameters. However, the number of dimensions estimated appears to be capped in the number of votes.

A second important observation is that the model is **conservative** in its estimation of new dimensions when the number of votes or voters is low. In all cases, the number of estimated dimensions is lower than the truth with small roll call voting data sets resulting in a one dimensional posterior. On the other hand, when presented with a data set that is truly one-dimensional, BPIRT accurately estimates that only one dimension is needed. This finding should assuage concerns that BPIRT uncovers spurious dimensions. All in all, BPIRT provides a useful tool for estimating the dimensionality of the underlying policy space. In particular, it is well suited to examine whether or not a roll call data set requires only one dimension to explain variance within the data set.

I also examine BPIRT's ability to correctly uncover structural zeros in the binary matrix. Recall that zeros in this matrix imply that a specific vote does not rely on variance explained by a given dimension when explaining the underlying utility functions that lead to certain vote outcomes. The proportion of correctly classified

TABLE 2.4. Proportion of Elements in R Correctly Classified From Simulated Data With Various Numbers of Voters, Votes, and Known Dimensionalities Using BPIRT

Votes Voters	100	450	900
100	0.93	0.99	1.00
250	0.97	1.00	1.00
400	0.97	1.00	1.00

(a) True Dimensionality = 1

Votes Voters	100	450	900
100	0.77	0.73	0.73
250	0.80	0.81	0.84
400	0.80	0.87	0.89

(b) True Dimensionality = 3

Votes Voters	100	450	900
100	0.81	0.77	0.77
250	0.82	0.84	0.86
400	0.83	0.89	0.91

(c) True Dimensionality = 5

Votes Voters	100	450	900
100	0.84	0.82	0.84
250	0.86	0.87	0.89
400	0.87	0.91	0.93

(d) True Dimensionality = 7

Note: Values reported are posterior medians.

elements of $\mathbf{R}$ for each simulation set is shown in Table 2.4.[19] The relationship between the number of votes, number of voters, and accuracy in recovering elements of the binary matrix is similar to the one seen in Table 2.3 - more votes and more voters results in more accurate estimation of the structural parameters in $\mathbf{R}$. However, unlike in the previous case, the accuracy of estimation seems to be driven by the number of voters. This finding makes sense, however, as each estimate within $\mathbf{R}$ is related to a specific vote/dimension combination. Thus, more voters means more information about which sources of variation best describe the vote. Even in smaller samples, BPIRT provides an accurate representation of the binary matrix. This is especially apparent in estimations with a true one-dimensional model. All in all, these

[19] When constructing this data, I attempted to recreate the number and structure of dimensions that are typically seen in roll call voting. This led to seven dimensions with a decreasing number of votes which took on each dimension. Every vote took on the first dimension while other dimensions were only required for a proportion of votes. This leads to the high hit rate when the true number of dimensions is one and the decrease in hit rate between one and three.

simulations show that BPIRT can accurately recover the underlying structures which drive voting under the varying dimensions model of vote choice.

B Multidimensionality in the U.S. Congress Over Time

For analysis in this section, I examine the roll call voting data sets for the $1^{st}-115^{th}$ $(1789-2017)$ sessions of both chambers of the U.S. Congress, separately. Over the set of roll call votes in each session, I analyzed votes that had at least 5 votes in the minority and voters that registered roll call votes for at least 50% of the votes analyzed. I chose to run the BPIRT procedure on each roll call data set for a burnin of 20,000 iterations and capture 1000 draws of the parameters from the stationary posterior distribution over two independently initialized chains. Assessments of convergence both within and across chains showed no evidence of lack of convergence.

The proportion of variance explained by a dimension for a vote can be derived using the properties of the BPIRT model. Recall that the marginal probability of the augmented data under BPIRT is:

$$P(\boldsymbol{x}_i) \sim \mathcal{N}_P(\boldsymbol{\alpha}, (\boldsymbol{R} \odot \boldsymbol{\Lambda})(\boldsymbol{R} \odot \boldsymbol{\Lambda})' + \boldsymbol{\mathcal{I}}_P) \tag{2.34}$$

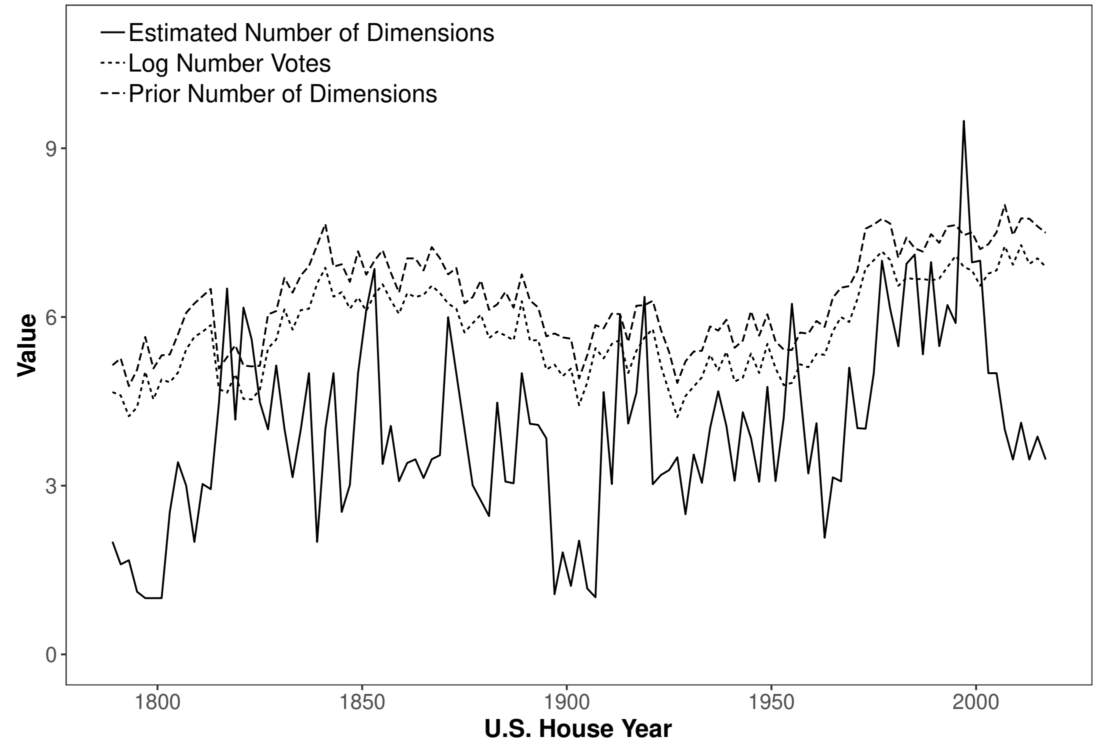
Note that under the marginal posterior, each voter has the same probability density function. This implies that the variance of the augmented data for a vote is:

$$V[\boldsymbol{x}_j] = \sum_{k=1}^{K} r_{j,k}\lambda_{j,k}^2 + 1 \tag{2.35}$$

Then, the proportion of variance explained (PVE) by a dimension on a vote can be defined as:

$$\text{PVE}_{j,k} = \frac{r_{j,k}\lambda_{j,k}^2}{\sum\limits_{h=1}^{K} r_{j,k}\lambda_{j,h}^2} \tag{2.36}$$

FIGURE 2.10. Relationship Between Number of Votes, Estimated Number of Dimensions, and Prior Number of Dimensions Implied by IBP for each Session of the U.S. House



Since each of these quantities presented in this section are, themselves, uncertain measures, the posterior means are also associated with a 95% highest posterior density interval. In practice, these ranges are very tightly bunched around the posterior means. As these quantities do not plot well, I have chosen not to include the error bars in many of the figures. These quantities can be found in the replication materials. The inclusion of 95% HPD intervals do not change the overall conclusions made from any of the graphs included in this section.

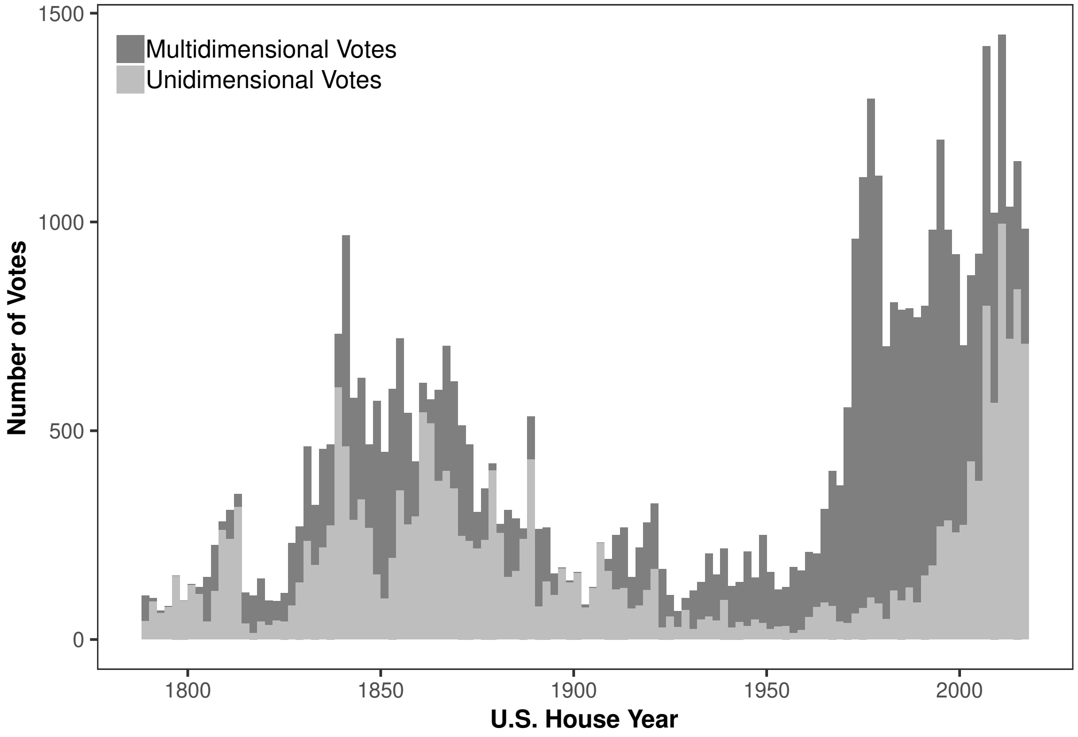
B.1 IBP and Dimensions Observed in the U.S. House

For many sessions of the U.S. House and U.S. Senate, the posterior distribution for the number of dimensions converged strongly on a single value. These sessions are indicated in Figure 2.1 by points with no error bars. Exploring the infinite set of features using the beta process prior is contingent on tuning parameters, which are outlined in the estimation procedure presented in the Appendix. Similarly, the discrete nature of the posterior distribution for number of dimensions can lead to results that appear to not have reached stationarity. Given the performance of BPIRT in simulation exercises, there is strong evidence that these values are equal to or below the truth and assuage any concerns related to not fully exploring the posterior. Other choices for these hyperparameters lead to results that share the same mode but have higher values included in the 95% HPDs.

The relationship between the number of voters, number of votes, and estimated dimensionality is well established in the simulation section. However, it is important to see how these relationships manifest in the actual roll call data used in these applied examples. Similarly, one might question whether or not these results are unduly driven by the choice of priors for the hyperparameters of the IBP prior. To address these concerns, I checked the relationship between the number of votes, the expected number of dimensions drawn from the IBP prior, and the estimated dimensionality.

Figure 2.10 shows the logarithm of the number of votes analyzed, the posterior mean number of dimensions estimated, and the number of dimensions implied by the IBP prior using posterior means values for the IBP hyperparameters for each session of the U.S. House. First, it is easy to see that there is s strong dependence in the number of votes and the prior number of dimensions implied by BPIRT - the prior number of dimensions scales almost perfectly with the logarithm of the number of votes with an additive constant implied by the IBP hyperparameters. On the other hand, there is not a strong correlation between the posterior mean number of dimensions estimated by BPIRT and the prior expected number of dimensions. This implies that the choice of prior is not unduly influencing the number of dimensions estimated by BPIRT. However, the choice of prior does seem to place a cap on the number of dimensions

FIGURE 2.11. Number of Unidimensional and Multidimensional Votes Analyzed in Each Session of the U.S. House

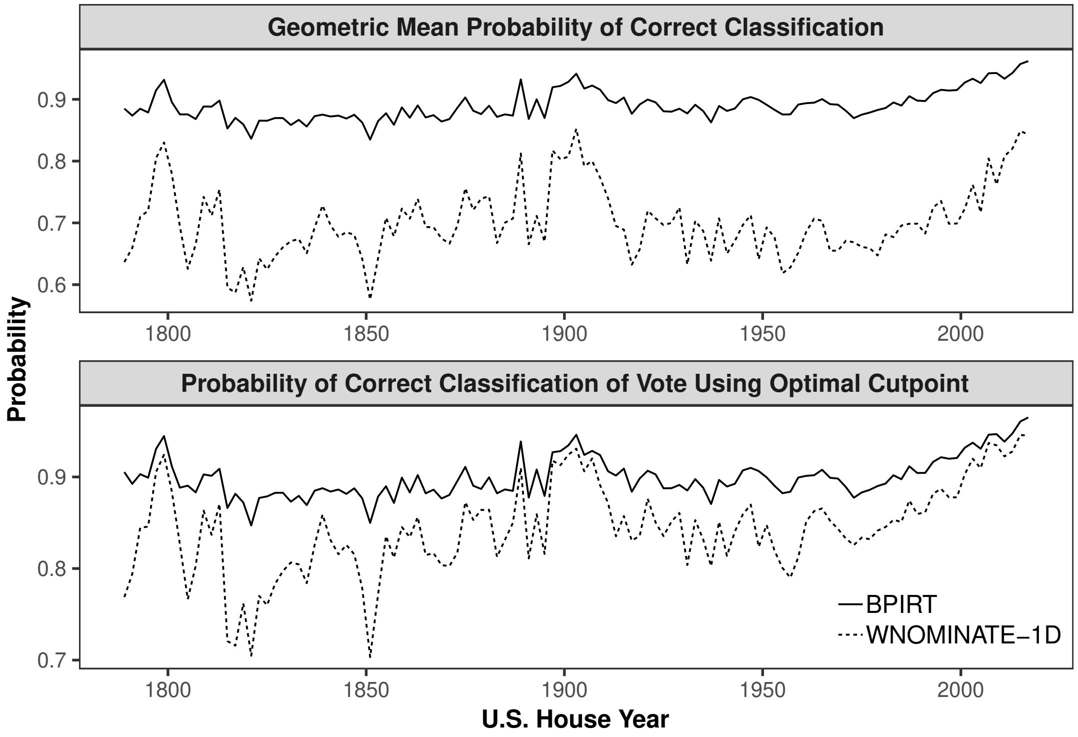


which can be estimated; the number of dimensions estimated rarely goes above the number of dimensions implied by the prior. This behavior is expected due to the properties of the IBP prior discussed in the simulations.

B.2 More Summaries of Multidimensional Voting in the U.S. House

Figure 2.11 shows the number of votes that were analyzed for each session of the U.S. House. These votes are then classified as unidimensional or multidimensional by the posterior mean probability that a vote requires more than the first dimension to best explain individual vote variation. This figure demonstrates two key concepts. First, the sheer number of votes that occur within each session dramatically increase after 1950. This allows BPIRT to better estimate the number of dimensions needed to model the roll call data set and, in turn, allows for more multidimensionality in votes to appear. Second, this plot makes it clear that the proportion of votes that

FIGURE 2.12. Aggregate GMP and Proportion Correctly Classified Votes for the 1st - 115th Sessions of the U.S. House



were classified as multidimensional by BPIRT also dramatically increases after 1950. While there is a downturn in this proportion in recent times, these proportions are generally on par with the proportion of votes that were classified as unidimensional in the mid-1800s. This is another way of showing that the unidimensional, party bloc voting of recent times is not a unique occurrence.

GMP is only one way in which the quality of ideal point models is assessed. The most common way in which models are compared is through correct classifications metrics. Using the ideal points and other structural parameters, the proportion of votes that are correctly classified can be used to demonstrate the ability of the model to partition “Yea” and “Nay” votes appropriately in different situations. This metric is a natural fit for predictive models like NOMINATE, but does not necessarily account for the uncertainty associated with each of the model parameter estimates (Carroll et al., 2009). The proportion of votes correctly classified using optimal cutpoints for BPIRT was estimated using the mean of the augmented posterior - $\mu_{i,j} = (r_j \odot \lambda_j)\omega_i - \alpha_j$. If $\mu_{i,j} < 0$, then the vote was a predicted "Nay" and "Yea" otherwise. A similar calculation is used for NOMINATE.

Figure 2.12 shows the aggregate GMP and proportion correct classification under the optimal cutpoint for both BPIRT and one-dimensional WNOMINATE model. Beginning with correct classification, it is easy to see that BPIRT and a unidimensional WNOMINATE model yield similar results throughout much of U.S. history, especially in recent sessions. While this could certainly be taken as evidence that the unidimensional model is sufficient, correct classification done in this sense is theoretically deficient and ignores the inherent uncertainty associated with ideal point measures while encouraging heavily overfit models (Aldrich et al., 2014; Roberts et al., 2016). Ideal points are rarely used to predict new votes; in fact, ideal points are almost always used to explain the voting behavior given the entirety of votes for an analyzed period. It is imperative that ideal point fits are treated with the same probabilistic rigor that any other inferential technique requires when attempting to explain behavior and, thus, important to consider more statistically rigorous approaches when making choices about the underlying parameters of ideal point models.

Proportion reduction in error metrics are similar to correct classification metrics (Carroll et al., 2009; Aldrich et al., 2014; Roberts et al., 2016). These approaches are central to previous discussions of how many dimensions are needed to model a roll call voting set. While these approaches provide some information of this topic, they are post-hoc statistics that require *a priori* assumptions about the structure of the underlying latent space. BPIRT estimates the dimensionality and necessity of dimensions at a vote-level within its statistical procedure and, therefore, is incompatible with the notion of adding or subtracting whole dimensions from the latent space. For

TABLE 2.5. Correlation between Dimensions Estimated for the 107th U.S. House

	Party	Procedural	Security	Budget	Rural	Foreign	DOD
Party	1.00	0.30	0.28	0.33	-0.07	-0.09	0.17
Procedural	0.30	1.00	0.35	0.05	-0.03	-0.07	0.25
Security	0.28	0.35	1.00	0.10	0.09	-0.04	0.16
Budget	0.33	0.05	0.10	1.00	0.13	-0.01	0.05
Rural	-0.07	-0.03	0.09	0.13	1.00	0.04	-0.06
Foreign	-0.09	-0.07	-0.04	-0.01	0.04	1.00	0.01
DOD	0.17	0.25	0.16	0.05	-0.06	0.01	1.00

Note: Values reported are posterior means for the Pearson correlation coefficient between estimated ideal points.

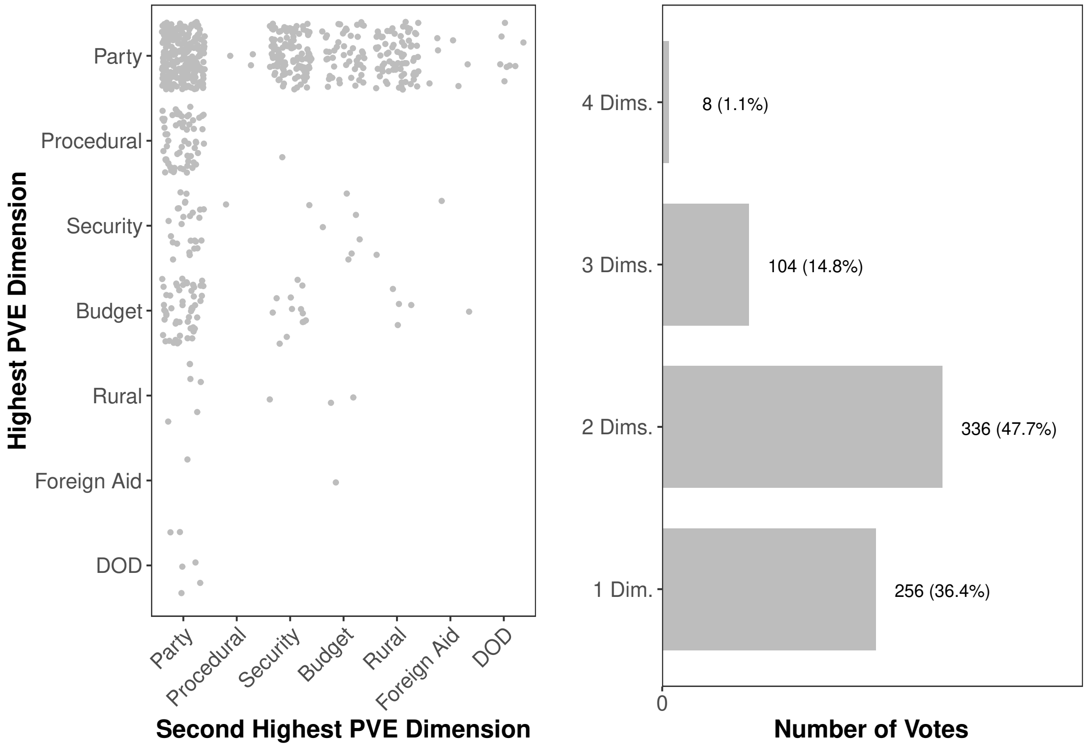
this reason, I choose to avoid inappropriate attempts to compare APRE and MPRE achieved from BPIRT.

B.3 More Summaries for the 107th U.S. House

I used a suite of tools to provide names for each of the dimensions estimated by BPIRT for the 107^{th} session of the U.S. House. First, I used vote classifications from `voteview.com` to analyze the set of votes which had non-zero contributions to a dimension and saw general trends in the content of votes that loaded in each dimension (Poole and Rosenthal, 2012). Second, I used non-negative matrix factorization and regularized logistic regression to extract important words from the bill summaries associated with each vote on each set of votes. These tools created a general picture of what each dimension was modeling. While still somewhat ad-hoc, this approach defines a general method that will be useful for future research attempting to find trends in what each dimension means, over time.

BPIRT tends to extract dimensions which run relatively orthogonal to one another. One of the advantages of the IBP prior is that it generally prevents against dimensions that are highly correlated with one another from arising during estimation. This can be attributed to the IBP draws, which check to see if there is any new information

FIGURE 2.13. Highest PVE Dimensions and Number of Dimensions For Each Vote in the 107th U.S. House



Note: Values reported are estimates of the binary matrix from the iteration of the MCMC procedure with the highest complete-data likelihood. Points on the main diagonal of the PVE graph required only one dimension to best model variation within the vote.

added by a new dimension while conditioning on previously existing dimensions. Table 2.5 shows the pairwise correlations between dimensions estimated for the 107^{th} session of the U.S. House. Correlations for this session range from .01 to .35. While this indicates that there are dimensions which are not orthogonal to one another, the correlation is still relatively low.

One advantage that comes from using an IBP prior is that votes are allowed to take on a collection of dimensions rather than one, and only one, dimension. Given the complexity of the topics that are being considered when votes are cast, this is a desirable property. BPIRT models this complexity and presents the set of dimensions which each vote requires to best model its variance in voting. To this end, it is interesting to examine the number of dimensions, and which dimensions, are chosen for votes within the analyzed set. Figure 2.13 illustrates this dynamic by showing the dimensions with the highest and second highest PVE for each vote analyzed. While around 35% of votes only require the party loyalty dimension to model the roll call votes, the other 65% of votes require at least one other dimension. However, all votes require a non-zero contribution from the party loyalty dimension. This dynamic is partially due to the "rich get richer" property of the IBP prior, the fact that most votes require the party loyalty dimension as the highest or second highest PVE dimension fits well with theories of party control in U.S. legislative voting.

C U.S. Legislative Voting and Multidimensionality

C.1 A Theory of Pivotal Voters

I constructed a data set of all cloture votes that took place between the 89^{th} and 115^{th} sessions of the U.S. Senate. I began by collecting all cloture votes recorded in the U.S. Senate records, which can be found at `https://www.senate.gov/legislative/cloture/clotureCounts.htm`. Using this set of votes, I attempted to create matches to roll call votes retrieved from `https://www.voteview.com` (Poole and Rosenthal, 2012). For each cloture vote, I attempted to match the roll call records via the session number and the rollcall number. However, this proved challenging in cases where the clerk roll number was not recorded. Thus, I pulled in supplementary information about each vote from the Propublica API to attempt to join cloture votes to roll call records using date and bill name (Leeper, 2015). This approach was used to give each cloture vote a date, roll call number, and relevant bill number.

Once each record was joined to a roll call vote, I created a series of cloture vote episodes for each set of motions to invoke cloture that involved the same bill in the same session of the U.S. Senate. I defined a cloture vote episode as any group of cloture votes that occurred within the same session for the same bill number that had at least 2 members. This led to 777 cloture votes and 257 cloture vote episodes. In each episode, I analyzed pairs of votes in sequence, meaning that I compared the votes from the first and second votes, second and third votes, etc. This led to 517 direct comparisons.

For each comparison, I defined a vote switch as a pair of votes from a U.S. Senator that were different (i.e. "yea" to "nay" or "nay" to "yea" on whether to invoke cloture). Another key component of this design is placing U.S. Senators in ideal point quartiles. Using the set of ideal points for a specific session (which are defined at the session level), I divided the ideal points into quartiles and assigned each member a quartile based on this measure. For this division, I used the ideal points retrieved from the iteration of the stationary posterior that had the highest complete data likelihood.[20] Corresponding switches to the ideal point quartiles were made when the vote was classified as a "President-Side Veto".

More recent sessions of the U.S. Senate introduced a new problem that was seemingly not present in the data set from Krehbiel (1998). In some votes pairs, there appeared to be an unusually large number of switches that occurred across both parties. Upon further examination, these cases corresponded to complicated cloture votes that were associated with various amendments to bills proposed by both parties. Given that these votes were occurring on the same bill, but had completely different underlying contexts, I removed these 40 vote pairs from the data set. I identified votes pairs to remove by finding all votes in which 50% of the votes from each party switched.

[20] I explored the idea that the results may change if the uncertainty of the ideal points was included in further empirical models using these measures. To do this, I ran the models in proceeding parts of this section using the full first-dimension ideal point posteriors and taking a random draw for each observation that is then used to create the quartiles. This bootstrap design essentially incorporates the uncertainty of the ideal points into the estimation procedure. The coefficients retrieved from the logistic regression models then represent a distribution of the full-error model. For a 1000-draw bootstrap design, I found no discernable difference between the full-error and single point estimate models.

Analysis of the full data set including these 40 vote pairs could not directly replicate the findings of Krehbiel (1998). Upon removal, I was left with 477 cloture vote pairs. Table 2.6 shows some summary statistics for this data.

The logistic regression models in this section were estimated using `MCMClogit` from `MCMCpack` (Martin et al., 2011). Priors on the coefficients were conjugate normal with zero mean and variance of 1000. Each model was estimated twice using diffuse starting values and convergence was checked using the suite of tools provided by the `superdiag` package (Tsai et al., 2012). After a generous burn-in of 100,000 iterations, there was no evidence of a lack of convergence to the stationary distributions. One consideration when estimating the models for cloture vote switching is choosing the threshold for multidimensionality in the threshold model. While a small value for this threshold makes sense, there is empirical justification for the decision. The threshold of .001 was chosen after using a grid search of all values between .001 and .999 as the cutpoint and estimating the logistic regression for vote switching for all values. Using .001 as the cutpoint yielded the model with the highest log-marginal likelihood. Code for the models in this section can be found in the replication materials.

C.2 A Theory on Party Control

Data on the set of all final passage votes in the U.S. House was retrieved from the political institutions and public choice roll-call database, which includes each roll call vote that occurred in the U.S. Congress and classifications relating to they type of vote that occurred (Crespin and Rohde, 2012). To match the analysis from Cox and McCubbins (2005), I only analyzed final passage votes which required a simple majority for passage. Table 2.2 shows some summary statistics for this data.

One modeling consideration which arose was what to do with final passage votes that were nearly unanimous. When estimating ideal points, I chose not to estimate vote-level dimensionalities for votes that had less than a 5 vote split in order to prevent singularity issues when inverting matrices. In order to best replicate the methods of Cox and McCubbins (2005), I chose to code these votes as unidimensional. Theoretically, this is justified - on a unanimous or nearly unanimous vote, the vote can be modeled using a single dimension with a cutpoint that exists outside of the range implied by the ideal points. This was done to attempt to faithfully recreate the data set and estimation strategy from Cox and McCubbins (2005).

Models for this section were estimated using `JAGS` (Plummer, 2003). `runjags` was used to run the models within `R` (Denwood, 2016). Diffuse normal priors were placed on the coefficients. Special care needed to be taken to appropriately model the intercept in these logistic regressions. Since the underlying model predicts that the probability of a roll is zero when the distance between the floor median and party median is zero, it is expected that the intercept will be highly negative and only partially locally identified. For this reason, the mean of the prior for the intercept is set to be -10 while other means are set to 0. This helped model convergence greatly.

Spike-and-slab priors on the regression coefficients were computed using a beta-Bernoulli specification. Specifics can be seen in the replication materials. I chose to use this prior specification given the nature of the test presented by Cox and McCubbins (2005) - under party cartel theory, the distance between the majority party median and the floor median should have zero effect on the probability of a majority party roll. While this specification is not a test of zero effect, directly, it forces the model to be fit under conditions where the value of a coefficient that is not statistically distinguishable from zero is forced to be represented as zero. This allows for easier interpretation of the expected null result.

As with the previous set of models, the value of .001 is both theoretically and empirically justified. For each of the models run in this section, I first ran a version of the threshold model which places a uniform prior, bounded between 0 and 1, on the cutline for the threshold model. Under the convergence guarantees of MCMC methods, this value should converge to a stationary distribution which dictates values of the cutline that maximize the posterior likelihood conditional on the priors. In each case, the posterior distribution had a single mode close to a small number around zero. This provides empirical justification for this choice. All models and `JAGS` code can be found in the replication materials for this paper.

TABLE 2.6. Summary statistics for cloture vote switches in the 89^{th} - 115^{th} Sessions of the U.S. Senate

	Number of	Number of	Proportion Switches by:			
Congress	Vote Episodes	Switches	Adj.	Adj. Mods	f-Quart.	Non-Adj.
89	3	7	0.14	0.29	0.14	0.43
90	3	12	0.50	0.25	0.00	0.25
91	1	2	0.00	0.00	0.50	0.50
92	9	49	0.20	0.16	0.33	0.31
93	13	59	0.15	0.22	0.37	0.25
94	12	46	0.09	0.15	0.20	0.57
95	7	16	0.06	0.38	0.38	0.19
96	8	68	0.25	0.26	0.44	0.04
97	14	65	0.17	0.38	0.32	0.12
98	5	28	0.46	0.18	0.18	0.18
99	8	43	0.37	0.23	0.40	0.00
100	20	60	0.08	0.42	0.25	0.25
101	6	42	0.24	0.24	0.36	0.17
102	17	129	0.06	0.41	0.20	0.33
103	19	40	0.23	0.10	0.65	0.03
104	24	52	0.25	0.21	0.48	0.06
105	22	83	0.27	0.20	0.39	0.14
106	18	126	0.03	0.44	0.06	0.47
107	27	237	0.22	0.30	0.19	0.29
108	17	216	0.29	0.20	0.25	0.26
109	11	89	0.30	0.19	0.29	0.21
110	44	508	0.15	0.32	0.25	0.28
111	32	154	0.29	0.18	0.45	0.08
112	20	327	0.41	0.05	0.47	0.07
113	28	293	0.40	0.04	0.54	0.02
114	68	501	0.17	0.25	0.22	0.36
115	13	111	0.30	0.12	0.43	0.15

TABLE 2.7. Summary statistics for final passage votes in the 83^{rd} - 115^{th} Sessions of the U.S. House

		Number of:		Distance between Floor Median and:	
Congress	Votes	Maj. Rolls	Min. Rolls	Maj. Median	Min. Median
83	48	0	9	0.21	0.33
84	46	3	7	0.31	0.26
85	58	5	8	0.31	0.26
86	72	2	24	0.18	0.39
87	95	1	30	0.21	0.33
88	92	3	36	0.25	0.33
89	169	0	45	0.11	0.46
90	214	2	27	0.27	0.31
91	168	8	19	0.28	0.31
92	173	8	15	0.26	0.33
93	245	5	34	0.26	0.34
94	255	8	62	0.12	0.47
95	209	4	56	0.13	0.46
96	169	1	52	0.15	0.45
97	126	4	30	0.26	0.36
98	121	10	55	0.16	0.49
99	116	6	48	0.20	0.47
100	116	1	47	0.19	0.48
101	117	2	42	0.19	0.48
102	117	8	50	0.17	0.50
103	107	3	63	0.17	0.54
104	130	3	72	0.21	0.56
105	124	8	62	0.23	0.54
106	166	10	61	0.26	0.52
107	111	4	41	0.24	0.54
108	129	1	48	0.19	0.59
109	119	5	59	0.18	0.62
110	149	5	98	0.23	0.58
111	104	1	81	0.16	0.62
112	126	1	104	0.20	0.67
113	139	3	112	0.23	0.65
114	159	2	136	0.20	0.69
115	140	0	110	0.23	0.67

CHAPTER III

A Bayesian Nonparametric Approach to Estimating Group Dynamics in Roll Call Scaling

3.1 Introduction

Studies of legislative behavior focus upon the relationship between legislative preferences, institutional structure, and legislative outcomes. Spatial models are a frequently used tool for studying these relationships. In a spatial model of voting bodies, policies are represented geometrically and votes occur as a function of individual legislators' *ideal points*. An ideal point represents a legislator's most preferred policy outcome and competing policies are judged based upon their distances from her most preferred policy. Under the assumption of utility-maximizing, rational legislators, the spatial model provides a consistent method for researchers to understand how ideal points and policy lead to specific legislative outcomes.

A common task in the legislative behavior literature is to estimate the set of ideal points for matrix of *roll call* data. In this data, the votes for each legislator on a variety of different proposals are recorded. Then, a scaling procedure is used to determine the ideal points for each legislator (Poole and Rosenthal, 1997; Clinton et al., 2004). Scaling procedures typically seek to represent each policy votes on in the roll call set in a low-dimensional Euclidian space. In turn, this allows estimation of ideal points in the same Euclidian space. Thus, the scaling procedure admits a consistent space in which all votes within the roll call set can be represented. Scaling roll call votes in this way implies that there exists a single policy space in which represents all roll call votes within the analyzed roll call set.

The policy space uncovered by scaling methods encompasses the various complexities of the legislator voting behaviors. While the ideal points, themselves, are generally of interest, the uncovered policy space is also substantively interesting. For example, McCarty et al. (2016) utilize ideal points estimated using NOMINATE methodology (Poole and Rosenthal, 1984) and the corresponding policy space to show increased polarization in elite voting over time. This result (and many others like it) relies on the assumption that meaningful parts of the policy space exist only in one dimension. This low-dimensionality conjecture is a key part of numerous theories relating to changes in Congress over time and is key to many other theories which utilize ideal point estimates.

McCarty et al. (2016) argue that there are between one and two dimensions in most session of Congress. The first dimension projects legislators' votes to a "liberal-conservative" dimension which corresponds mostly to economic issues. The second dimension, if needed, corresponds to social issues of the time, typically questions related to race. Over time, NOMINATE shows that the need for a second dimension has disappeared and most roll call voting behavior can be described by the liberal-conservative dimension. The single dimension argument has been the basis for many formal models and empirical findings about Congress (Aldrich and Battista, 2002; Bafumi and Herron, 2010; Binder, 1999; Cameron, 2000; Cox and McCubbins, 2005; Jessee, 2009, 2010; Krehbiel, 1992). However, many of these results are incredibly sensitive to changes in this assumption; if the dimensionality of the congressional vote choice model is any value greater than one, then median voter theorem no longer holds and the results no longer hold (Kramer, 1973). Thus, strong evidence for the one-dimensional model should be in place.

Along with concerns about the dimensionality of the policy space, there has been recent research that examines the notion of independence that is required for roll call scaling models to make estimates. In particular, there are questions about the role of groups within the data. A base assumption that must be made for procedures like NOMINATE is that the errors associated with each individual vote are independent and identically distributed conditional on the ideal point. While roll call scaling techniques are attempting to decompose dependence among the votes to find ideal points, dependence among errors is still a potential problem. This problem is discussed in depth by Spirling and Quinn (2010), showing that there is cause to be concerned about correlated errors in parliamentary voting in the U.K. This problem is mostly due to party incentives and can cause dependence among the votes that is not accounted for in the standard spatial model. Spirling and Quinn (2010) utilize an infinite mixture model approach which models the policy space as a number of latent clusters, but forgo the standard ideal point interpretation which is often used in studies of U.S. legislatures.

Continuous measurements of ideal points are key to numerous studies of the U.S. Congress, particularly those related to party control and party effects in the legislative chambers. Following Krehbiel (1992) and Krehbiel (2010), understanding the effects of party membership and the role that parties play in voting became an important endeavor in legislative studies. Through careful qualitative accounts and empirical studies, theories of party control emerged and posited that members of Congress have incentives to vote with their party and the groups assess situations where strong control over members is needed (Aldrich and Rohde, 2000; Aldrich, 1995; Cox and McCubbins, 2007). These studies and many that followed used NOMINATE scores to empirically test these points. NOMINATE scores were also used to show that parties do not have much influence over the voters of their members (Cox and Poole, 2002). These opposite findings raise causes for concern when using NOMINATE scores to test theories of party effects.

The main problem that arises when using NOMINATE scores and other roll call scaling techniques to test theories of party control is that, *a priori*, groups are assumed to be orthogonal noise. These stimation procedure require that errors are assumed to be independent and identically distributed across both individuals and bills, conditional on the uncovered latent variables. Given that there appears to be grouping that arises according to party membership in these latent scores, the possibility that group effects exist should, at least, be considered a possibility when estimating the model. Thus, flexible models that allow for less restrictive assumptions are needed in order to better test theories of voting in Congress.

Using the clustering approach proposed by Spirling and Quinn (2010) as a starting point, I propose a new model which allows for both individual ideal points and group ideal points and membership to be measured. This model, Clustered Beta Process IRT (C-BPIRT) utilizes a combination of beta process priors on the number of dimensions for the uncovered ideal point space (Knowles and Ghahramani, 2011) and Dirichlet process priors (Ferguson, 1973) on group membership within the set of roll call votes. This model flexibly estimates continuous level ideal points for each individual voter with unknown dimensionality while also uncovering a set of discrete groupings within the data and estimating the group effect on voting. This model uses Bayesian nonparametric approaches to let the data dictate which groups and dimensions are needed to best model the voting behavior of members of Congress. By not constraining the model to fit a rigid set of assumptions, the data dictates which potential mechanisms of vote behavior are important and estimates the corresponding measures.

3.2 A Model for Roll Call Analysis

For a legislature, assume there are N voting members that cast P votes over the course of time analyzed. For any given vote $j \in (1, P)$, legislator $i \in (1, N)$ must choose between two alternatives: cast a "Yea" vote for the proposed motion ($\boldsymbol{\vartheta}_j \in \mathbb{R}^K$) or cast a "Nay" vote for the proposed motion ($\boldsymbol{\varphi}_j \in \mathbb{R}^K$). Behavior in this legislature is assumed to be describable in a K-dimensional policy space - all votes that are made by legislator i can be described by the K-dimensional point locations of $\boldsymbol{\vartheta}_j$ and $\boldsymbol{\varphi}_j$ within the space and a K-dimensional ideal point, $\boldsymbol{\omega}_i$, which encapsulates the policy preferences of legislator i.

A legislator must choose whether to vote for $\boldsymbol{\vartheta}_j$ or $\boldsymbol{\varphi}_j$. Using a utility maximization model that assumes quadratic loss in distance from her ideal point, assume that she chooses the alternative which grants the highest utility:

$$\begin{aligned} U_i(\boldsymbol{\vartheta}_j) &= -\|\boldsymbol{\omega}_i - \boldsymbol{\vartheta}_j\|^2 + \eta_{ij} \\ U_i(\boldsymbol{\varphi}_j) &= -\|\boldsymbol{\omega}_i - \boldsymbol{\varphi}_j\|^2 + \nu_{ij} \end{aligned} \tag{3.1}$$

where η_{ij} and ν_{ij} are stochastic elements of the utility functions. This model is completely specified if a known structure is placed on η_{ij} and $\nu_{i,j}$ (Heckman and Snyder Jr, 1996; Poole and Rosenthal, 1997; Clinton et al., 2004).

Let $\boldsymbol{Y}$ be a matrix of roll call votes and $y_{i,j}$ be the vote choice that legislator i makes on proposal $j : y_{ij} = 1$ if legislator i votes "Yea" on vote j and $y_{ij} = 0$ if she casts a "Nay" vote. Given the model construction, the probability that legislator i votes for $\boldsymbol{\vartheta}_j$ can represented as:

$$P(y_{ij} = 1) = F(\boldsymbol{\lambda}_j' \boldsymbol{\omega}_i - \alpha_j) \tag{3.2}$$

where $F(\cdot)$ is the CDF associated with the chosen error structure, $\alpha_j = \frac{\boldsymbol{\vartheta}_j' \boldsymbol{\vartheta}_j - \boldsymbol{\varphi}_j' \boldsymbol{\varphi}_j}{\sigma_j^2}$, and $\boldsymbol{\lambda}_j = \frac{2(\boldsymbol{\vartheta}_j - \boldsymbol{\varphi}_j)}{\sigma_j^2}$.

This construction admits a corresponding statistical model that allows for estimation of the *structural parameters* $\boldsymbol{\alpha}$ and $\boldsymbol{\Lambda}$ and the ideal points, $\boldsymbol{\Omega}$. Assuming the errors are independent and identically distributed, a likelihood function can be derived:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Omega}|\boldsymbol{Y}) = \prod_{i=1}^{N}\prod_{j=1}^{P} F(\boldsymbol{\lambda}_j'\boldsymbol{\omega}_i - \alpha_j)^{y_{ij}} \times \left(1 - F(\boldsymbol{\lambda}_j'\boldsymbol{\omega}_i - \alpha_j)\right)^{1-y_{ij}} \tag{3.3}$$

Bayesian implementations of this model place priors on all of the structural parameters and estimation proceeds using Markov Chain Monte Carlo methods (Clinton et al., 2004). With minor changes, this model is equivalent to the NOMINATE procedure (Poole and Rosenthal, 1997).

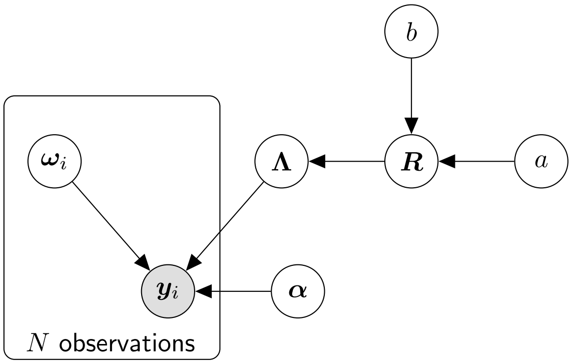
McAlister (2018) derives a similar model that allows for appropriate estimation of the dimensionality of the resulting ideal point space. Unlike previous approaches that require post-hoc Scree tests of the number of dimensions (Cattell, 1966), often erroneously leading to conclusions that one dimension is appropriate for describing the policy preferences of members of the U.S. Congress, a Bayesian nonparametric approach is used that places a prior on the number of dimensions in the model. Using the Indian Buffet Process prior for the number of dimensions needed to represent the ideal point space, a slightly altered formulation is proposed:

$$P(y_{i,j} = 1) = F((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'\boldsymbol{\omega}_i - \boldsymbol{\alpha}_j) \tag{3.4}$$

where r_j is a binary vector that takes a value of zero if bill j does not influence a dimension and a one otherwise. This specification induces a spike-and-slab prior on the bill loadings and the infinite nature of the priors allows for estimation of a sparse set of dimensions. Figure 3.1 shows a graphical representation of the beta process IRT model.

The main workhorse of the BPIRT model is the *Indian Buffet Process*, which places

FIGURE 3.1. Graphical Representation of Beta Process IRT Model



a prior on the number of dimensions needed to best explain the latent ideal point space (Ghahramani and Griffiths, 2006; Paisley and Carin, 2009). The Indian Buffet Process is a version of a stochastic Beta process that places a prior on binary matrices. This approach allows for each bill-dimension pair to be explicitly tested against the hypothesis of no effect; a test that tells whether the collection of votes on bill j provides any information about the location of the ideal point on dimension k. The dimensionality of the set of ideal points is a substantively meaningful variable (McCarty et al., 2016) and proper estimation is a meaningful task (Aldrich et al., 2014). For these reasons, this specification of the ideal point model will serve as a starting point for the model derived in this paper.

3.3 A Model for Roll Call Analysis with Group Influences

Under a model with group influence, legislator i must choose between $\boldsymbol{\vartheta}_j$ and $\boldsymbol{\varphi}_j$, as before. She uses a utility maximization model that assumes quadratic loss in distance from her ideal point to choose the best alternative. However, unlike the previous model, assume that she belongs to a group, $g \in (1, ...G)$ that influences her vote choice.[1] Members of g share common policy goals, policy preferences, and

[1] Throughout this paper, let g_i refer to the group with which voter i is associated. Each voter is associated with one, an only one, group. This is a strict assumption, but one that generally holds in U.S. legislatures (Aldrich, 1995).

leadership that rewards loyalty to the group as well as individual policy and career goals. As such, g can be viewed as a set of legislative actors that have multiple influences driving vote choice. Like individual actors, it is assumed that groups have an ideal point, $\boldsymbol{\tau}_g$, that encapsulates its policy preferences in a K-dimensional policy space. Each actor also has an individual ideal point, $\boldsymbol{\xi}_i$. Some convex combination of these two ideal points makes up the observed ideal point, $\boldsymbol{\omega}_i = \beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i$, where β_i is a mixing component between zero and one that dictates how heavily i's ideal point leverages the group ideal point. Formally, this implies that the new utility functions are:

$$\begin{aligned} U_i(\boldsymbol{\vartheta}_j) &= -\|\beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i - \boldsymbol{\vartheta}_j\|^2 + \eta_{ij} \\ U_i(\boldsymbol{\varphi}_j) &= -\|\beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i - \boldsymbol{\varphi}_j\|^2 + \nu_{ij} \end{aligned} \tag{3.5}$$

where $\eta_{i,j}$ and $\nu_{i,j}$ are stochastic elements of the utility functions.

As before, these equations can be rearranged to give the standard IRT formulation:

$$P(y_{i,j} = 1) = \boldsymbol{\Phi}^{-1}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'(\beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i) - \boldsymbol{\alpha}_j) \tag{3.6}$$

where $\boldsymbol{\Phi}^{-1}$ is the inverse normal CDF.[2] This formulation can also be expressed in a more familiar regression-like form:[3]

$$\begin{aligned} P(y_{i,j} = 1) &= \int_0^\infty \mathcal{N}(x_{i,j} \; ; \; (\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'(\beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i) - \boldsymbol{\alpha}_j, 1) dx_{i,j} \\ &= P((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'(\beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i) - \boldsymbol{\alpha}_j + \epsilon_{i,j} > 0) \; ; \; \epsilon_{i,j} \sim \mathcal{N}(0,1) \end{aligned} \tag{3.7}$$

where the loadings, $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}$, can be seen as the regression coefficients and the K-dimensional ideal points, $\boldsymbol{\omega}_i = \beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i$, as the observed data.

[2] Throughout this paper, probit link functions are used to link the binary outcomes to the data generating function. The formal model and following empirical procedures can be altered to use the logistic link function. This leads to a generally similar formal model and empirical procedure. For more information, see Carroll et al. (2009); Goplerud (2019).

[3] $\mathcal{N}(x \; ; \; \mu, \sigma^2)$ is the normal density of x given mean μ and variance σ^2.

Assuming $\epsilon_{i,j}$ is independent across all observations and items, a joint likelihood over the data can be specified:

$$\begin{aligned}\mathcal{L}(\boldsymbol{Y}) = \prod_{i=1}^{N}\prod_{j=1}^{P} & P((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'(\beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i) - \boldsymbol{\alpha}_j + \epsilon_{i,j} > 0)^{y_{i,j}} \\ & + P((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'(\beta_i \boldsymbol{\tau}_{g_i} + (1-\beta_i)\boldsymbol{\xi}_i) - \boldsymbol{\alpha}_j + \epsilon_{i,j} < 0)^{1-y_{i,j}}\end{aligned} \tag{3.8}$$

The standard roll call scaling model can be seen as a special case of this model; specifically, assuming that $\boldsymbol{\omega_i} = \boldsymbol{\xi}_i$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{Y}) = \prod_{i=1}^{N}\prod_{j=1}^{P} & P((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'\boldsymbol{\xi}_i - \boldsymbol{\alpha}_j + \epsilon_{i,j} > 0)^{y_{i,j}} \\ & + P((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'\boldsymbol{\xi}_i - \boldsymbol{\alpha}_j + \epsilon_{i,j} < 0)^{1-y_{i,j}}\end{aligned} \tag{3.9}$$

meaning that each ideal point is only made up of an individual component.

This points to a key question - what happens if only individual components are estimated when a group component to the ideal point truly exists? The answer depends wholly on the relationship of $\boldsymbol{\tau_{g_i}}$ and $\boldsymbol{\xi}_i$. Omitted variable bias is a well-known phenomenon in regression analysis and establishes that a variable that is left out of the regression specification that is correlated with one of the included regressors leads to biased estimates of the coefficients. This identity is a key result that has led to an entire body of work on detecting and correcting for missing correlates in regression analyses.[4]

In the roll call scaling setting, similar results are expected. Significant work has shown that parties and other legislative groups influence roll call vote outcomes (Krehbiel, 1992; Krehbiel et al., 2005; Cox and Poole, 2002; Aldrich, 1995; Aldrich et al., 2014). Under the grouped model of roll call scaling, this work expresses doubt that $\boldsymbol{\tau}_{g_i}$ is

[4] See Ferrari (2020) for a very thorough discussion of this problem and solutions in the regression context.

independent of $\boldsymbol{\xi}_i$. This can manifest in the estimates of the structural parameters in several ways. Formally, assume that there is a meaningful effect from a common group, but $\beta_i = 0 \; \forall \; i \in (1, ..., N)$. For any legislator where $\beta_i \neq 0$:

$$P(y_{i,j} = 1) = P((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'\boldsymbol{\xi}_i) - \boldsymbol{\alpha}_j + \epsilon_{i,j} > 0) \; ; \; \epsilon_{i,j} \sim \mathcal{N}(0,1) - (\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)'\beta_i \boldsymbol{\tau}_{g_i} \quad (3.10)$$

If $\epsilon_{i,j}$ is treated as standard normal idiosyncratic noise, the residual left by $\boldsymbol{\tau}_{g_i}$ must go somewhere.

A first problem with this misspecification is that errors can be pushed to the ideal point estimates, themselves. Since the model assumes that ideal points are conditionally independent given all other parameters, these estimates will take on some of the residuals. In the theoretically most likely scenario that $\boldsymbol{\tau}_{g_i}$ and $\boldsymbol{\xi}_i$ are positively correlated, uncertainty will be underestimated and the ideal points will be biased away from zero. This result is important since a large body of work relies on ideal point estimates for the U.S. Congress to study polarization (McCarty et al., 2016; Tausanovitch and Warshaw, 2017; Fiorina et al., 2006). If the ideal points estimated in this way are biased away from zero, then estimates of polarization are overstated.

There is also evidence that misspecification of group effects in the ideal point model can lead to poor inference about the number of dimensions needed to model the ideal point space. Aldrich et al. (2014) show that in simulations of ideal point spaces with a known number of policy dimensions, the standard roll call scaling model tends to underestimate the dimensionality of the space when there are strong group effects. This problem can be mitigated by scaling any known groups separately. While the American legislative context may appear to lend itself to prior specification of voting groups, this assumption might be too strong. In order to appropriately estimate dimensionality, systematic correlations among votes due to groups should be estimated, then taken into account.

Beyond problems of dependence in errors, roll call scaling using the standard model also suffers from attempting to place a discrete latent variable on a continuous scale. Theories of the role of groups in the U.S. Congress posit that there is a tiered approach to vote decision making - first members of the party consult with the party desire then decide whether or not they should depart from the party wishes (Aldrich and Rohde, 2000; Rohde, 2010; Cox and McCubbins, 2007). While there are certainly cases where this model is not appropriate, there is evidence that party plays a role in the vote decision. The first dimension of the NOMINATE model is correctly interpreted as party loyalty (Lee, 2009), and is often made up of distinctive clusters of ideal points. These clusters are interpreted as parties and the distances are used to make statements related to the behavior of parties in Congress (McCarty et al., 2016). However, this usage of ideal points is not supported by the theory of the model - if party is a cause of votes, then party should be accounted for in the ideal point estimates. However, party is a discrete covariate. Thus, the continuous model is unequipped to properly estimate these effects. The effect of mapping a discrete latent variable to a continuous manifold is explored by Aldrich et al. (2014) and is shown t0 lead to an understatement of the dimensionality of the ideal point space. The party effect model derived above provides an approach to properly estimating the underlying group latent variables.

This theoretical exercise demonstrates the importance of ensuring that potential group effects are taken into account when modeling roll call outcomes. However, there are two significant challenges in estimating and interpreting the group model. First, it is too strong of an assumption to explicitly specify the number of groups and their internal structure in a roll call model. While there are theories of voting in legislative chambers that posit that various groupings matter, there are corresponding theories which say the opposite. For this reason, it is important that any estimation procedure that attempts to model group effects is flexible and allows for the possibility that group effects may not exist - specifying group effects when group effects do not exist leads to inefficient estimations of the ideal points, which lead to inefficient estimates of the other structural parameters in the roll call model. Similarly, relying on party labels alone does not sufficiently or flexibly explore group effects.

Second, given the problems of rotational invariance present in estimates of the structural parameters of the IRT model, the estimates for $\boldsymbol{\xi}_i$, $\boldsymbol{\tau}_{g_i}$, and β_i are not uniquely identifiable - $\boldsymbol{\xi}_i$ has infinite solutions, even if $\boldsymbol{\omega}_i$ would be uniquely identified. Attempting to estimate two components of an additive effect leads to more problems and requires more stringent constraints in order to identify a unique solution. For these reasons, it is worthwhile to derive an estimation strategy for the group effects model that allows for an infinitely exchangeable prior specification; rather than estimating $\boldsymbol{\xi}_i$ and $\boldsymbol{\tau}_{g_i}$ separately, the sum of the two is estimated, $\boldsymbol{\omega}_i = \boldsymbol{\xi}_i + \boldsymbol{\tau}_{g_i}$, such that $\epsilon_{i,j}|\boldsymbol{\omega}_i, g_i \perp \epsilon_{i',j}|\boldsymbol{\omega}_i', g_{i'} \; \forall \; j \in (1, ..., P)$, $i \neq i'$.

3.4 Clustered Beta Process IRT

A key assumption for the latent variable models previously discussed is that each of the N observations are *independently and identically distributed.* Each of the i random vectors associated with the observed data are modeled separately and assumed to constitute a joint likelihood:

$$P(Y|-) = \prod_{i=1}^{N} P(y_i|-) \tag{3.11}$$

Put another way, each observation is assumed to be exchangeable conditional on the structural parameters. In the context of ideal point estimation, the exchangeability assumption arises in the estimation of the latent variables, $\boldsymbol{\Omega}$. *A priori*, $\boldsymbol{\omega}_i$ is assumed to follow:

$$\boldsymbol{\omega}_i \sim \mathcal{N}_K(0, \boldsymbol{I}_K) \; \forall \; i \in (1, ..., N) \tag{3.12}$$

where $\boldsymbol{I}_K$ is a $K \times K$ identity matrix. This implies that each observation has the same prior on the ideal point regardless of group and directly leads to the i.i.d. result on the errors.

As pointed out previously, there is reason to doubt this assumption. Previous work on roll call scaling in a variety of legislatures has sought to address this problem. Ramey (2016), Aldrich et al. (2014), and Bernhard and Sulkin (2018), just to name a few pieces, have sought to perform a combination of within and between party scaling to account for the group incentives present in roll call voting. In each of these cases, however, scaling was done as an iterative procedure: ideal points were uncovered for each party and then mapped back to the latent space using ex-post assumptions. This approach is not ideal for a number of reasons - 1) strong prior theory is used to say that parties are the only groups that matter, 2) the link back to a common space assumes that strict separation exists between the parties on all issue dimensions, 3) the dimensionality of the policy space was assumed to be known, *a priori*, or was tested using ad-hoc post-processing approaches. Similar, but distinct, a *hierarchical* IRT formulation has been explored in recent years (Rai and Daumé, 2009; Gruhl et al., 2013), but the applied literature has given very little attention to this class of models. While hierarchical IRT addresses the issue of combining the groups into a common latent space, it still ignores that many applied situations have data where groups are unknown.

An alternative scaling approach is presented by Spirling and Quinn (2010). Rather than seeking to extract continuous manifolds, Spirling and Quinn (2010) seek to find latent clusters within binary roll call votes from the U.K. parliament. This approach addresses many of the issues presented by avoiding explicit scaling and assuming that members are wholly loyal to a single group. This allows for the estimation of beta-Bernoulli clusters from the binary data. However, this comes at the cost of continuous ideal points in a reduced dimensional setting - clusters are still associated with a P-dimensional mean and covariance. Since one of the goals of ideal point scaling is to project high dimensional roll call data into a lower-dimensional substantively meaningful space, Similarly, this model sacrifices specification of individual departures from the overall group. While this is certainly appropriate for the strong party systems present in the U.K. parliament, there has been significant work in the U.S. legislature that shows that individual legislators mix individual and party incentives when making vote choices (Sulkin, 2005; Lee, 2009).

3.4.1 Uncovering Unknown Latent Groups

I seek to derive an approach to estimating groups in the ideal point model that places each of the i legislators in one of G groups:

$$P(y_i) = \int_{0}^{\infty} \int_{\Theta} \int_{\omega_i} \int_{G} \mathcal{N}_p(x_i \; ; \; (\mathbf{\Lambda} \odot \boldsymbol{R})\boldsymbol{\omega}_i - \boldsymbol{\alpha}, \boldsymbol{\mathcal{I}}_P) P(\boldsymbol{\omega}_i|g_i) P(g_i|\boldsymbol{g}_i^*) dG d\omega_i d\Theta dx_i \quad (3.13)$$

where g_i is a categorical variable with probability densities corresponding to the probability that observation i is part of cluster $g \in (1, ..., G)$, $\boldsymbol{g}_i^*$ and $\boldsymbol{\Theta}$ is the collection of all other parameters not explicitly written above.

In the case where g_i is unknown, but the number of clusters, $|G|$, is known, then a prior can be placed on g_i such that:

$$P(g_i) \sim Cat(\zeta) \quad (3.14)$$

where Cat(ζ) corresponds to a $|G|$-dimensional categorical distribution. ζ is a set of $|G|$ probabilities that sums to 1.

However, the case of interest in this context is one where the number of clusters within the data is *unknown*. Thus, a standard categorical distribution is inappropriate. For this situation, a Bayesian nonparametric approach will be used and g_i

assumed to follow a Dirichlet process (Ferguson, 1973). The Dirichlet process can be seen as a prior over distributions. Each $g \in (1, ..., G)$ constitutes a cluster with an associated distribution. The Dirichlet process prior admits a probability density over the membership for observation i. In other words:

$$P(g_i) \sim DP(\beta, H)$$

where β is a mixing parameter and H is a base measure. Given this choice of prior, the posterior distribution, $P(G|\Theta)$ can be defined. Due to the conjugacy of the Dirichlet prior to the categorical distribution $P(\Theta|G)$, the resulting posterior follows a Dirichlet distribution.

Given the form of the posterior distribution, we know the joint probability that $G = G^*$ for some permutation of group labels. However, this is not the quantity of interest. Rather, the inferential task asks to find $P(g_i|\Theta)$. Given that the Dirichlet process is *infinitely exchangeable*, a predictive distribution can be defined and used to examine the individual probability distributions:

$$P(g_i) = \frac{1}{\beta + N}\left(\beta H + \sum_{j \neq i} \delta_{g_j}\right) \tag{3.15}$$

where H is the base measure, β is the DP hyperparameter, and $\delta_{g_j} = 1$ if observation j is in group g. In words, this says that the probability that $i \in g$ is proportional to the number of other observations that are already in group g. There is also some probability that i belongs in its own cluster, which is quantified by the base measure H and the DP hyperparameter, only.

Using this structure, a process can be defined for the case of interest- a case where the number of groups is unknown *a priori*. By driving $G \rightarrow \infty$, a process can defined where g_i is modeled as an *infinite mixture*. In words, this means that there are infinite possible cluster components within the data. In order for the model to be tractable, a finite number of these components should be used within the data. Thus, of the $G = \infty$ possible clusters that g_i can join, only G^+ have a positive probability of occurrence. This process constitutes the *Chinese Restaurant Process* (Rasmussen, 2000; Pitman et al., 2002; Blackwell and MacQueen, 1973; Escobar and West, 1995). Under this prior, each observation is assumed to take on a currently defined cluster with probability proportional to the number of other observations that are currently in that cluster and take a new cluster assignment with probability proportional to β. This prior is sparsity inducing as the expected number of clusters grows with N:

$$E[G^+] \approx O(\log(N)) \tag{3.16}$$

This implies that the complexity of the inferred mixture can only be as complex as the sample size allows. This is desirable as it prevents against overfitting.

As with all mixture models, the cluster assignments achieved using the CRP are unique up to a permutation of labels; the model is identified, but the individual cluster assignments are not. However, the groupings are identifiable. For this reason, inference performs not on the individual cluster labels, themselves, but the probability that $g_i = g_j \forall i \neq j$.

3.4.2 Latent Groups within Ideal Point Estimation

The key quantity of interest in this model is $P(g_i|\boldsymbol{g}_{-i})$. Using Bayes rule, this quantity can be defined as:

$$P(g_i = g|\boldsymbol{g}_{-i}) \propto \int\limits_{\omega_i} \int\limits_{g} \mathcal{N}_p(x_i\ ;\ (\boldsymbol{\Lambda} \odot \boldsymbol{R})\boldsymbol{\omega}_i - \boldsymbol{\alpha}, \boldsymbol{\mathcal{I}}_P) P(\boldsymbol{\omega}_i|g_i = g) d\omega_i dg \tag{3.17}$$

Note that finding g_i amounts to evaluating the probability of legislator i's roll call record given the model for each of the groups currently active under the CRP and integrating over the ideal point. Without further specification of the distributional structure of $\omega_i|g_i$, this model is intractable.

For the purposes of ideal point analysis, multivariate normal clusters are assumed. For each $g \in G$, the corresponding cluster is assumed to follow a K^+ dimensional multivariate normal cluster with mean vector μ_g and diagonal covariance matrix Σ_g, where K^+ is dictated by the beta process prior on the loadings matrix (McAlister, 2018). Standard conjugate priors are assumed on these parameters. Plugging this into (3.17), we can redefine the model as:

$$\begin{aligned} P(g_i = g|\boldsymbol{g}_{-i}) \propto \int\limits_{\omega_i}\int\limits_{\mu_g}\int\limits_{\Sigma_g} &\mathcal{N}_p(x_i\ ;\ (\boldsymbol{\Lambda} \odot \boldsymbol{R})\boldsymbol{\omega}_i - \boldsymbol{\alpha}, \boldsymbol{\mathcal{I}}_P) \\ &\qquad\qquad \mathcal{N}_{K^+}(\boldsymbol{\omega}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\mathcal{N}_{K^+}(\boldsymbol{\mu}_g; \boldsymbol{\mu}_0, (\kappa_0\boldsymbol{\Sigma}_g)^{-1}) \\ &\qquad\qquad \prod_{k=1}^{K^+} IG(\sigma_{g,k}; \alpha_0, \beta_0) d\Sigma_g d\mu_g d\omega_i \end{aligned} \tag{3.18}$$

This quantity encapsulates the probability that i belongs in cluster g given all other observations that are currently in g.

This integral can be simplified by recognizing that the latter part of the integral is the posterior predictive distribution for the conjugate multivariate normal model. Simplifying gives:

$$P(g_i = g|\boldsymbol{g}_{-i}) \propto \int\limits_{\omega_i}\int\limits_{\mu_g}\int\limits_{\Sigma_g} \mathcal{N}_p(x_i\ ;\ (\boldsymbol{\Lambda} \odot \boldsymbol{R})\boldsymbol{\omega}_i - \boldsymbol{\alpha}, \boldsymbol{\mathcal{I}}_P)\mathcal{T}_{K^+;2\alpha_g}\left(\mu_g, \frac{\beta_g(\kappa_g + 1)}{\alpha_g\kappa_g}\right) d\omega_i \tag{3.19}$$

where $\mathcal{T}_{K;d}(\mu, \Sigma)$ is the K dimensional multivariate t-distribution with location μ, scale matrix Σ, and d degrees of freedom. Let $\boldsymbol{\omega}_{-i,g}$ be the values of the latent variable for all observations in cluster g **not including** observation i and n_g be the number of observations currently in g **excluding** observation i. Define:

$$\bar{\boldsymbol{\omega}}_{-i,g} = \frac{\sum \boldsymbol{\omega}_{-i,g}}{n_g}$$

$$\mu_g = \frac{\kappa_0 \mu_0 + n_g \bar{\boldsymbol{\omega}}_{-i,g}}{\kappa_0 + n_g}$$

$$\kappa_g = \kappa_0 + n_g$$

$$\alpha_g = \alpha_0 + \frac{n_g}{2}$$

$$\beta_g = \beta_0 + \frac{1}{2}\sum_{j=1}^{n_g} (\boldsymbol{\omega}_{-i,g,j} - \bar{\boldsymbol{\omega}}_{-i,g})^2 + \frac{\kappa_0 n_g (\bar{\boldsymbol{\omega}}_{-i,g} - \mu_0)^2}{2(\kappa_0 + n_g)}$$

Unfortunately, this integral has no analytical solution; numerical methods must be used. An accurate approach utilizes a Laplace approximation to the multivariate t-distribution. Using the quadrature approach, define the approximate normal distribution as:

$$\mathcal{T}_{K^+,2\alpha_g}\left(\boldsymbol{\omega}_i \ ; \ \boldsymbol{\mu}_g, \frac{\beta_g(\kappa_g+1)}{\alpha_g \kappa_g}\right) \approx \mathcal{N}_{K^+}\left(\boldsymbol{\omega}_i ; \boldsymbol{\mu}_g, \frac{\beta_g(\kappa_g+1)}{\left(\alpha_g + \frac{1}{2}\right)\kappa_g}\right) \tag{3.20}$$

This allows the integral to be solved analytically:

$$P(g_i = g|\boldsymbol{g}_{-i}) \propto \int\limits_{\omega_i}\int\limits_{\mu_g}\int\limits_{\Sigma_g} \mathcal{N}_p(x_i \ ; \ (\boldsymbol{\Lambda}\odot\boldsymbol{R})\boldsymbol{\omega}_i - \boldsymbol{\alpha}, \boldsymbol{\mathcal{I}}_P)\mathcal{T}_{K^+;2\alpha_g}\left(\mu_g, \frac{\beta_g(\kappa_g+1)}{\alpha_g\kappa_g}\right) d\omega_i \approx \exp(q^* - q) \tag{3.21}$$

where

$$\Xi_g = \frac{\beta_g(\kappa_g+1)}{\left(\alpha_g + \frac{1}{2}\right)\kappa_g}$$

$$A = -\frac{1}{2}(\Lambda'\Lambda + \Xi_g^{-1})$$

$$b = (y_i + \alpha)'\Lambda + \mu_g'\Xi_g^{-1}$$

$$q^* = -\frac{1}{2}((y_i+\alpha)'(y_i+\alpha) + p\log(2\pi) + \mu_g'\Xi_g^{-1}\mu_g + K^+\log(2\pi) + \log(det(\Xi_g)))$$

$$q = \frac{1}{4}bA^{-1}b' - \frac{K^+}{2}\log(2\pi) - \frac{1}{2}\log\left(\det(-\frac{1}{2}A^{-1})\right)$$

Given the choice to use a CRP prior on g_i, a base distribution must also be specified from which new clusters are proposed. In the context of factor analysis, it makes sense to define the base distribution:

$$H \sim \mathcal{N}_{K^+}(0, \boldsymbol{\mathcal{I}}_{K^+}) \tag{3.22}$$

which is equivalent to the standard prior used in factor analysis procedures. This allows the probability that g_i belongs in a new cluster, g_{new}, to be defined as:

$$P(g_i = g_{new}) \propto \int_{-\infty}^{\infty} \mathcal{N}_p(x_i\ ;\ (\boldsymbol{\Lambda}\odot\boldsymbol{R})\boldsymbol{\omega}_i - \boldsymbol{\alpha}, \boldsymbol{\mathcal{I}}_P)\mathcal{N}_k(\boldsymbol{\omega}_i; 0, \boldsymbol{\mathcal{I}}_{K^+})d\omega_i = \exp(\dot{q}^* - \dot{q}) \tag{3.23}$$

where

$$\dot{A} = -\frac{1}{2}(\Lambda'\Lambda + I_k)$$

$$\dot{b} = (y_i+\alpha)'\Lambda$$

$$\dot{q}^* = -\frac{1}{2}((y_i+\alpha)'(y_i+\alpha) + p\log(2\pi) + K^+\log(2\pi))$$

$$\dot{q} = \frac{1}{4}\dot{b}\dot{A}^{-1}\dot{b}' - \frac{K^+}{2}\log(2\pi) - \frac{1}{2}\log(\det(-\frac{1}{2}\dot{A}^{-1}))$$

Given these quantities, we can define the conditional posterior distribution for g_i. Due to the conjugacy of the DP prior and the categorical likelihood, the posterior will follow a Dirichlet distribution:

$$\begin{gathered}P(g_i|-) \sim Dirichlet(\boldsymbol{g}_i^*) \\ g_{g,i}^* = n_q P(g_i = g|\boldsymbol{g}_{-i}) \; \forall \; g \in (1, ..., G^+) \\ g_{g+1,i}^* = \beta P(g_i = g_{new})\end{gathered} \tag{3.24}$$

The probability that g_i belongs to group g is a weighted average of the number of other observations in g and the probability that g_i is in the same cluster as the other like observations. To assign legislator i to a single group, a standard Dirichlet draw is taken from $\boldsymbol{g}_i^*$.

The differences between this approach and the standard continuous latent variable model arise in how each latent variable is taken into account. In the standard procedure, each $\boldsymbol{\omega}_i$ is assumed to be *independent* of the other observations. In the clustered case, each $\boldsymbol{\omega}_i$ is assumed to belong to a cluster and information about the location of the latent variable is shared between observations in the same cluster. In other words, latent variables are essentially estimated in groups rather than individually. Using the exchangeability properties of the Dirichlet process, this addresses the initial exchangeability problem by partitioning the set of observations into conditionally exchangeable groups - observations are assumed to be exchangeable within the same cluster. This allows group properties to propagate through the latent variable rather than emerging as its own set of orthogonal latent manifolds.

This property can be expressed in another way - the latent variable associated with each individual is a function of both group and individual dynamics. Given the construction of the latent variable, there are two influences on the posterior distribution of each idea point: a group effect and an individual effect. Here, the group effect can be seen as a shared prior on the latent variable that is common across members of the same group. Then, the individual effect can be seen as a noise component that is separate from the group effect. Using the Dirichlet process approach allows the individual effect to be modeled as conditionally independent from the group effect, given the specification of the model. While it is not possible to directly estimate these two effects due to the rotational invariance that is inherent in latent variable models, post-processing procedures can be used to estimate the two effects given a specific rotation of the latent structural parameters.

A final consideration for the latent variable generating process is to examine the CRP hyperparameter, β. This parameter controls how likely it is for each observation to dictate a new cluster. When β is small, the *a priori* likelihood that an observation in cluster g finds a new unique cluster decreases. On the other hand, when β is large, clusters will have lower overall membership and be less "stable" when estimated. Finding the appropriate value for β is key for estimation. Escobar and West (1995) shows that β follows a gamma distribution and a conditional posterior can be defined as:

$$P(\beta|-) \sim \text{Gamma}(1 + C^{+}, 1 + \gamma_c log(N)) \tag{3.25}$$

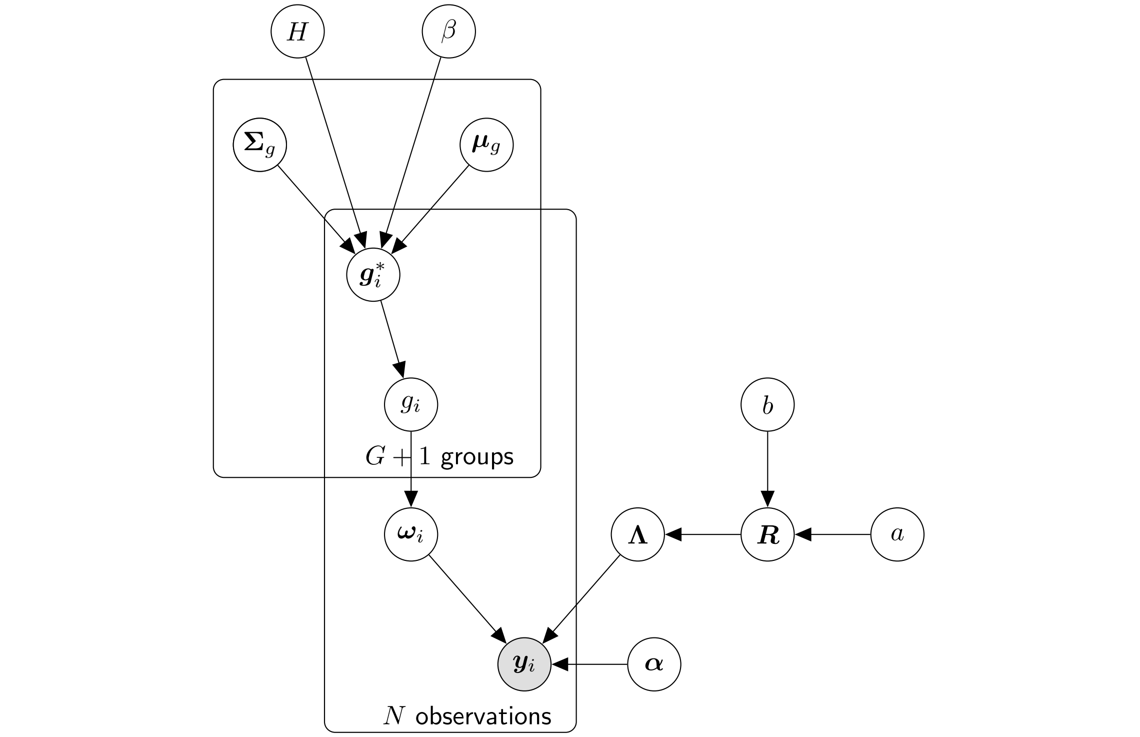
where C^{+} is the number of active clusters and γ_c is Euler's constant.

3.4.3 Estimation of the C-BPIRT Model for Binary Outcomes

Using the theory above, a roll call scaling procedure utilizing sparsity inducing beta process priors for measuring the dimensionality of the space and Dirichlet process priors on the latent variables is derived. Let $\boldsymbol{X}$ be a continuous latent mapping of the observed roll call votes $\boldsymbol{Y}$ such that:

$$x_{ij} \sim \begin{cases} \mathcal{TN}_{-\infty,0}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega_i} - \alpha_j, 1) \text{ if } y_{ij} = 0 \\ \mathcal{TN}_{0,\infty}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{ij} = 1 \\ \mathcal{N}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{ij} \text{ is missing} \end{cases} \tag{3.26}$$

FIGURE 3.2. Graphical Representation of C-BPIRT



where $\mathcal{TN}_{l,u}(\mu, \sigma^2)$ is the truncated normal distribution truncated between l and u. Placing the IBP prior on the loadings as in McAlister (2018), a Dirichlet process prior on $\boldsymbol{\Omega}$, and standard conjugate priors on the other parameters completes the full model specification. Figure 3.2 shows a graphical representation of the C-BPIRT model.[5]

Due to the conjugacy between the Dirichlet Process prior/ideal points and Indian Buffet Process prior/factor loadings, Gibbs sampling can be used to estimate the joint posterior of all parameters in the C-BPIRT model. Gibbs sampling steps are given in Appendix A.

This model has many of the same features as the basic factor analysis model. The manifest variables are decomposed into the loadings matrix, Λ, and the latent vari-

[5] In contrast to McAlister (2018), this paper uses a slightly modified finite variant of the Indian Buffet Process. Details on this approach can be found in Paisley and Carin (2009). The major difference between these approaches regards the posterior over R. Rather than defining a full posterior over R, a lower bound is found on K^+ and the maximum a posteriori arrangement of zeros and ones in R is found. This approach compares favorably to the full IBP specification and there is little reason to believe that it produces substantively different results (Doshi et al., 2009).

ables, Ω. The dimensions of the latent variables are assumed to be orthogonal. Marginally, $P(\boldsymbol{x}_i + \boldsymbol{\alpha}) \sim \mathcal{N}_{K+}(\mathbf{0}, \boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \boldsymbol{\mathcal{I}}_K)$. The new additions, however, provide interesting properties for the factor analysis procedure. First, the addition of the infinite binary matrix, $\boldsymbol{R}$, allows learning about the true number of latent dimensions within the manifest data. $\boldsymbol{R}$ represents whether feature $k \in (1, ..., \infty)$ is a meaningful summary of the data. Similarly, the CRP prior on $\boldsymbol{\Omega}$ sorts the latent variables into meaningful groups, allowing simultaneous estimation of latent groupings with corresponding location and spread information. In contrast to simply clustering the data, this model grants the desirable properties of knowing latent groups while also having the same continuous measure properties that make factor analysis an attractive approach to estimating a latent space.

A problem that is common to all ideal point specification is that the estimates for the structural parameters are not uniquely identified without further constraints - identical estimates of $\boldsymbol{\Omega}$ can be achieved by multiplying $\boldsymbol{\Lambda}$ by an orthonormal matrix, $\boldsymbol{M}$, such that $\boldsymbol{M}\boldsymbol{M}' = \boldsymbol{\mathcal{I}}_P$. Following a common convention to ensure identifiability, many implementations of Bayesian factor analysis assume that $\boldsymbol{\Lambda}$ has a full-rank lower triangular structure with positive elements on the diagonal (Geweke and Zhou, 1996). The spirit of this recommendation relies on the notion that structural zeroes can be placed in the loadings matrix in accordance with theory. However, this is rarely achievable as the theory behind a latent space, especially those with more than one dimension, is difficult to put in terms of the loadings matrix. Similarly, the resulting prior on $\boldsymbol{\Lambda}$ is no longer exchangeable. When these constraints are placed in an ad-hoc manner, they can lead to significant multimodalities in the resulting posterior.

C-BPIRT places zero constraints by virtue of the Indian Buffet Process prior on the loadings matrix. The properties of the IBP prior lead to a sparse solution for the loadings matrix that requires that the number of non-zero elements for each dimension increases as each dimension is added. While this prior **does not guarantee** a uniquely identifiable rotation of the latent space, most applications of C-BPIRT result in unimodal posteriors for the loadings and ideal points, indicating that enough zeros have been placed in the loadings matrix to uniquely identify a rotation. As always, it is prudent to check the resulting posteriors for evidence of rotational reflections. Should these reflections arise, post-processing methods presented by Gruhl et al. (2013) can be used.

A key quantity of interest using this model is the set of group assignments, $g_i \in (1, ..., G) \; \forall \; (1, ..., N)$. As mentioned previously, inference on g_i is difficult - when estimated via MCMC, the model is only identifiable up to a permutation of the cluster labels. From iteration to iteration in the Gibbs sampler, there is no guarantee that the cluster labels are significant. However, the membership of the clusters is consistent. Though there are numerous solutions to the label switching problem when the number of clusters is known and fixed, the fact that the number of clusters is stochastic proves problematic for these approaches. Thus, inference about group membership should proceed examining the probability that two or more observations share the same cluster. Using this information, many interesting techniques can be used to establish cluster membership. An alternative method uses the cluster labels from the *maximum a posteriori* iteration from the posterior Monte Carlo draws. While this approach loses some of the information about posterior uncertainty around the cluster moments and cluster membership, it allows for a simple sorting of observation into classes.

A final point is that the C-BPIRT model allows for estimation of a meaningful posterior predictive interval on the ideal points given group membership. Specifically:

$$P(\boldsymbol{\omega}_{N+1}|g^*_{N+1} = g \in G, \boldsymbol{\Omega}, \boldsymbol{G}, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \sim \mathcal{T}_{K^+,n_g}(\boldsymbol{\omega}_N; \hat{\boldsymbol{\mu}}_g; \hat{\boldsymbol{\Sigma}}_g)P(\boldsymbol{\omega}_N) \qquad (3.27)$$

Under the reasonable assumption that the prior on the new ideal point has an improper uniform prior, the posterior predictive distribution on the new ideal point is simply a multivariate $\mathcal{T}$ distribution with location and scale equal to the mean of the posterior distributions of the mean and variance of the group's multivariate normal cluster. Posterior predictive distributions are useful for examining new ideal points and a common criticism of current roll call scaling methods is that new legislators cannot be easily placed in the common space without strong prior assumptions about their relative rankings compared to other legislators. Group membership, however, is much easier to assign - new legislators are elected under a party label, so assigning them to the group that is most commonly associated with that party would provide reasonable predictions on their votes. Similarly, this approach can be used to think of common group ideal points. Assuming that all members of a group share the same underlying distribution over their ideal points, ideal points can be drawn from the cluster distribution or all legislators in the group can be assumed to have a common ideal point. Either choice can lead to interesting models that could not be achieved using coll call scaling methods that assume independent and identically distributed ideal points and errors.

3.5 Application to Roll Call Scaling in the U.S. Congress

I illustrate the benefits of C-BPIRT in two legislative voting settings. First, I explore roll call voting in the 107^{th} session (2001 - 2003) of the U.S. House. I use C-BPIRT to uncover interesting group related voting, particularly regarding the Bipartisan Campaign Reform Act of 2002. Second, I explore group dynamics in the 88^{th} session of the U.S. House (1963-1965). This session saw major civil rights and social welfare reforms and is known, historically, as a session where parties splintered on these issues. Each legislative setting explored features a situation where group dynamics beyond party loyalty played a key role in dictating policy outcomes. Unlike standard roll call scaling techniques that ignore these group dependencies, C-BPIRT allows for both group and individual ideal point estimation and provides rich insight into the sets of votes and individuals that shape legislation through non-party coalition behavior.

Roll call scaling using C-BPIRT uses the set of all roll call votes and U.S. Representatives for each session. Votes that were unanimous or nearly unanimous were excluded (i.e. less than five voters in the minority). Similarly, legislators that cast a "Yea" or "Nay" vote in less than 75% of the total roll calls were excluded from the analysis. For each session, two chains of the MCMC procedure were run with a healthy burnin of 10,000 iterations and 10,000 Monte Carlo samples were taken from the posterior distribution. Across all chains, there were no apparent issues of convergence.

3.5.1 107th U.S. Congress

I begin by exploring roll call voting behavior in the 107^{th} session of the U.S. House. This session began in 2001 and ended in 2003. During this time, members of the U.S. Congress dealt with a variety of pressing issues regarding the September 11^{th} attacks and the resulting war on terror. These issues ranged from funding of the war efforts, security on domestic soil, immigration reform, and budget reform to ensure that the resulting economic depression did not interrupt vital government services. Along with issues of national security, the U.S. Congress also addressed the numerous financial scandals in previous years and sought to address the relatively unfettered power large corporations had in donating money to political campaigns. These two issue areas defined much of the within and between party conflict of this session.

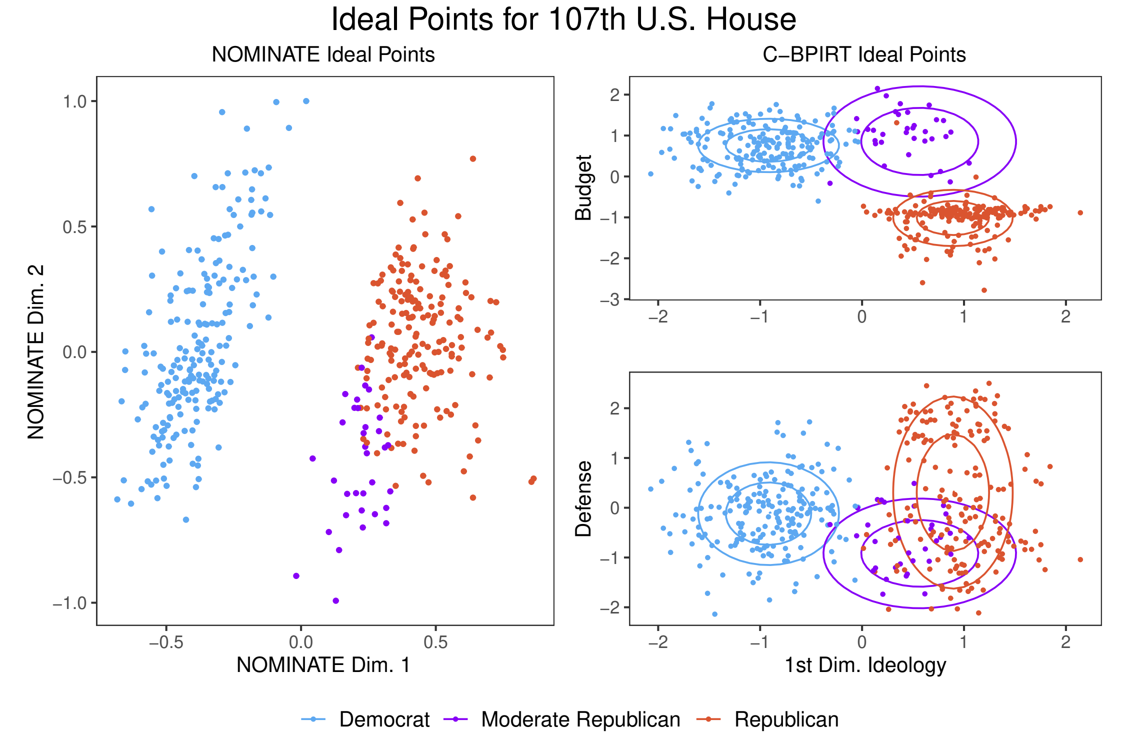
Legislative actions taken in this session generally followed party lines. Democrats had a slim majority of the seats in this session and Republicans often sought to draw a number of more conservative Democrats across the aisle to pass more conservative legislation. However, finance reform and issues related to counterterrorism and the war effort created strong cleavages in both parties. Given their slim majority, however, Democrats controlled much of the roll call agenda and most votes that went to the floor resulted in Democrats voting along the party line. Republicans, on the other hand, were often split in roll call votes - regardless of them being needed to create a winning coalition, Republicans showed splits in voting on issues of defense and financial reform.

Even with these complications, current roll call scaling methods like NOMINATE and Bayesian ideal point estimation show that the 107^{th} U.S. House is mostly described by a one-dimensional ideal point model with a second dimension needed only on a few votes (Carroll et al., 2009; Clinton et al., 2004). Beyond the simple count of dimensions, there does not appear to be a strong correlation between necessity of the second dimension in NOMINATE and the topic of the roll call vote. While this is supported through generic classification quality metrics, the interpretation of the 107^{th} U.S. House granted by NOMINATE ideal points does not align with the multifaceted and group driven narrative which actually drove this session of Congress.

McAlister (2018) finds that one aspect of this misalignment comes from inappropriate testing of dimensionality. While NOMINATE finds that the 107^{th} U.S. House was largely one dimensional, McAlister (2018) finds strong evidence that there are many dimensions which lead to roll call outcomes. While the first dimension shown by BPIRT correlates heavily with the first dimension of NOMINATE ideal points, six other meaningful dimensions are uncovered by the beta process priors on the latent space. These six dimensions show votes where intraparty disagreement in roll call votes occurs and are largely related to the issue sets that are known to have created variation in roll call voting within parties - budgetary issues and topics related to the September 11^{th} attacks.

The analysis using BPIRT still suffers from the assumption that all errors, even in many dimensions, are independent and identically distributed after accounting for the individual ideal points. While the new dimensions are certainly informative to

FIGURE 3.3. Ideal point estimates from NOMINATE and C-BPIRT for the 107th U.S. House.



Note: Plotted C-BPIRT ideal points are posterior means for each individual ideal point estimate. Group labels are given by the *maximum a posteriori* group uncovered by C-BPIRT. Plotted contours assume multivariate-$\mathcal{T}$ distributed clusters and show 66% and 95% ellipses given by the posterior mean and covariance for each group.

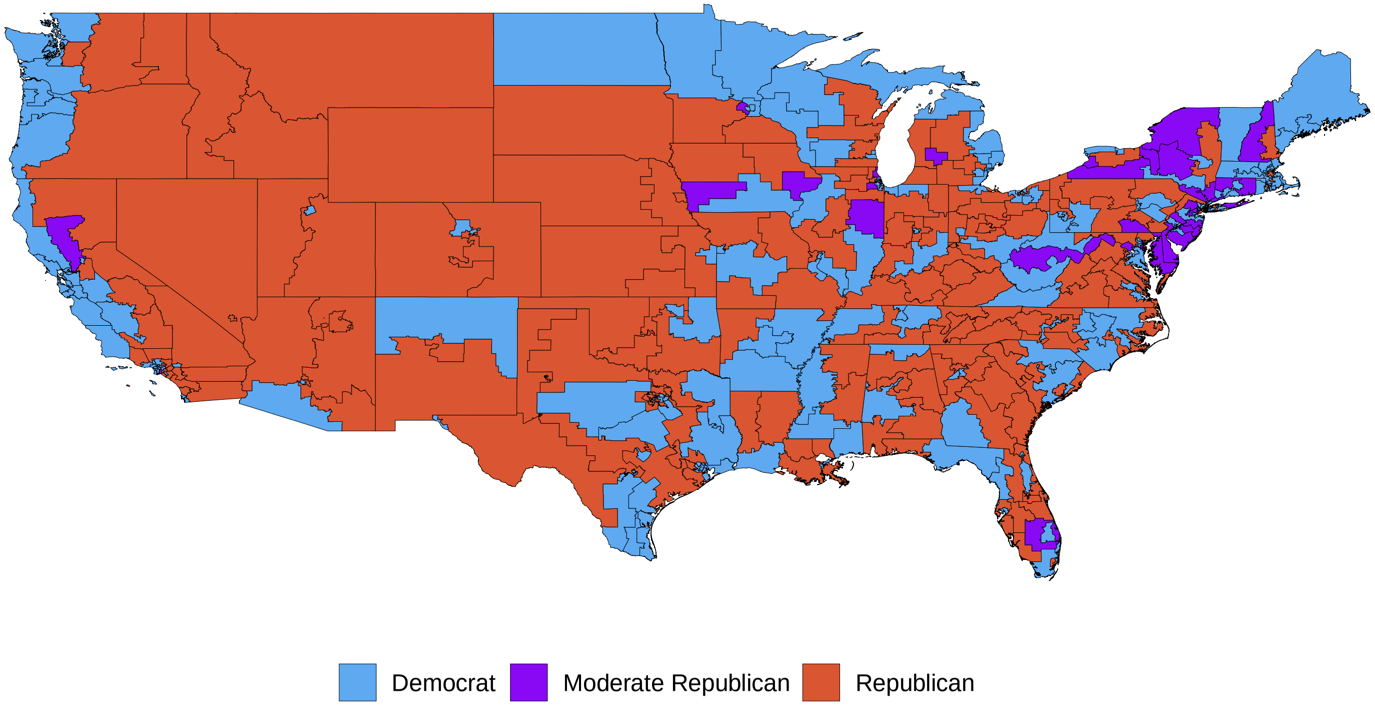
complexity in the roll call votes, there is still evidence of correlated errors due to shared party incentives. This can be seen most clearly in the estimates that make up the first dimension of roll call voting as they appear to simply map a notion of party loyalty (Lee, 2009) and shows a strong separation between parties. While some literature argues that this is simply a feature of legislative preferences (Cox and Poole, 2002; Krehbiel, 1992), modern literature on Congressional voting shows that this feature of the first dimension is an artifact of group correlations (Aldrich et al., 2014; Ramey, 2016). In an ideal scenario, groups would be mapped to a discrete measure in the ideal point space and variations in voting beyond groups would be mapped to the continuous latent space.

C-BPIRT offers an alternative to NOMINATE and BPIRT that accounts for these correlations and, theoretically, maps the party loyalty of a voter to a discrete measure. In turn, this procedure creates conditionally exchangeable errors that are more in line with the necessary independent and identically distributed assumption that is baked into roll call scaling procedures. C-BPIRT also provides many useful pieces of information about what group dynamics exist in the legislature and how they influence roll call voting outcomes. Applying C-BPIRT to the 107^{th} U.S. House and comparing its results to those attained from NOMINATE and BPIRT should demonstrate how accounting for group voting changes interpretations ideal point estimates and provides new insights into legislative voting behavior.

Roll call scaling using C-BPIRT produces three substantively interesting sets of estimates: 1) multivariate $\mathcal{T}$ groupings in the ideal point space and the corresponding means and covariances, 2) the number of dimensions and the votes that load on each dimension, and 3) individual legislator ideal points. For the 107^{th} U.S. House, this information can be seen in Figure 3.3. C-BPIRT scaling on the 107^{th} U.S. House reveals three distinct voting groups: Democrats, Republicans, and a set of 38 Moderate Republicans. These voting groups strictly follow party lines with no crossover between parties. A map of U.S. House districts and the corresponding group uncovered by C-BPIRT is shown in Figure 3.4. Figure 3.4 shows that the group of Moderate Republicans is not a surprising finding; the members of the 107^{th} U.S. house in this group generally represent Republican districts in New England. Defined by more liberal views on social issues and a commitment to social welfare programs, this group of Republican representatives is well known in U.S. legislative history and has long played the part of a pivotal group in determining landmark legislative outcomes (Lee, 2009; Poole and Rosenthal, 1984).

In contrast to NOMINATE's 1-2 dimensional space and BPIRT's 7 dimensional space, C-BPIRT uncovers three dimensions. These dimensions generally correlate to three

FIGURE 3.4. Map of U.S. House districts and group labels for the 107^{th} Session of the U.S. House.



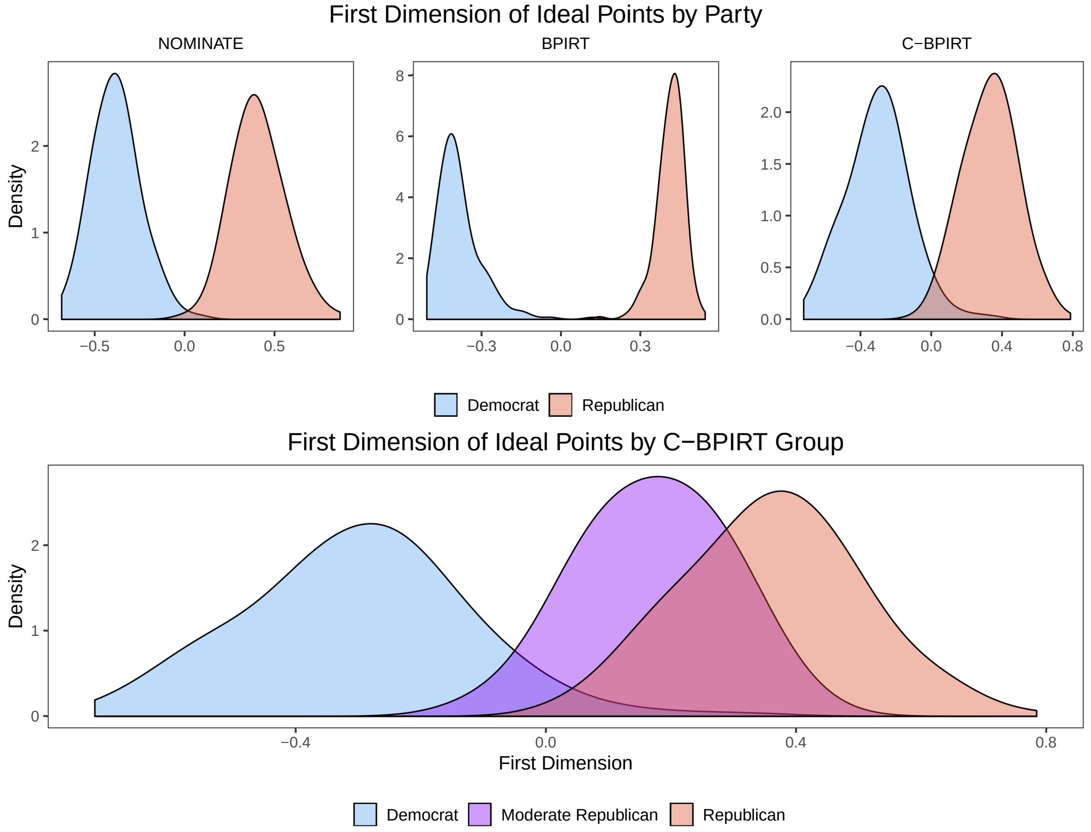
Note: Districts are colored by the *maximum a posteriori* group labels uncovered by C-BPIRT.

areas: an overall ideology dimension, similar to NOMINATE's first dimension, a dimension that includes votes on the budget and other financial issues, and a dimension related to votes on defense spending and counterterrorism in response to the 9/11 attacks. Figure 3.3 demonstrates the relationship between groups and uncovered dimensions and shows that dimensions emerge as a combination of two scenarios. First, a dimension emerges when a distinct voting coalition needs to be modeled. The first dimension corresponds to votes where the moderate Republican cluster votes with the Republican cluster and against the Democrat cluster. On the other hand, the budget dimension corresponds to votes where the Moderate Republicans coalesce with the Democrats. The defense dimension shows the second scenario - dimensions emerge when clustered voting does not seem to model voting outcomes. In other words, the defense dimension includes a set of votes where assuming similar voting within clusters provides a poor model of roll call voting outcomes.

In many ways, this behavior is in line with the party driven roll call scaling approaches of Ramey (2016) and Aldrich et al. (2014) and the numerous theories of party organization in U.S. legislative voting (Aldrich, 1995; Aldrich and Rohde, 2000; Rohde, 2010). While NOMINATE and BPIRT scores assume that all individuals compute the utility calculus for each vote independent of one another, C-BPIRT assumes that there is some shared component to these utility computations. C-BPIRT leads to a representation of the policy space uncovered by roll call scaling and ideal point estimates that are more in line with modern conceptions of the many influences which drive legislative voting.

A first important difference between the estimates of independent and identically distributed ideal points and C-BPIRT is the structure of the first dimension. Unlike NOMINATE and BPIRT, C-BPIRT estimate that not all votes load on the first dimension; of the 645 roll call votes scaled in this session, only 438 load on the

FIGURE 3.5. Comparison of first dimension ideal point estimates from NOMINATE, BPIRT, and C-BPIRT for the 107th U.S. House



Note: Plotted densities are the posterior means for each individual ideal point estimate. For each of the estimation procedures, ideal points are grouped by party label. For C-BPIRT, points are also grouped by the *maximum a posteriori* voting group uncovered by C-BPIRT.

first dimension. This aspect of C-BPIRT's ideal points can be attributed to the grouped estimation. Since much of the party line voting is included in the estimates of the discrete grouping measure, a vote only loads on a dimension when there is additional variation that is unexplained solely by group voting. Put another way, the first dimension attained from C-BPIRT models votes where the votes of Moderate Republicans align with those from the Republican cluster and are opposite of those in the Democrat cluster.

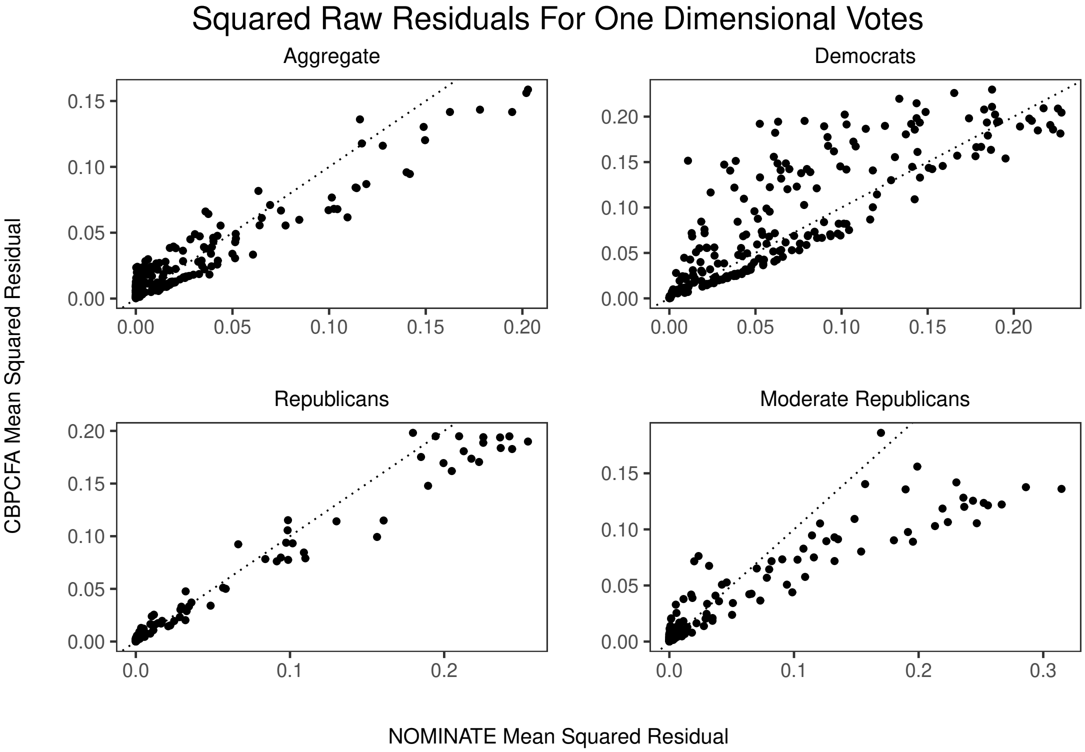
The effect of clustering can be seen in the shape and distribution of the first dimension ideal points. Figure 3.5 shows the distribution of ideal points broken out by party label on the first dimension as estimated by NOMINATE, BPIRT, and C-BPIRT. The first dimension of NOMINATE and BPIRT ideal points show a distinct division between Republican and Democrat voting. C-BPIRT, on the other hand, shows more crossover between the two parties. Looking at the bottom panel, it is clear that this is due to C-BPIRT's discovery that there exists a group of moderate Republicans. This difference between the main Republican cluster and the Moderate Republicans manifests in different ways across the models. The left panel of Figure 3.3 shows NOMINATE scores colored by the group estimated by C-BPIRT and demonstrates that the first dimension situates on votes that have party line voting while the second dimension gets the disagreement within the Republican party. BPIRT behaves in a similar way, but attributes these differences to multiple dimensions. C-BPIRT finds that Moderate Republicans cluster votes similarly across many different votes and ensures that their ideal points are close in all cases.

A key point is that breaking the ideal points into groups creates conditional exchangeability of errors and leads to better estimates of the structural parameters. While it is difficult to show which scaling method produces the best estimates, the residuals for each model can be examined to see if they better fit the independence of errors assumption that underlies all of the scaling procedures. To accomplish this, I explore 239 roll call votes that C-BPIRT estimates are one-dimensional votes and use only the first dimension of ideal points. I estimate the raw squared residuals for each individual vote implied by the first dimension NOMINATE ideal points and the posterior mean of the first dimension C-BPIRT ideal points.[6]

Figure 3.6 shows NOMINATE residuals compared to C-BPIRT residuals. Looking at the average squared residual for all votes taken together, C-BPIRT and NOMINATE perform similarly with C-BPIRT performing slightly better on votes where

[6] In this context, the raw squared residual is $(y_{ij} - \Phi^{-1}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \boldsymbol{\alpha}_j))^2$. Results using deviance residuals and studentized residuals yielded similar results.

FIGURE 3.6. Comparison of squared residuals for 239 one dimensional votes in the 107^{th} U.S. House.



Note: Raw squared predictive residuals are estimated for each vote. Residuals are compared for NOMINATE and C-BPIRT ideal points. Dotted lines are at 45 degrees and indicate equality.

NOMINATE struggles to accurately classify votes. A similar result is seen for the Republican cluster. However, NOMINATE trades accuracy in the Moderate Republican cluster for better accuracy in the larger Democrat cluster; the residuals are not independent of classification. C-BPIRT, however, corrects for this by allowing larger residuals in the Democratic cluster and passing this accuracy to the Moderate Republican cluster. In turn, this allows the residuals to uncorrelated with group labels and lead to more reasonable ideal point arrangements.

A similar effect can be seen on the second dimension of the NOMINATE ideal points, shown in Figure 3.3. The second dimension of NOMINATE correlates most heavily with membership in the Moderate Republican cluster - lower scores on the second dimension correlate with being a Moderate Republican. This implies that a vote that only used the second dimension would separate Democrats and Moderate Republicans while splitting the main Republican cluster. However, this orientation is substantively backwards. Members that are classified as Moderate Republicans in the 107^{th} U.S. House, such as Chris Shays (R,CT) and Judy Biggert (R,IL), were generally noted for their more liberal positions on social policies and issues of financial reform. In fact, much of the bipartisan legislation proposed in this session of the U.S. House was sponsored by members within this cluster. Under a group interpretation of ideal points, this would imply that there were some votes where a portion of Democrats and Republicans cast the same vote while Moderate Republicans voted opposite the Democrats and some portion of Republicans.

Examining the roll call record for this session reveals that this kind of voting never happened. The closest that a specific roll call vote gets to this kind of coalition occurs on a vote for an amendment to the Farm Bill and Rural Security Act of 2002 (Roll Call #364), where 35% of Democrats, 45% of Republicans, and 61% of Moderate Republicans cast votes to approve the amendment. Aside from this single vote, however, no other vote within this legislative session yielded a clear vote where Republicans and Democrats coalesced against the preferences of the Moderate Republicans. This sentiment is echoed by the dimensions uncovered by C-BPIRT - the only meaningful coalitions were Republicans vs. Democrats and Democrats/Moderate Republicans vs. Republicans. This result indicates that, on its own, the second NOMINATE dimension has no meaningful interpretation and only serves as a pivot to separate Moderate Republicans from Republicans on those votes which they voted against the rest of the party.

C-BPIRT corrects for this result by ensuring that ideal points are meaningfully arranged in the ideal point space. One way to demonstrate this is to examine one particularly important roll call vote in the 107^{th} U.S. House. Following his 2000 Presidential bid, John McCain led the U.S. Senate to pass a version of the Bipartisan

TABLE 3.1. Votes on the Shays Amendment to the Bipartisan Campaign Reform Act presented during the 107^{th} session of the U.S. House

	Yes	No
Democrats	200	10
Republicans	38	177

(a) By Party

	Yes	No
Democrats	200	10
Republicans	10	171
Moderate Republicans	28	6

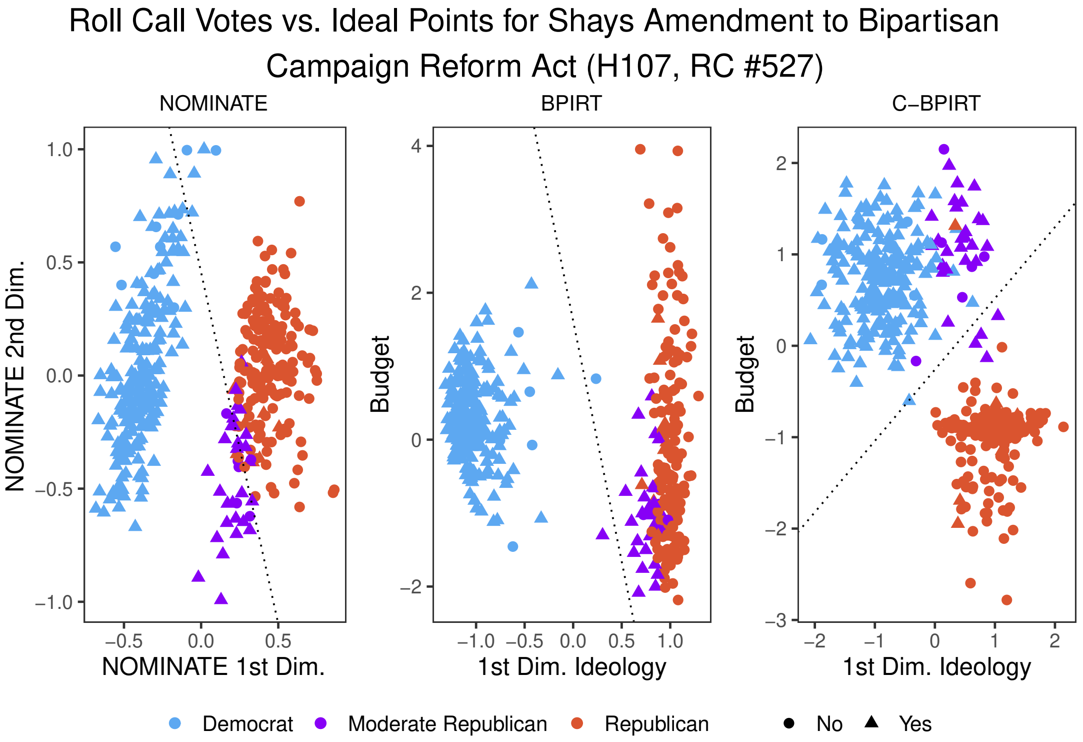
(b) By C-BPIRT Group

Campaign Finance Reform Act, frequently known as the McCain-Feingold act. This bill, written as a way to limit the influence of corporations in political campaigns, was designed by Chris Shays (R,CT) and Marty Meehan (D,MA) of the U.S. House and John McCain (R,AZ) and Russ Feingold (D, WI). Since the legislation was crafted by a bipartisan group, it was expected to have support from Moderate Republicans in the U.S. House and U.S. Senate. This support was critical for passage in the U.S. Senate as the Democrats did not have 60 votes to invoke cloture on debate. Ultimately, the Bipartisan Campaign Reform Act of 2002 passed both chambers and was signed into law.

A critical vote that occurred in the U.S. House regarded the Shays Amendment to the Bipartisan Campaign Reform Act (RC #527). This amendment sought to set further limitations on soft money given to political parties. Though the vote was championed by Chris Shays (R,CT), Republicans were not supportive of the amendment or the bill, as a whole. Table 3.1 shows the raw counts of Yes and No votes by party and by C-BPIRT groups. While party alone does not explain certain Republican support for the amendment, C-BPIRT explains this support as a function of members clustering with other Moderate Republicans.

Figure 3.7 shows the votes on this amendment as a function of ideal points estimated by NOMINATE, BPIRT, and C-BPIRT grouped by C-BPIRT groups. For BPIRT and C-BPIRT, this vote used ideal points on two dimensions: the first party line dimension and a budget dimension. Both NOMINATE and BPIRT generally explain

FIGURE 3.7. Ideal point estimates and roll call votes associated with the Shays Amendment to the Bipartisan Campaign Reform Act presented during the 107^{th} session of the U.S. House.



Note: For BPIRT and C-BPIRT, this vote was associated with only two dimensions: the overall first dimension and a dimension associated with votes related to the federal budget and campaign finance reform. Dotted lines indicate points in the ideal point space where a legislator would have a .5 probability of casting a "Yea" vote.

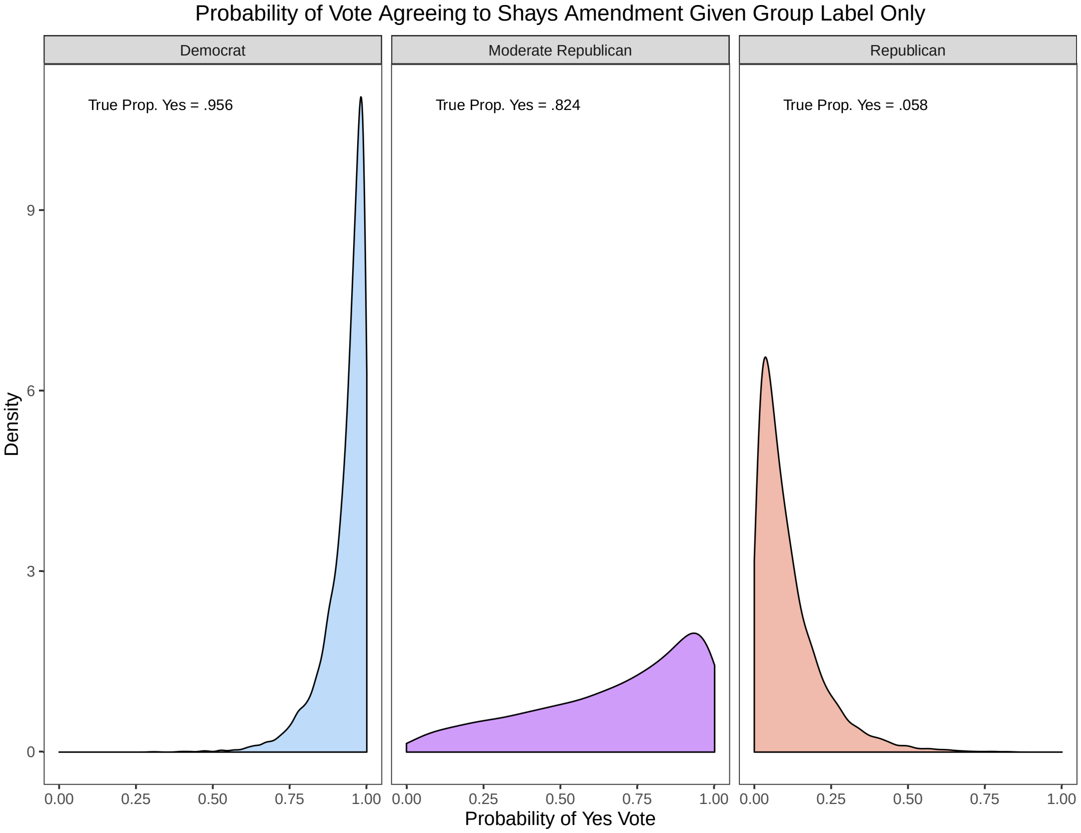
the vote in the same way - the cutline between Yes and No is mostly controlled by the first dimension and the point of indecision is within the set of Republican representatives. In both cases, the members that are close to the cutline are Moderate Republicans. Both methods, however, require some contribution from the second dimension to fit the best cutline. For NOMINATE, this second dimension ranks Moderate Republicans below Republicans and has Democrats spread across almost all viable ideal point values. BPIRT ranks Democrats in the middle with Moderate Republicans below Democrats and Republicans spread over the entire set of viable values. Neither of these ideal point arrangements points to a particularly informative ideal point space, especially for explaining votes on Shays amendment.

C-BPIRT presents estimates that align with the theoretical arrangement of ideal points that explain voting for campaign finance reform. On this vote, and other votes that have similar voting patterns, there is a two dimensional ideal point space. The first dimension accounts for party line voting while the second dimension accounts for votes where Democrats and Moderate Republicans coalesce. Given knowledge of how the series of votes regarding campaign finance reform played out in the U.S. House, this arrangement fits the prior understanding of which legislators cast Yes and No votes to shape the legislation. Unlike NOMINATE and BPIRT, C-BPIRT finds an arrangement of ideal points and cutline that places Moderate Republicans with Democrats while still explaining party line votes in a meaningful way.

The roll call vote on the Shays amendment can also be used to explore a model where individual ideal points are irrelevant and only the group ideal points matter. While C-BPIRT groups do not perfectly predict voting outcomes in this roll call vote, Table 3.1 shows that groups alone do a pretty good job of explaining the vote outcomes. This notion can be formalized by replacing the individual ideal points with the distribution of ideal points implied by the cluster moments and assuming that all members in that cluster vote in accordance with that ideal point distribution. Using C-BPIRT estimated factor loadings for this roll call vote, I used a post-processing procedure to simulate from the posterior predictive distribution of probabilities for casting a Yes vote on the Shays amendment assuming only group label is known. This simulation seeks to assess how well using only the group distribution of ideal points explains the roll call vote outcome.

Figure 3.8 shows the resulting distribution of probabilities that a member from each cluster casts a Yes vote in support of the the Shays amendment. Each density can be interpreted as a distribution over probabilities that a single voter casts a Yes vote given only their group label. Under this group model, the vote distributions

FIGURE 3.8. Posterior predictive probability distribution that a new legislator casts a "Yea" vote to approve the Shays amendment to the Bipartisan Campaign Reform Act in the 107^{th} session of the U.S. House given only a C-BPIRT group label.



Note: Posterior predictive probabilities were computed using 100,000 random samples from the multivariate-$\mathcal{T}$ clusters using the *maximum a posteriori* cluster means and variances. The true proportion of members in each cluster casting a "Yea" vote is also reported.

TABLE 3.2. Expected number of Yes and No votes on Shays amendment to the Bipartisan Campaign Reform Act in the 107^{th} session of the U.S. House given only a C-BPIRT group label.

	Yes	**No**
Democrats	195.27	14.73
	[188,202]	[8,22]
Republicans	20.38	161.62
	[12,29]	[153,170]
Moderate Republicans	23.01	11.00
	[18,28]	[6,16]

Note: Brackets under each value show 95% credible intervals.

TABLE 3.3. Proportion of Votes Correctly Classified and Geometric Mean Probability of Correct Classification for all roll call votes in the 107^{th} U.S. House.

	Correct Class	GMP
NOMINATE	0.92	0.80
BPIRT	0.92	0.81
C-BPIRT	0.92	0.84
C-BPIRT Groups	0.89	0.82

Note: Results are presented for two dimensional NOMINATE ideal points, BPIRT ideal points, C-BPIRT ideal points, and C-BPIRT votes conditional on groups alone.

generally align with those actually observed in Table 3.1 - Democrats are almost 100% likely to cast a Yes vote while Republicans are almost 100% likely to cast a No vote. Moderate Republicans show more uncertainty with a posterior mean of .65 and spread over the entire unit interval. This echoes the fact that the cutline for this vote went through the Moderate Republican cluster. Using these distributions, Table 3.2 shows expected Yes and No counts for the roll call vote using the groups only model. While there are some departures from the observed counts, this model performs admirably and estimates vote counts that are generally close to the truth. This finding is encouraging for one of the core benefits of C-BPIRT - group voting matters in U.S. legislative voting and can often yield results that are similar to those from more complicated ideal point models.

As a final point of comparison, I examine the ability of each roll call scaling model to correctly classify vote outcomes. I compare the two dimensional NOMINATE model, BPIRT with seven dimensions and corresponding binary matrix, C-BPIRT with three dimensions and individual ideal points, and C-BPIRT with three dimensions and group ideal points only using the posterior predictive ideal point distribution. For each model, I compute the proportion of votes correctly classified and the geometric mean of correct classification over all 426 representatives and 645 roll call votes analyzed.[7]

[7] The proportion of votes correctly classified finds the probability that each legislators' vote is a Yes given each models' structural parameters. A prediction is considered correct if $y_{ij} = 1$ & $P(y_{ij} =$

Table 3.3 shows the results of these comparisons. In terms of correct classification, BPIRT, NOMINATE, and C-BPIRT perform similarly. C-BPIRT using groups only yields a correct classification rate that is slightly lower, but performs admirably given that any individual deviations from assigned groups is assumed to be baked into the cluster moments. The geometric mean probability of correct classification shows similar results. However, a surprising result is that the groups only model outperforms NOMINATE and BPIRT. This speaks to the importance that group dynamics played in the legislative voting during the 107^{th} session of the U.S. House.

3.5.2 88th U.S. Congress

The 88^{th} U.S. Congress took place between 1963 and 1965. This session of Congress is often noted as one of the most important sessions due to the amount of landmark legislation that was considered and ultimately approved during this session. Over the course of these two years, major reforms to the U.S. social welfare system were made, including the Food Stamps Act of 1964, which crafted the modern social welfare system that still exists in the U.S. today. Along with social welfare issues, the 88^{th} U.S. Congress saw passage of two important civil rights acts, the Civil Rights Act of 1964 and the Voting Rights Act of 1965, which established a number of important laws that ensured equal rights and access to U.S. citizens, regardless of race. These bills are often seen as an important turning point in the U.S. Civil Rights movement of the 1950s and 1960s.

Beyond the importance of the legislation considered in this sessions, the 88^{th} U.S. Congress saw a significant splintering of parties. Historically, four groups are known to have existed and created voting coalitions within the Congressional session: Democrats,

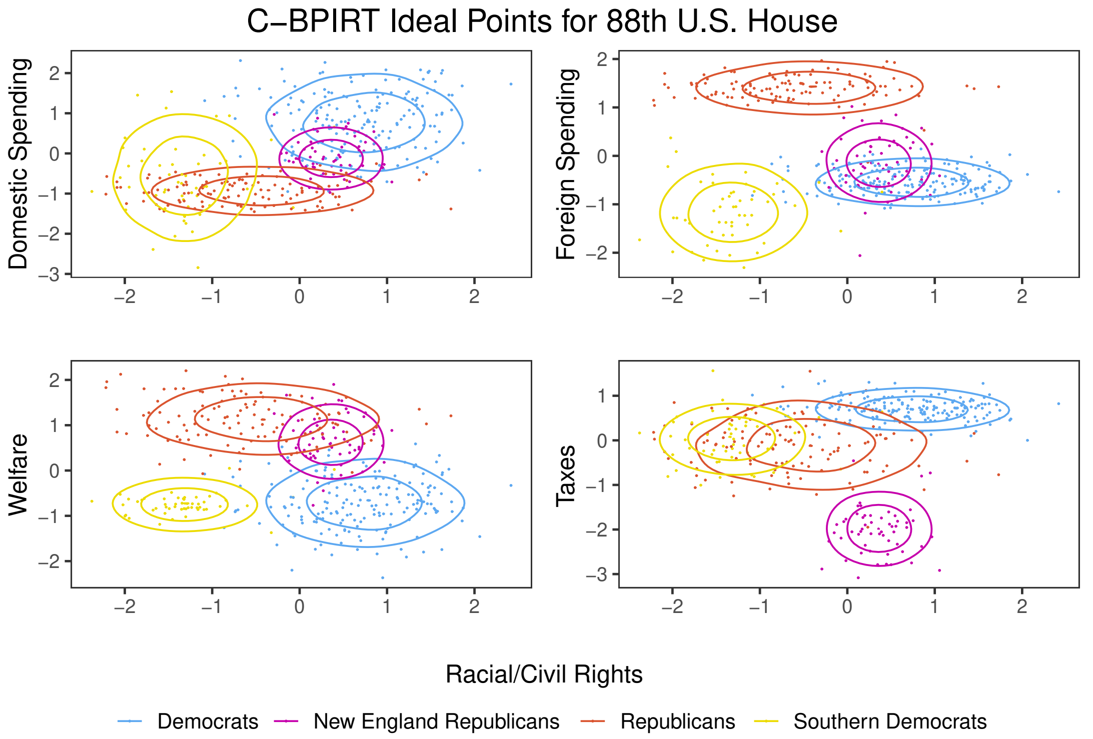
1) > .5 or $y_{ij} = 0$ & $P(y_{ij} = 1) < .5$. The geometric mean of correct classification for the entire session $\exp\left[\frac{1}{NP}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{P} y_{ij}\log(P(y_{ij}=1)) + (1-y_{ij})\log(1-P(y_{ij}=1))\right]$. This metric rewards models where incorrect classification is accompanied by probabilities closer to .5.

Republicans, Southern Democrats, and New England Republicans. Like other sessions in the 20^{th} century, all members were elected under either Democrat or Republican party membership. However, divides within each party existed. Southern Democrats were Democrats elected from the the southern United States that held strong views on racial equality and social welfare programs. While they did caucus with the Democrats, there were a number of issues where Southern Democrats split with the main Democratic party and cast votes that were against the party desires. This cleavage was particularly important to civil rights legislation, as Democrats supported sweeping reforms to the law which established equal rights regardless of race. New England Republicans were also an important group. More liberal on many issues than their Republican counterparts (and often the Democrats), New England Republicans supported sweeping civil rights changes to ensure equality for all and programs that ensured that all Americans received government assistance, when needed. However, they were largely anti-Socialist and wanted to design programs that ran efficiently and supported large tax hikes to prioritize a balanced budget.

The 88^{th} U.S. House is well-known among legislative scholars as a session where group dynamics outweighed individual voting preferences. Poole and Rosenthal (1987) demonstrates that the NOMINATE model performs better than a party/group driven model in all U.S. House sessions other than the 88^{th} U.S. House - there is strong evidence that simply modeling parties as cohesive units performs better than the standard two dimensional NOMINATE model. I argue that this finding is partially true, but the failure of NOMINATE comes from assuming that all errors are independent and identically distributed without taking group membership into account and assuming only two groups. Using C-BPIRT, I show that there is substantial variation across parties and individuals and that continuous ideal point measures provide information that would not be attainable simply from looking at party label.

As a starting point, Figure 3.9 shows the estimates of ideal points, groups, and dimen-

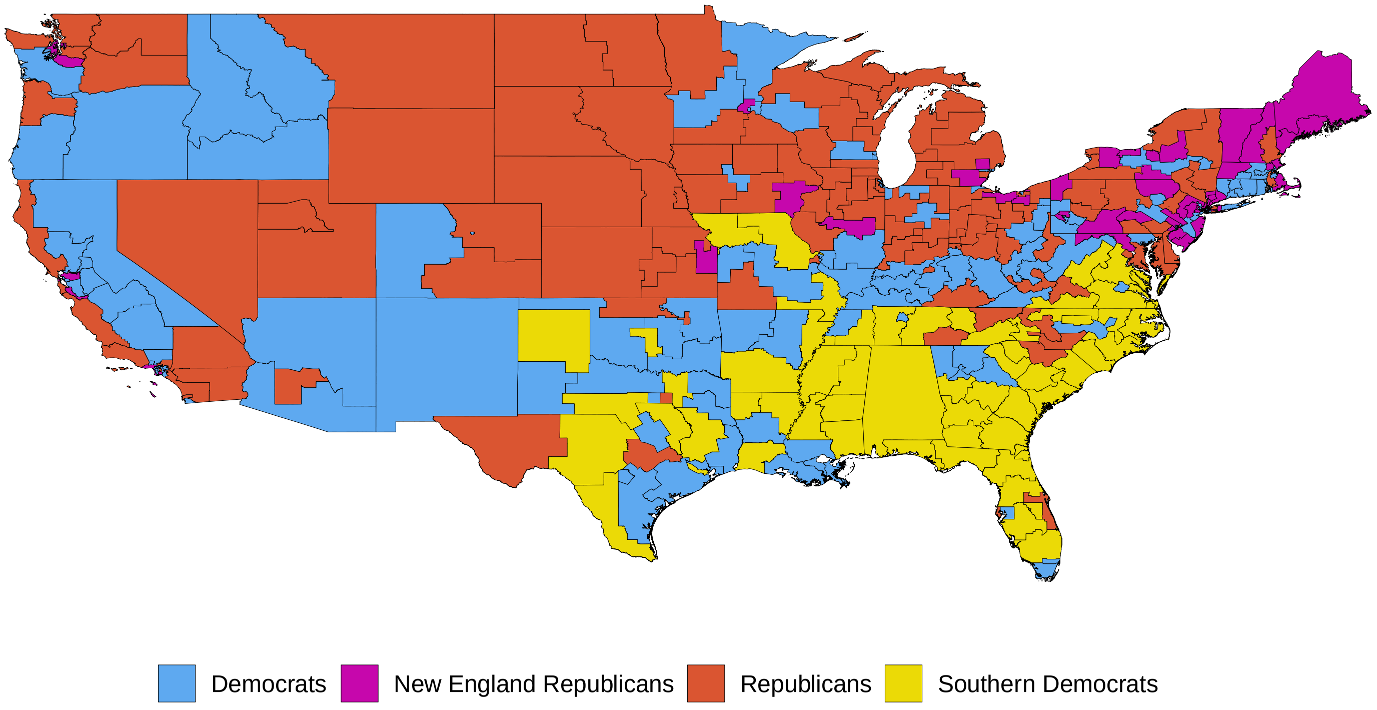
FIGURE 3.9. Ideal Points and Dimensions estimated by C-BPIRT for the 88^{th} session of the U.S. House.



Note: Individual ideal points are plotted with group labels and cluster densities associated with the highest complete data likelihood iteration of the MCMC procedure.

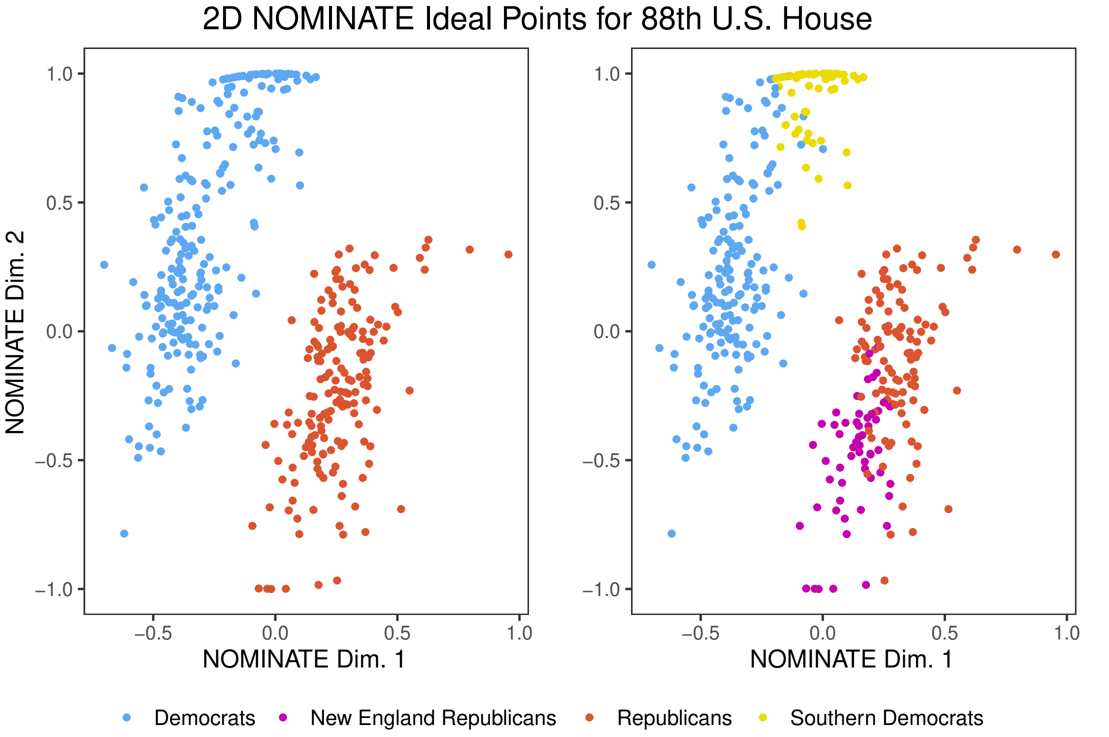
sion uncovered by C-BPIRT for the 88^{th} U.S. House. C-BPIRT uncovers four groups that align with historical expectations: Democrats, Republicans, New England Republicans, and Southern Democrats. As shown in Figure 3.10, these groups have a strong correlation with geography in the U.S. As shown previously, C-BPIRT group discovery requires finding groups and dimensions where there are unique overall arrangements of votes. To this end, C-BPIRT discovers five dimensions of voting in the 88^{th} U.S. House. As with the 107^{th} U.S. House, C-BPIRT dimensions are more sparse than their BPIRT counterparts - the first dimension, domestic spending, loads on 134 of the 190 roll call votes analyzed in contrast to 100% of votes loading on the first dimension in BPIRT. C-BPIRT uncovers four other dimensions - Racial/Civil Rights, Social Welfare, Foreign Spending, and Taxes. Each of the five dimensions uncovered

FIGURE 3.10. Map of U.S. House districts for the 88^{th} Session of the U.S. House.



Note: Districts are colored by the emphmaximum a posteriori group labels uncovered by C-BPIRT.

FIGURE 3.11. Two Dimensional NOMINATE Ideal Points for the 88^{th} U.S. House.



Note: Points are divided by party and groups uncovered by C-BPIRT.

by C-BPIRT correspond to different legislative coalition combinations. For example, the Racial/Civil Rights dimensions shows votes where New England Republicans and Democrats opposed Southern Democrats and Republicans were split. In another case, the tax dimension pits the low tax desires of the Democrats, Republicans, and Southern Democrats against the higher tax preferences by New England Republicans. The domestic spending dimension corresponds more closely to a continuous dimensions, having a loose notion of coalitional structure, but still showing significant variation within clusters. These ideal points demonstrate a dimension where individual variation is important. The consistency of dimensions with coalitions aids in interpretation of dimensions and ensures that ideal points align in meaningful ways.

Figure 3.9 demonstrates the importance of groups in determining roll call outcomes in the 88^{th} U.S. House. While the explicit clustering is unique to C-BPIRT, these groups are discussed and shown in roll call analysis using NOMINATE. Figure 3.11 shows two dimensional NOMINATE ideal points for the 88^{th} U.S. House. The first dimension of NOMINATE ideal point estimates is generally separated by party label - all Democrats, regardless of membership in the Southern Democratic coalition, are to the left of Republicans. The second dimension of NOMINATE ideal points corresponds to roll call votes corresponding to Racial/Civil Rights and the parties share similar structures; New England Republicans support civil rights legislation more than the bulk of Democrats and Republicans while Southern Democrats oppose civil rights legislation en masse. While the second dimension of NOMINATE closely aligns with the Racial/Civil Rights dimension from C-BPIRT, the first dimension seems to correspond most closely to the domestic spending dimension. However, the interpretation of the first dimension from NOMINATE ideal point estimation differs - the first dimension accounts for the largest portion of variation in roll call voting across all votes.

This interpretation is problematic when considering that the first dimension of NOMINATE is often interpreted as member ideology not including preferences on civil rights. In words, NOMINATE scores imply that Democrats are more liberal than the roughly equivalent Southern Democrats and New England Republicans which are more liberal than the Republicans. This orientation of ideal points implied by NOMINATE has two problems: 1) Southern Democrats and New England Republicans have roughly equivalent voting records outside of racial issues and 2) Southern Democrats were moderate on some set of issues. Neither of these points fit with the historical view of Southern Democrats.

To explore the idea implied by NOMINATE that Southern Democrats and New England Republicans cast similar votes on a number of issues, I analyze the probability that a representative from each C-BPIRT group casts the same roll call vote across

TABLE 3.4. Probability of same vote cast on a roll call vote for each C-BPIRT group.

	Dems.	NE Reps.	Reps.	South Dems.
Dems.	.87	.53	.33	.55
NE Reps.	.53	.76	.67	.33
Reps.	.33	.67	.77	.55
South Dems.	.55	.33	.55	.71

(a) All Votes

	Dems.	NE Reps.	Reps.	South Dems.
Dems.	.80	.51	.45	.51
NE Reps.	.53	.98	.88	.04
Reps.	.45	.88	.82	.15
South Dems.	.51	.04	.15	.93

(b) Votes on Civil Rights

	Dems.	NE Reps.	Reps.	South Dems.
Dems.	.84	.52	.33	.58
NE Reps.	.52	.77	.67	.35
Reps.	.33	.67	.75	.55
South Dems.	.58	.35	.55	.69

(c) All Votes, No Civil Rights

Note: Votes are broken out in three ways: 1) All 190 roll call votes, 2) 18 roll call votes on civil rights, 3) 172 Votes on non-civil rights issues. Vote classifications were retrieved from the Policy Agenda Project (Adler and Wilkerson, 2006).

votes as both those within the same group and those from other groups. Along with aggregate probabilities, I also find the probability that they share common votes on Civil Rights votes and non-Civil rights votes. Probabilities from this analysis can be seen in Table 3.4. From this analysis, it is clear that there is very little evidence for Southern Democrats and New England Republicans to share similar ideal points on an aggregate ideology dimension. For all votes, New England Republicans and Southern Democrats cast the same vote the same proportion of times that Republicans and Democrats cast the same vote. As the goal of roll call scaling to place like voting patterns in similar locations of the latent space, placing Southern Democrats and New England Republicans together is just as bad as placing Democrats and Republicans in the same location. Removing the 18 votes on civil rights from the analysis yields a similar conclusion. This examination shows that there is no meaningful way that the first dimension of ideal points can be twisted to make them align in a meaningful way. This further shows the inappropriateness of using the first dimension of NOMINATE scores as a substitute for legislator ideology.[8]

Beyond placing New England Republicans and Southern Democrats in the same place on the first dimension, NOMINATE implies that Southern Democrats represent a moderate group, outside of racial issues. Aside from the domestic spending dimension, which shows no clear group divides across all groups, Figure 3.9 shows no evidence that Southern Democrats had ideal points that are in the middle on any issue set. On foreign spending, Southern Democrats represent a policy position that strongly opposes defense and foreign investment - more extreme than Democrats and New England Republicans. On welfare issues, their position is similar to that of Democrats. Finally, on taxation issues, they exhibit preferences similar to those seen by Republicans and Democrats. Given that there is no clear place where the Southern Democrats represent a moderate policy position, NOMINATE provides a misleading characterization of Southern Democrat preferences. The implied moderate position can be chalked up to overaggregation and Simpson's paradox. Since the vast majority of legislators are in the Democrat or Republican cluster (289 out of 395 legislators, to be exact), NOMINATE weights its ideal point orientation around these major groups. Since the Southern Democrat cluster mixes being more liberal and more conservative than the Democrats on each distinct issue set, the average comes out towards the middle when the truth is that Southern Democrats represented an extreme group in the U.S. House session. C-BPIRT avoids this issue by modeling groups first, then ideal points. This ensures that the orientation of ideal points is consistent with the group model when groups are an important part of the voting calculus.

[8] See Lee (2009) and Aldrich et al. (2014) for further discussion of why this is a really bad practice.

The 88^{th} U.S. House shows a case where accounting for group similarities provides ideal point estimates that are more consistent with the truth than the independent and identically distributed counterparts. The group dynamics that dominate the voting in this session come to the forefront and provide consistent ideal point estimates that make sense in the historical context. Unlike NOMINATE, C-BPIRT requires no postprocessing or stringent *a priori* assumptions that the legislators belong to specific groups. This aspect of C-BPIRT demonstrates promise as an estimation approach that provides a richer and more consistent view of historical roll call voting for explorations of U.S. legislative voting behavior.

3.6 Conclusion

In this article, I derive and show how to implement an ideal point estimation procedure that allows for Bayesian nonparametric estimation of both issue dimensions and group-consistent ideal points from roll call voting records. I show the importance of this model from both a formal perspective regarding lack of independence in errors and from a historical perspective by analyzing the 107^{th} and the 88^{th} sessions of the U.S. House. I show that C-BPIRT can uncover substantively meaningful group voting behavior that is ignored by NOMINATE and BPIRT. I also show that NOMINATE recovers ideal point orientations that do not align with historical or observational evidence for how voting groups actually chose to vote in each of these sessions. The results provided by C-BPIRT show that meaningful dimensions and groups can be extracted from roll call vote data sets without trading off accuracy for interpretability.

Extensions to the C-BPIRT model exist that can provide even richer inference about group voting in the U.S. Congress. First, C-BPIRT assumes that each session of the U.S. Congress is independent of those that occur before and after. This assumption is generally incorrect since there is, historically, relatively low turnover for members of Congress between sessions. Since voting behavior is unlikely to change session to session, time-varying dynamic models of ideal points will certainly improve model accuracy and provide more consistent insights about legislative voting behavior. Aside from the ideal points, themselves, information can be shared about issue dimensions and groups across Congresses to provide a picture of how Congressional issue sets and voting coalitions evolve over time. Novel advances in tree-based priors in the Bayesian nonparametric literature provide a route for this dynamic model. Second, there are numerous covariates which correlate with group membership apart from raw roll call voting records. Including these covariates in the group estimation procedure would provide opportunities to model relationships between donor records, legislative speech, etc. and voting while also assuming that these actions occur with group incentives in mind. Finally, this model represents an empirical analogue to work in social choice theory about group preference aggregation and outcomes. Given that the U.S. Congress is unique in its combination of group incentives and individual policy incentives due to varied constituent preferences, C-BPIRT can be extended to uncover ideal points that include constituent preferences and estimate the impact of these preferences of roll call voting outcomes.

Bibliography

Adler, E Scott and John Wilkerson (2006). Congressional bills project. *NSF 880066*, 00880061.

Aldrich, John H (1995). *Why parties?: The origin and transformation of political parties in America*. University of Chicago Press.

Aldrich, John H and James S Coleman Battista (2002). Conditional party government in the states. *American Journal of Political Science*, 164–172.

Aldrich, John H , Jacob M Montgomery, and David B Sparks (2014). Polarization and ideology: Partisan sources of low dimensionality in scaled roll call analyses. *Political Analysis 22*(4), 435–456.

Aldrich, John H and David W Rohde (2000). *The logic of conditional party government: Revisiting the electoral connection*. PIPC.

Bafumi, Joseph and Michael C Herron (2010). Leapfrog representation and extremism: A study of american voters and their members in congress. *American Political Science Review 104*(3), 519–542.

Bernhard, William and Tracy Sulkin (2018). *Legislative style*. University of Chicago Press.

Binder, Sarah A (1999). The dynamics of legislative gridlock, 1947–96. *American Political Science Review 93*(3), 519–533.

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

Cameron, Charles M (2000). *Veto bargaining: Presidents and the politics of negative power*. Cambridge University Press.

Carroll, Royce , Jeffrey B Lewis, James Lo, Keith T Poole, and Howard Rosenthal (2009). Measuring bias and uncertainty in dw-nominate ideal point estimates via the parametric bootstrap. *Political Analysis 17*(3), 261–275.

Cattell, Raymond B (1966). The scree test for the number of factors. *Multivariate behavioral research 1*(2), 245–276.

Clinton, Joshua , Simon Jackman, and Douglas Rivers (2004). The statistical analysis of roll call data. *American Political Science Review 98*(2), 355–370.

Cox, Gary W and Mathew D McCubbins (2005). *Setting the agenda: Responsible party government in the US House of Representatives*. Cambridge University Press.

Cox, Gary W and Mathew D McCubbins (2007). *Legislative leviathan: Party government in the House*. Cambridge University Press.

Cox, Gary W and Keith T Poole (2002). On measuring partisanship in roll-call voting: The us house of representatives, 1877-1999. *American Journal of Political Science*, 477–489.

Doshi, Finale , Kurt Miller, Jurgen V Gael, and Yee W Teh (2009). Variational inference for the indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pp. 137–144.

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

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

Ferrari, Diogo (2020). Modeling context-dependent latent effect heterogeneity. *Political Analysis 28*(1), 20–46.

Fiorina, Morris P , Samuel J Abrams, and Jeremy Pope (2006). *Culture war?: The myth of a polarized America*. Longman Publishing Group.

Geweke, John and Guofu Zhou (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies 9*(2), 557–587.

Ghahramani, Zoubin and Thomas L Griffiths (2006). Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pp. 475–482.

Goplerud, Max (2019). A multinomial framework for ideal point estimation. *Political Analysis 27*(1), 69–89.

Gruhl, Jonathan , Elena A Erosheva, and Paul K Crane (2013). A semiparametric approach to mixed outcome latent variable models: Estimating the association between cognition and regional brain volumes. *The Annals of Applied Statistics*, 2361–2383.

Heckman, James J and James M Snyder Jr (1996). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. Technical report, National Bureau of Economic Research.

Jessee, Stephen A (2009). Spatial voting in the 2004 presidential election. *American Political Science Review 103*(1), 59–81.

Jessee, Stephen A (2010). Partisan bias, political information and spatial voting in the 2008 presidential election. *The Journal of Politics 72*(2), 327–340.

Knowles, David and Zoubin Ghahramani (2011). Nonparametric bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 1534–1552.

Kramer, Gerald H (1973). On a class of equilibrium conditions for majority rule. *Econometrica: Journal of the Econometric Society*, 285–297.

Krehbiel, Keith (1992). *Information and legislative organization.* University of Michigan Press.

Krehbiel, Keith (2010). *Pivotal politics: A theory of US lawmaking.* University of Chicago Press.

Krehbiel, Keith , Adam Meirowitz, and Jonathan Woon (2005). Testing theories of lawmaking. In *Social choice and strategic decisions*, pp. 249–268. Springer.

Lee, Frances E (2009). *Beyond ideology: Politics, principles, and partisanship in the US Senate.* University of Chicago Press.

McAlister, Kevin (2018, Sep). *Disagreement and Dimensionality: A Beta Process Approach to Roll Call Scaling for the U.S. Congress.*

McCarty, Nolan , Keith T Poole, and Howard Rosenthal (2016). *Polarized America: The dance of ideology and unequal riches.* mit Press.

Paisley, John and Lawrence Carin (2009). Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 777–784. ACM.

Pitman, Jim et al. (2002). Combinatorial stochastic processes.

Poole, Keith T and Howard Rosenthal (1984). The polarization of american politics. *The Journal of Politics 46*(4), 1061–1079.

Poole, Keith T and Howard Rosenthal (1987). Analysis of congressional coalition patterns: A unidimensional spatial model. *Legislative Studies Quarterly*, 55–75.

Poole, Keith T and Howard Rosenthal (1997). *Congress: A political-economic history of roll call voting.* Oxford University Press on Demand.

Rai, Piyush and Hal Daumé (2009). The infinite hierarchical factor regression model. In *Advances in Neural Information Processing Systems*, pp. 1321–1328.

Ramey, Adam (2016). Vox populi, vox dei? crowdsourced ideal point estimation. *The Journal of Politics 78*(1), 281–295.

Rasmussen, Carl Edward (2000). The infinite gaussian mixture model. In *Advances in neural information processing systems*, pp. 554–560.

Rohde, David W (2010). *Parties and leaders in the postreform House.* University of Chicago Press.

Spirling, Arthur and Kevin Quinn (2010). Identifying intraparty voting blocs in the uk house of commons. *Journal of the American Statistical Association 105*(490), 447–457.

Sulkin, Tracy (2005). *Issue politics in Congress.* Cambridge University Press.

Tausanovitch, Chris and Christopher Warshaw (2017). Estimating candidates' political orientation in a polarized congress. *Political Analysis 25*(2), 167–187.

A Gibbs Sampling for C-BPIRT

Estimation using Gibbs sampling proceeds in the following way:

1. **Sample the latent variable, $\boldsymbol{X}$.** For each $i \in (1, ..., N)$ and $j \in (1, ..., P)$, sample $x_{i,j}$ from a truncated normal distribution according to:

$$x_{i,j} \sim \begin{cases} \mathcal{TN}_{-\infty,0}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{i,j} = 0 \\ \mathcal{TN}_{0,\infty}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{i,j} = 1 \\ \mathcal{N}((\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\omega}_i - \alpha_j, 1) \text{ if } y_{i,j} \text{ is missing} \end{cases} \tag{3.28}$$

2. **Sample R and $\boldsymbol{\Lambda}$ jointly.**[9] Define K^+ as the current number of active features. For each $j \in (1, ..., P)$ and $k \in (1, ..., K^+)$ define:

$$\begin{aligned} t_{j,k} &= \frac{P(r_{j,k} = 1|\boldsymbol{Y}, -)}{P(r_{j,k} = 0|\boldsymbol{Y}, -)} \\ &= \frac{P(\boldsymbol{Y}|r_{j,k} = 1, -)}{P(\boldsymbol{Y}|r_{j,k} = 0, -)} \frac{P(r_{j,k} = 1)}{P(r_{j,k} = 0)} \end{aligned} \tag{3.29}$$

$$\frac{P(\boldsymbol{Y}|r_{j,k} = 1, -)}{P(\boldsymbol{Y}|r_{j,k} = 0, -)} = \sqrt{\frac{\gamma_k}{\gamma}} \exp\left(\frac{1}{2}\gamma\mu^2\right) \tag{3.30}$$

$$\frac{P(r_{j,k} = 1)}{P(r_{j,k} = 0)} = \frac{m_{-j,k}}{p - m_{-j,k} + 1} \tag{3.31}$$

where $\gamma = \boldsymbol{\omega}_k'\boldsymbol{\omega}_k + \gamma_k$, $\mu = \frac{1}{\gamma}\boldsymbol{\omega}_k'\hat{E}_j$, $\hat{E}_j = x_j - (\boldsymbol{r}_j \odot \boldsymbol{\lambda}_j)\boldsymbol{\Omega} + \alpha_j$ setting $\lambda_{j,k} = 0$, and $m_{-j,k} = -r_{j,k} + \sum_{h=1}^{p} r_{h,k}$. Let

$$p_{r=1} = \frac{t_{j,k}}{1 + t_{j,k}}$$

[9] See Paisley and Carin (2009) for more info on the details of this step. For the purposes of roll call scaling, the number of dimensions is expected to be low and much less than 100. As such, I choose to set the initial number of dimensions to 100. This gets the correct number of dimensions in simulations.

then sample $P(r_{j,k}|-) \sim \text{Bern}(p_{r=1})$. If $r_{j,k} = 1$, then sample $P(\lambda_{j,k}|-) \sim \mathcal{N}(\mu, \gamma^{-1})$. Otherwise, set $\lambda_{j,k} = 0$.

3. **Sample $\boldsymbol{\Omega}$.** For each $i \in (1, ..., N)$, let $\boldsymbol{\mu}_i$ be a K^+-column vector including the cluster means for g_i and $\boldsymbol{\Sigma}_i$ a $K^+ \times K^+$ diagonal covariance matrix with the variances for g_i. Sample $\boldsymbol{\omega}_i$ from:

$$\boldsymbol{\omega_i}|- \sim \mathcal{N}_{K^+}(m_i, V_i) \tag{3.32}$$

where:

$$\boldsymbol{V}_i = \left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_i^{-1}\right)^{-1}$$

$$m_i = \boldsymbol{V}_i \left(\Lambda'(\boldsymbol{x}_i + \boldsymbol{\alpha}) + \boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}\right)$$

4. **Remove Inactive Features, Normalize $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$.** For each $k \in (1, ..., K^+)$, if $r_{j,k} = 0 \; \forall \; (1, ..., P)$, remove k from the analysis. Recalculate K^+.

Post-process Λ to normalize the variance. For each $j \in (1, ..., P)$ and $k \in (1, ..., K^+)$ set $\lambda_{j,k}$:

$$\lambda_{j,k} = \frac{\lambda_{j,k}}{\sqrt{1 + \sum\limits_{h=1}^{K^+} \lambda_{j,h}^2}} \tag{3.33}$$

Post process Ω to normalize location and variance. For each $k \in 1 \leq k \leq K^+$, set $\omega_{i,k}$:

$$\omega_k = \frac{\omega_k - \bar{\omega}_k}{sd(\omega_k)} \tag{3.34}$$

5. **Sample Latent Groups, $\boldsymbol{G}$.** At each iteration, let G^+ be the current set of active clusters. Shuffle the order of observations. For each $i \in (1, ..., N)$, find the probability that $\boldsymbol{\omega}_i \in g \; \forall \; (1, ..., G^+)$. For each existing cluster, calculate

$P(g_i = g|\omega_{-i,g})$ where:

$$P(g_i = g|\boldsymbol{g}_{-i}) \propto \int\limits_{\omega_i}\int\limits_{\mu_g}\int\limits_{\Sigma_g} \mathcal{N}_p(x_i\ ;\ (\boldsymbol{\Lambda}\odot\boldsymbol{R})\boldsymbol{\omega}_i-\boldsymbol{\alpha},\boldsymbol{\mathcal{I}}_P)\mathcal{T}_{K^+;2\alpha_g}\left(\mu_g,\frac{\beta_g(\kappa_g+1)}{\alpha_g\kappa_g}\right)d\omega_i \tag{3.35}$$

One method of solution is outlined in 3.4.2.

Determine the probability that observation i belongs in a new cluster:

$$P(g_i = g_{new}) \propto \int\limits_{-\infty}^{\infty} \mathcal{N}_p(x_i\ ;\ (\boldsymbol{\Lambda}\odot\boldsymbol{R})\boldsymbol{\omega}_i-\boldsymbol{\alpha},\boldsymbol{\mathcal{I}}_P)\mathcal{N}_k(\boldsymbol{\omega}_i;0,\boldsymbol{\mathcal{I}}_{K^+})d\omega_i \tag{3.36}$$

This method is also outlined in 3.4.2.

Draw g_i from:

$$P(g_i|-) \sim \text{Cat}(n_h P(g_i = h|\boldsymbol{g}_{-i})\ \forall\ h\ \in\ (1,...,G^+)\ ,\ \beta P(g_i = g_{new})) \tag{3.37}$$

If $g_i = g_{new}$, update G^+ and include the new cluster in proceeding iterations. If there are zero observations in an existing group, update G^+ and remove that cluster from the analysis.

6. **Sample β.** Sample β from:

$$\beta|- \sim \text{Gamma}(G^+, \gamma_c\log(N)) \tag{3.38}$$

where γ_c is Euler's number.

7. **Sample $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ for all clusters.** For each cluster $g \in (1,...,G^+)$, sample the mean, $\boldsymbol{\mu}_g$, and covariance, $\boldsymbol{\Sigma}_g$, for each Gaussian cluster. $\boldsymbol{\Sigma}_g$ is assumed to

be a diagonal covariance matrix. Let $\boldsymbol{\Omega}_g$ be the collection of latent variables that are currently placed in cluster g. For each $k \in (1, ..., K^+)$, draw:

$$\mu_{g,k} \sim \mathcal{T}_{2\alpha_g}\left(\bar{\omega}_{g,k}, \frac{\beta_g}{\alpha_g \kappa_g}\right) \tag{3.39}$$

$$\Sigma_{g,k} \sim \mathrm{IG}(\alpha_g, \beta_g) \tag{3.40}$$

8. **Sample Item Level Intercepts, $\boldsymbol{\alpha}$.** For each $j \in (1, ...p)$, sample the item level intercept from:

$$P(\alpha_j|-) \sim \mathcal{N}\left(\bar{\mu}_j, \frac{1}{N^2}\sum_{i=1}^{N}(\mu_{i,j} - \bar{\mu}_j)^2\right) \tag{3.41}$$

where $\mu_{i,j} = \boldsymbol{\lambda}_j'\boldsymbol{\omega}_i - x_{i,j}$ and $\bar{\mu}_j = \frac{1}{N}\sum\limits_{i=1}^{N}\mu_{i,j}$.

9. **Sample Factor Precisions, γ_k.** For each $k \in 1 \leq k \leq K^+$, sample γ_k from:

$$P(\gamma_k|-) \sim Gamma\left(c + \frac{m_k}{2}, d + \sum_{j=1}^{p}\lambda_{j,k}^2\right) \tag{3.42}$$

where m_k is the number of sources for which feature k is active.

10. **Sample d.** Sample d from:

$$d|- \sim \mathrm{Gamma}\left(c_0 + cK^+, d_0 + \sum_{k=1}^{K^+}\gamma_k\right) \tag{3.43}$$

CHAPTER IV

Interval Estimation on the Marginal Likelihood

4.1 Introduction

Model selection is an important step for statistical modeling, particularly in the social sciences. Often, models are derived from theories that are developed leveraging previous research and other information gathered in the scientific process (Raftery, 1995). The goal of the model selection step is to determine which of a number of possible candidate models best fit the data. While many information criteria and fit statistics exist for this task, they are rife with weaknesses and frequently lead to overfit or underfit models due to improper handling of uncertainty about the locations of model parameters (Raftery, 1995; Gelman et al., 2014).

The Bayesian approach to model comparison presents a model selection criteria that inherently accounts for model complexity due to the introduction of priors - integrating the joint posterior distribution over the priors gives a measure of the model evidence. This metric of model evidence, called the marginal likelihood, represents the probability that the data was generated under the specified model taking into account the prior information. Under relatively accurate priors or with large amounts of data, comparing marginal likelihoods across models accurately selects the correct model in many cases where other fit statistics do not (Zellner, 1971; Fong and

Holmes, 2019). Since many models used in the social sciences have meaningful and widely used Bayesian versions, model comparison via the marginal likelihood is a routine that will accurately distinguish between and select the best models and their corresponding underlying theories.[1]

A problem that arises is that an analytical expression for the marginal likelihood can only be derived for a small number of restrictive examples. While methods exist for taking samples from an posterior distribution when analytic evaluation of the marginal likelihood is intractable, explicit computation of the marginal likelihood is often seen as a secondary task and other methods for assessing overall model fit are used.[2] In spite of the benefits that marginal likelihoods bring to model comparison, the desire to use other criteria stems from the computational burden required to compute the marginal likelihood using Monte Carlo draws from the joint posterior distribution. When the chosen model has a simple sampling scheme, the gold-standard Candidate's estimator can be used to compute the marginal likelihood to an arbitrary desired accuracy (Chib, 1995; Chib and Jeliazkov, 2001). However, the Candidate's estimator is not easily computed for most models of interest.

To overcome the computational problem, several Monte Carlo sampling methods have been proposed to approximate the marginal likelihood (Hammersley and Handscomb, 1964; Raftery and Banfield, 1991; Gamerman and Lopes, 2006; Newton and Raftery,

[1] Bayesian model selection typically falls into two classes: model averaging and best model selection (Zellner, 1971). The model averaging approach uses the marginal likelihood to derive a probability distribution over all candidate models and averages the parameters values using these posterior probabilities. On the other hand, best model selection seeks to find the best model over a set of candidate models - the model with the highest marginal likelihood is selected. In this paper, I present methods for assessing the marginal likelihood with the goal of choosing a single best fit model. However, the marginal likelihood estimation approaches outlined are applicable to the model averaging approach with minor changes.

[2] Samples from a joint posterior distribution can be taken in a variety of ways that have been studied and discussed extensively. See Betancourt (2017) for an excellent overview of Markov Chain Monte Carlo, sequential Monte Carlo, and Hamiltonian Monte Carlo methods. Information criteria are often used as substitutes for the marginal likelihood and assess both within model and predictive ability. See Gelman et al. (2014) for an overview of existing information criteria.

1994; Meng and Wong, 1996; Gronau et al., 2017). Theoretically, Monte Carlo approaches seek to find a point estimate to the marginal likelihood where the asymptotic variance of the estimator disappears as the number of samples taken from the posterior distribution approaches infinity. The bridge sampling estimator of Meng and Wong (1996) and Gronau et al. (2017) represents a state of the art application of this approach. Monte carlo estimators have a general "one-size-fits-all" computational recipe, but require specifying a distribution with known form that approximates the observed joint posterior. While these approaches sometimes compare favorably to the Candidate's estimator, work has shown that point estimates of the marginal likelihood from Monte Carlo approaches suffer when the posterior is not approximately normal and the approximation does not match the high density points of the posterior. This is of particular concern since the quality of the point estimates depend on assumptions of asymptotic consistency and negligible approximation error (Meng and Wong, 1996; Gelman and Meng, 1998). Algorithms have been proposed to normalize posterior draws via warping, but these approaches are often even more computationally expensive than estimation via the Candidate's estimator (Wang and Meng, 2016).

Another approach that attempts to solve the marginal likelihood problem is variational Bayesian inference, which seeks to find an approximation of a convenient form to the true joint posterior that maximizes the evidence lower bound on the marginal likelihood (Jordan et al., 1999; Jordan, 2004). Typically, this approach is used to estimate a posterior using a Bayesian expectation maximization approach, but evidence lower bound is often used as a proxy for the marginal likelihood. However, model comparison using this quantity alone unfairly favors models that have low posterior variance (Chérief-Abdellatif, 2019). Ji et al. (2010) propose a corresponding evidence upper bound that leverages Monte Carlo sampling and demonstrates that the two quantities sandwich the true marginal likelihood. The interval implied by this inequality shares many of the same theoretical underpinnings as the Monte

Carlo approaches to estimating the marginal likelihood (Pradier et al., 2019; Dieng et al., 2017). Unlike Monte Carlo approaches, variational bounds make no assumptions about approximation error, but no work has been done to show how the bounds shrink as a function of approximation quality.

In this paper, I bridge the gap between variational methods of estimating the marginal likelihood and Monte Carlo approaches. I show that the marginal likelihood can be treated as a random variable and a posterior distribution with corresponding intervals can be derived. These intervals share many of the same qualities as the Monte Carlo estimators, but approximation error does not disappear as the number of posterior samples approaches infinity. This allows the width of the intervals to adjust to the quality of the approximation - good approximations have smaller intervals while bad approximations result in less certainty about the location of the marginal likelihood. The concepts of intervals on the marginal likelihood has been presented, but little work has been done to compare the interval estimates to known marginal likelihoods and have mostly been explored in the setting of variational minimization problems (Grosse et al., 2015; Dieng et al., 2017; Pradier et al., 2019).

I derive two sets of intervals. The first set, called JSW bounds, directly leverage the variational inequalities from Ji et al. (2010). In simulations where the true marginal likelihood is known, I show that these inequalities consistently surround the true marginal likelihood while the bridge sampling estimator is often incorrect. However, JSW bounds are wide and cover the marginal likelihood 100% of the time. I show that the interval implied by the evidence upper and lower bounds includes values that cannot exist due to the relationship between forward and reverse Kullback-Leibler divergences. I then propose a new inequality on the ratio of the forward and reverse KL divergences that allows for a finer tuned interval on the marginal likelihood, called kappa bounds. I derive a computationally efficient estimation procedure for these bounds and demonstrate that kappa bounds improve on JSW bounds and the bridge sampling estimator. I show the flexibility and accuracy of kappa bounds in both simulated and real-world data using both linear regression models and more complicated ordered factor analysis models.

4.2 Computing the Marginal Likelihood

For a specified model with parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_K\} \in \boldsymbol{\Theta}$ and prior density, $\pi(\boldsymbol{\theta})$, a common goal in Bayesian inference is to learn a posterior distribution for each of the parameters from observed data, $\mathcal{X} = \{x_1, x_2, ..., x_n\}$. By Bayes rule, the posterior can be defined as:

$$\mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) = \frac{\mathcal{P}(\mathcal{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\mathcal{P}(\mathcal{X})} \tag{4.1}$$

where $\mathcal{P}(\mathcal{X}|\boldsymbol{\theta})$ is the likelihood of the data given the model's parameters. For ease of exposition, $\mathcal{P}(\mathcal{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \mathcal{P}(\mathcal{X}, \boldsymbol{\theta})$ is referred to as the complete data likelihood which captures the combined likelihood of $\boldsymbol{\theta}$ with respect to the data and the prior. While $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ can be difficult to estimate directly, various Monte Carlo methods can be used to take a set of draws from the posterior.[3]

$\mathcal{P}(\mathcal{X})$ is referred to as the marginal likelihood or model evidence and dictates the probability that the data is observed conditional on the model parameters averaged over the prior. The marginal likelihood is defined as:

$$\mathcal{P}(\mathcal{X}) = \int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\mathcal{X}, \boldsymbol{\theta})d\boldsymbol{\theta} = \int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\mathcal{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{4.2}$$

The marginal likelihood can be seen as a measure of model fit and is frequently

[3] Throughout this paper, let $\mathcal{P}()$ refer to the density associated with the true posterior, let $\mathcal{Q}()$ refer to the density associated with an approximation to the posterior, let $\pi()$ be a generic prior density, and let $\mathbb{P}()$ refer to a generic probability function.

used to choose a model that best fits the data across a variety of candidate models. Recent research has also shown that marginal likelihood are a limiting case for leave-p-out and k-fold cross validation, showing that marginal likelihoods can supplant the need for computationally intensive validation checks (Fong and Holmes, 2019). For a number of simple problems, this integral can be evaluated analytically. However, the vast majority of interesting models require integration that is analytically intractable. As such, computational integration methods are used. The following standard approaches are of interest.

4.2.1 Candidate's Estimator

Often considered the gold standard for marginal likelihood computation, the Candidate's estimator (Besag, 1989; Chib, 1995) leverages the fact that:

$$\mathcal{P}(\mathcal{X}) = \frac{\mathcal{P}(\mathcal{X}, \boldsymbol{\theta})}{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})} \tag{4.3}$$

by Bayes' Theorem. Choosing a meaningful set of plugin estimates, $\boldsymbol{\theta}^*$, the Candidate's estimator is:

$$\hat{\mathcal{P}}(\mathcal{X}) = \frac{\mathcal{P}(\mathcal{X}, \boldsymbol{\theta}^*)}{\mathcal{P}(\boldsymbol{\theta}^*|\mathcal{X})} \tag{4.4}$$

For models that are specified with a set of conjugate priors and known posterior forms, this estimate can be evaluated analytically. For any steps that are not easily solved, further Monte Carlo estimates can be drawn allowing for evaluation of any intermediate parameter values as a function of the plugin estimates. For models that are not fully conjugate and use Metropolis-Hastings steps, a similar method is proposed by Chib and Jeliazkov (2001).

The Candidate's estimator is known to produce estimates that are arbitrarily close to the true marginal likelihood in the number of draws taken from the true posterior. While the estimate is accurate, the method is computationally demanding - any quantities that cannot be analytically derived require more simulation following the initial simulations that lead to draws from the true posterior. The amount of added computation increases as a function of the complexity of the model making this approach too costly in most applied situations. However, this method, when possible, can provide an exact estimate of the marginal likelihood.

4.2.2 Naive Monte Carlo Estimator

A simple method for estimating the marginal likelihood that requires no additional significant computation uses a simple Monte Carlo integral approximation (Hammersley and Handscomb, 1964; Raftery and Banfield, 1991). Using the identity:

$$\mathcal{P}(\mathcal{X}) = \int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\mathcal{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_{\pi(\boldsymbol{\theta})}[\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})] \tag{4.5}$$

Monte Carlo integration can be used to assess the marginal likelihood. A naive Monte Carlo estimate of the marginal likelihood is given by:

$$\hat{\mathcal{P}}(\mathcal{X}) = \frac{1}{W} \sum_{w=1}^{W} \mathcal{P}(\mathcal{X}|\theta_w^*) \ ; \ \theta_w^* \sim \pi(\boldsymbol{\theta}) \tag{4.6}$$

Note that no posterior draws are needed to assess the marginal likelihood with this estimator. The naive Monte Carlo estimate works well if the prior and the posterior have a similar shape and have significant overlap. However, the Monte Carlo approach works poorly when the posterior is more concentrated than the initial prior (Gamerman and Lopes, 2006). As the majority of the mass from the prior maps to likelihoods close to zero, the naive Monte Carlo estimate oversamples unimportant regions of the posterior and the estimate of the marginal likelihood is dominated by values close to zero. As such, the naive Monte Carlo estimator works, but requires so many posterior draws, especially with high dimensional posteriors, that getting an accurate estimate is computationally infeasible.

4.2.3 Importance Sampling Estimators

An improved Monte Carlo estimator is presented by Newton and Raftery (1994). The harmonic mean estimator observes that importance sampling methods can improve on the uniform search over the parameter space given by simple Monte Carlo methods. Using a similar identity for the marginal likelihood:

$$\mathcal{P}(\mathcal{X}) = \int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\mathcal{X}|\boldsymbol{\theta})^{-1}\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})d\boldsymbol{\theta} \tag{4.7}$$

the harmonic mean estimator is:

$$\hat{\mathcal{P}}(\mathcal{X})^{-1} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{P}(\mathcal{X}|\theta_i^*)^{-1} \ ; \ \theta_i^* \sim \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \tag{4.8}$$

Newton and Raftery (1994) show that the harmonic mean estimator arises as a consequence of using importance sampling defining the importance distribution as the true posterior. The harmonic mean estimate of the marginal likelihood is widely used as an approximation to the marginal likelihood, but has been shown to perform poorly in many cases where the complete data likelihood is low - the harmonic mean is known to be dominated by relatively low values. This leads to an estimator that can have infinite variance and is, therefore, unsuitable for most applications.

Following the work in Newton and Raftery (1994), the harmonic mean estimator can be improved by more accurately making assessments of the complete data likelihood in the posterior parameter space while also using an importance sampling approach to stabilize the variance of the estimator. In particular, by using an importance sampling density that avoids likelihoods close to zero while also allowing for draws that are not necessarily near the regions of highest mass in the posterior, the variance of the resulting estimator becomes finite and allows for consistent estimation of the marginal likelihood (Pajor, 2016; Neal, 2001).

Let $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ be a distribution over $\boldsymbol{\theta}$ parameterized by $\boldsymbol{\gamma}$ that closely approximates the posterior density. The importance sampling estimator stems from the identity:

$$\mathcal{P}(\mathcal{X}) = \int_{\boldsymbol{\Theta}} \frac{\mathcal{P}(\boldsymbol{\theta}, \mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) d\boldsymbol{\theta} = \mathbb{E}_{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} \left[\frac{\mathcal{P}(\boldsymbol{\theta}, \mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} \right] \tag{4.9}$$

This yields the importance sampling estimator:

$$\hat{\mathcal{P}}(\mathcal{X}) = \sum_{j=1}^{M} \frac{\mathcal{P}(\theta_j^*, \mathcal{X})}{\mathcal{Q}(\theta_j^*|\boldsymbol{\gamma})} \ ; \ \theta_j^* \sim \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \tag{4.10}$$

where M draws are taken from the approximating importance density. Rather than directly performing numerical integration on the posterior draws, the importance density is used to stabilize the complete data likelihood assessments, ensuring that the majority of draws come from regions of the posterior with high density while also allowing assessments outside of the posterior regions of highest mass.

A suitable importance density should (1) be easy to evaluate; (2) have the same domain as the posterior distribution; (3) closely resemble the posterior distribution; and (4) have fatter tails than the posterior distribution (Neal, 2001; Vandekerckhove et al., 2015). The final condition ensures that the estimator is not dominated by the tails of the distribution like the naive Monte Carlo estimator and the harmonic mean estimator (Neal, 2001).

Similar to the improvement on the Monte Carlo estimator by the harmonic mean estimator, a harmonic mean importance density estimator has been proposed (Gelfand and Dey, 1994):

$$\hat{\mathcal{P}}(\mathcal{X}) = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{Q}(\theta_i^*|\boldsymbol{\gamma})}{\mathcal{P}(\theta_i^*, \mathcal{X})} \right)^{-1} \ ; \ \theta_i^* \sim \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \tag{4.11}$$

Like the harmonic mean estimator, the proposal density is the true posterior. However, the improvement comes from assessing the complete data likelihood under the stabilization of an approximating importance density. Like the importance sampling estimator, the harmonic mean importance density estimator performs better than its non-stabilized counterpart. However, there is a requirement that the tails of the importance density are thinner than the posterior distribution (Gelfand and Dey, 1994; Newton and Raftery, 1994; DiCiccio et al., 1997).

4.2.4 Bridge Sampling Estimator

Both the importance sampling estimator and the generalized harmonic mean estimator impose strong constraints on the tail behavior of the importance density relative to the posterior distribution to guarantee a stable estimator. Such requirements can make it difficult to find a suitable importance density, especially when a high-dimensional posterior is considered. The bridge sampler, on the other hand, alleviates such requirements (Fruhwirth-Schnatter, 2004). Originally, bridge sampling was developed to directly estimate the Bayes factor, that is, the ratio of the marginal likelihoods of two models (Kass and Raftery, 1995). However, the bridge sampling estimator can be used to estimate the marginal likelihood of individual models (Overstall and Forster, 2010; Gronau et al., 2017).

The bridge sampling estimator presented by Meng and Wong (1996) leverages the identity:

$$\mathcal{P}(\mathcal{X}) = \frac{\int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\boldsymbol{\theta}, \mathcal{X}) \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\mathbb{E}_{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}\left[\mathcal{P}(\boldsymbol{\theta}, \mathcal{X}) h(\boldsymbol{\theta})\right]}{\mathbb{E}_{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}\left[\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) h(\boldsymbol{\theta})\right]} \tag{4.12}$$

where $h(\boldsymbol{\theta})$ is a *bridge function* that is used to induce desirable properties on the corresponding estimator. This admits an estimator:

$$\hat{\mathcal{P}}(\mathcal{X}) = \frac{\frac{1}{M} \sum\limits_{j=1}^{M} \mathcal{P}(\theta_j^*, \mathcal{X}) h(\theta_j^*)}{\frac{1}{N} \sum\limits_{i=1}^{N} \mathcal{Q}(\theta_i^*|\boldsymbol{\gamma}) h(\theta_i^*)} \; ; \; \theta_i^* \sim \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \; ; \; \theta_j^* \sim \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \tag{4.13}$$

The bridge sampling estimator can be seen as a middle point between the harmonic mean estimator and the importance sampling estimator - as discussed previously, taking taking the expectation with respect to the approximation requires thick tails while the expectation with respect to the posterior requires thin tails. Combining these two elements creates an estimator that minimizes the influence of approximation choice on the resulting estimate.

Meng and Wong (1996) show that there exists an optimal choice for $h(\boldsymbol{\theta})$. Specifically:

$$h(\boldsymbol{\theta}) = \mathcal{C} \cdot \left[\frac{N}{N+M} \mathcal{P}(\boldsymbol{\theta}, \mathcal{X}) + \frac{M}{N+M} \mathcal{P}(\mathcal{X}) \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \right]^{-1} \tag{4.14}$$

where $\mathcal{C}$ is an arbitrary constant. This choice of bridge function is optimal in the sense that it induces an estimator that minimizes the relative mean squared error of all possible estimators with respect to the true marginal likelihood (Meng and Wong, 1996). Note that this proof requires the assumption that bridge sampling estimator is asymptotically unbiased and is independent of the choice of approximation. The optimal bridge function depends on the marginal likelihood which is the very entity we want to approximate. This issue can be resolved by applying an iterative scheme that updates an initial guess of the marginal likelihood until the estimate of the marginal likelihood has converged according to a predefined tolerance level (Gronau et al., 2017).

The bridge sampling estimator has been shown to perform well in a variety of complex applied Bayesian models where the Candidate's estimator cannot be easily computed (Lopes et al., 2003; Gronau et al., 2017). A particular benefit of this approach is that the same computational recipe can be used in many settings:

1. Samples are taken from the true posterior, $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$.

2. Moments of the joint posterior are estimated and a corresponding multivariate normal distribution is specified, $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$.

3. Some large number of samples are taken from $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$.

4. The iterative scheme outlined in (Gronau et al., 2017) is used to estimate the marginal likelihood.

Recent work on the bridge sampling estimator has discussed the role of choice of approximation. While the optimal bridge sampling estimator does not rely on choice of approximation, better approximations lead to better estimates. In high dimensions, multivariate normal approximations might produce unstable estimates in case of high-dimensional posterior distributions that clearly do not follow a multivariate normal distribution. In such a situation, it might be advisable to consider more sophisticated versions of bridge sampling (Fruhwirth-Schnatter, 2004; Meng and Schilling, 2002; Wang and Meng, 2016).

4.2.5 Laplace-Metropolis Estimator

Perhaps the most commonly used approach to estimating the marginal likelihood, the Laplace-Metropolis estimator extends traditional Laplace asymptotics to usage with posterior samples (Tierney and Kadane, 1986). Described in Lewis and Raftery (1997), the Laplace-Metropolis estimator for the marginal likelihood is:

$$\hat{\mathcal{P}}(\mathcal{X}) = (2\pi)^{\frac{K}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\mathcal{P}(\tilde{\boldsymbol{\theta}}, \mathcal{X}) \tag{4.15}$$

where K is the dimensionality of the posterior, $\boldsymbol{\Sigma}$ is an estimate of the inverse Hessian, and $\tilde{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\ \mathcal{P}(\theta, \mathcal{X})$ is the maximum *a posteriori* (MAP) value of $\boldsymbol{\theta}$. Lewis and Raftery (1997) suggests estimating $\boldsymbol{\Sigma}$ as the covariance matrix of the posterior samples and letting $\tilde{\boldsymbol{\theta}}$ be the value from the posterior Monte Carlo draws with the maximum complete data likelihood. More modern approaches to computing the Laplace-Metropolis approximation feed the posterior samples to an iterative quadrature approach that simultaneously estimates the MAP value of $\boldsymbol{\theta}$ and the corresponding Hessian.

The Laplace-Metropolis estimator is an application of the well-known Laplace approximation to integrals of a specific form. However, this estimator is also an example of approximating the marginal likelihood by specifying a known distributional form that approximates the posterior. In this case, the posterior is assumed to be approximately normally distributed with moments that match those displayed by the posterior samples. Frequently, the normal approximation is sufficient - under certain conditions and low prior influence, the posterior converges to a normal distribution (Eberly and Casella, 2003). However, in cases where the posterior is skewed or has fatter tails than the normal distribution, the Laplace-Metropolis estimator will overestimate the fit of the data given the model and reward specifications that induce sharp-peaked posteriors.

4.2.6 Bounds on the Marginal Likelihood via Variational Approaches

Mean-field variational methods, initially developed in statistical physics and extensively studied by machine learning and Bayesian learning communities for deterministic approximation of marginal distributions (MacKay, 1995; Jordan et al., 1999; Jaakkola and Jordan, 2000; Humphreys and Titterington, 2000; Ueda and Ghahramani, 2002; Jordan, 2004), have been implemented in the model selection context (Corduneanu and Bishop, 2001; Beal, 2003). A core identity for variational methods is the evidence lower bound (ELBO):

$$\log \mathcal{P}(\mathcal{X}) \geq \int\limits_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \left[\log P(\boldsymbol{\theta}, \mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})\right] d\boldsymbol{\theta} \tag{4.16}$$

or stated a different way:

$$\log \mathcal{P}(\mathcal{X}) = \int\limits_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \left[\log P(\boldsymbol{\theta}, \mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})\right] d\boldsymbol{\theta} + \int\limits_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \left[\log \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})\right] d\boldsymbol{\theta} \tag{4.17}$$

where $\int\limits_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \left[\log \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})\right] d\boldsymbol{\theta}$ is the Kullback-Leibler divergence from the true posterior, $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$, to an approximate posterior, $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$, $KL(\mathcal{Q}||\mathcal{P})$. By Gibbs' inequality, $KL(\mathcal{Q}||\mathcal{P}) \geq 0$, leading to the equivalence of these two statements.

While the ELBO is most frequently used in variational estimation strategies, Ji et al. (2010) present a similar inequality that brackets the marginal likelihood from above. Specifically:

$$\log \mathcal{P}(\mathcal{X}) \leq \int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \left[\log P(\boldsymbol{\theta}, \mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})\right] d\boldsymbol{\theta} \tag{4.18}$$

or stated a different way:

$$\log \mathcal{P}(\mathcal{X}) = \int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \left[\log P(\boldsymbol{\theta}, \mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})\right] d\boldsymbol{\theta} - \int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \left[\log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) - \log \mathcal{P}(\boldsymbol{\theta}|\mathcal{X})\right] d\boldsymbol{\theta} \tag{4.19}$$

where $\int\limits_{\boldsymbol{\Theta}} \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \left[\log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) - \log \mathcal{P}(\boldsymbol{\theta}|\mathcal{X})\right] d\boldsymbol{\theta}$ is the Kullback-Leibler divergence from the approximate posterior, $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$, to the true posterior, $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$, $KL(\mathcal{P}||\mathcal{Q})$. This upper bound, often referred to as the EUBO, has received significant attention in recent machine learning literature as an alternative approach to fitting variational models on the ELBO (Dieng et al., 2017; Pradier et al., 2019).

Taken together with the ELBO, these two identities sandwich the true log marginal likelihood. Previous work has used these bounds separately as fitting criteria - given a choice of approximating distributions, typically in the exponential family, find the parameters, $\boldsymbol{\gamma}$, that minimize the Kullback-Leibler divergence between the posterior and the approximation. The resulting expectation-maximization fitting scheme is less computationally costly than corresponding Monte Carlo approaches, but there have been significant questions about the accuracy of variational approaches (Wang and Blei, 2019). However, the identities that are key to variational estimation provide a method for bounding the marginal likelihood and, when properly used in conjunction with a method that is guaranteed to sample from the true posterior, can provide powerful bounding equations for model comparison.

4.3 Estimators, Bounds, and Divergence

Aside from the variational bounding method, all of the commonly used methods for estimating the marginal likelihood of a model present a single point estimate. Unlike posterior summary statistics, which are point estimates that encompass the shape and location of the full posterior distribution, the marginal likelihood is an integral that has a single constant value. Hence, in the case where the posterior form is known, the correct constant value of the marginal likelihood can be derived exactly. However, as previously discussed, most interesting models do not admit a simple analytical evaluation of the model evidence.

Posterior approximation is the most common method for achieving reliable estimates of the marginal likelihood. Bridge sampling, in particular, is a cutting-edge approach which has recently seen a resurgence as a one-size-fits-all approach to approximating the marginal likelihood, particularly with a renewed interest in posterior warping to create posterior approximations that better fit non-normal posterior distributions (Gronau et al., 2017; Wang and Meng, 2016). Bridge sampling can be seen as a general importance sampling approach that contains many other marginal likelihood estimation methods as special cases, including the Candidate's estimator (Mira and Nicholls, 2004).

While point estimates of the marginal likelihood can be close to the true marginal likelihood, methods that use a posterior approximation introduce approximation error. In the case of bridge sampling, Meng and Wong (1996) demonstrate that under a set of restrictive assumptions, the bridge sampling estimator is the mean-squared error minimizing estimator of the marginal likelihood. More importantly, Meng and Wong (1996) derive the asymptotic variance associated with the bridge sampling estimator as a function of approximation error:

$$\lim_{N,M\to\infty} \mathbb{V}[\hat{\mathcal{P}}(\mathcal{X})] = \left(\frac{1}{N} + \frac{1}{M}\right)\left((1 - H_A(\mathcal{P},\mathcal{Q}))^{-1} - 1\right) + \mathcal{O}\left(\frac{1}{N+M}\right) \tag{4.20}$$

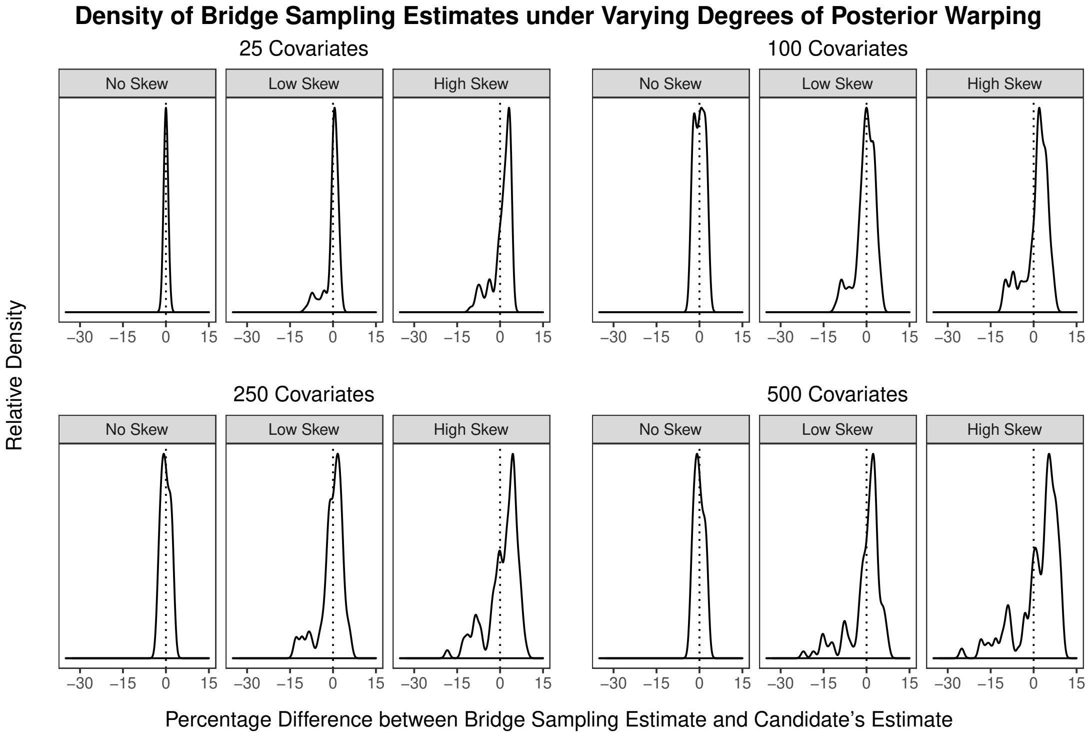
where $H_A(\mathcal{P},\mathcal{Q})$ is the harmonic divergence between the true posterior and the approximation. This equation demonstrates that the approximation error disappears as N and M get large. Similarly, assuming unbiasedness, the bridge sampling estimator is a maximum likelihood estimate of the true marginal likelihood (Wang and Meng, 2016). However, the rate at which influence of divergence between the posterior and approximation disappears as the number of posterior samples increases is ill-defined and these problems are much more apparent in high dimensional posteriors. While work has shown that these assumptions are met when the posterior approximation is essentially the same as the true posterior, it is worthwhile to explore the realistic scenario where the posterior and approximation are not perfectly aligned. Notions of bias are mentioned and a solution is proposed in the form of posterior warping to increase overlap between the approximation and true posterior (Meng and Schilling, 2002; Wang and Meng, 2016). However, these works provide no exploration of how the standard bridge sampling estimator performs under less-than-ideal assumptions. An asymptotically biased estimator of the marginal likelihood is of particular concern since skewness in the posterior can lead to mismatched modes between the posterior and approximation. [4]

In order to explore the accuracy of the bridge sampling estimator, I use a series of simulated data sets for a Bayesian linear regression model with a fixed error variance component. In particular, I seek to explore the performance of the bridge sampling estimator as a function of two attributes - the dimensionality of the posterior and the quality of the approximation to the posterior. Bayesian linear regression is an ideal model for this kind of exploration since the resulting posterior is exactly multivariate normal and the Candidate's estimator exactly computes the marginal likelihood with enough posterior samples. I simulated 500 data sets with 1000 observations and 25, 100, 250, or 500 covariates and produced 10,000 posterior draws from the corresponding linear regression posteriors. For each of these data sets, I computed the Candidate's estimator of the marginal likelihood and the bridge sampling estimator for the corresponding linear regression model under the true posterior. To assess performance with a somewhat inaccurate approximation, I induced positive skewness without changing the area under the true posterior using the warping method presented by Meng and Schilling (2002). Following the work of Gronau et al. (2017), I used multivariate normal approximations fit with moment matching for all posteriors and take 10,000 draws from the approximation. In order to assess the estimator's variance, I replicated the MCMC procedure 100 times for a number of the simulations to get a bootstrap estimate of the variance associated with the bridge sampling estimator.[5]

[4] As a quick point of clarification, the variance for the point estimate of the marginal likelihood is generally very small since the marginal likelihood is bounded between zero and one and generally very close to zero. However, marginal likelihoods on this scale are rarely useful since the difference between values that are close to zero are difficult to detect. For this reason, the logarithm of the marginal likelihood is the preferred quantity. For the rest of this paper, any estimates of the marginal likelihood will be presented and discussed on the natural logarithm scale.

[5] The initial intention was to bootstrap the estimator's variance for all simulations. However, after a few pilot runs, it became increasingly apparent that no more simulations were needed and the results from Meng and Wong (1996) and Wang and Meng (2016) hold.

FIGURE 4.1. Density plots of 500 simulated bridge sampling estimates compared to the true marginal likelihood.



Note: Dotted lines show the mean of the 500 estimates. Skewness is induced using posterior warping techniques from Meng and Schilling (2002).

In line with the findings of Meng and Wong (1996), the simulations show that the bootstrap variance quickly shrinks to zero when the number of draws taken from the true posterior and the approximation is large. Asymptotically incorrect estimates, on the other hand, appear to be a problem. Figure 4.1 shows the results of these simulations and assesses the incorrectness of the estimator across all simulated models. In the case where the posterior and approximation are both essentially multivariate normal, the bridge sampling estimator performs very well - the bridge sampling estimator is generally accurate and any errors are generally quite small. Things are not quite as neat when the approximation to the posterior is further from the truth. When the posterior has low amounts of skew, the modal bridge sampling estimator is generally close to the true value. However, the distribution of values is not symmetric;

the distribution is left-skewed and the bulk of estimates overestimate the true value of the marginal likelihood. When the posterior exhibits high amounts of skew, the same pattern is seen, but the modal estimate is no longer close to the true marginal likelihood and the left tail is even more extreme. These effects become even more pronounced when the dimensionality of the posterior increases. Interestingly, across all simulations, the mean of the estimates is quite close to zero. This implies that over a large number of models, the bridge sampling estimator is unbiased. Individual runs of the bridge sampling procedure, though, lead to estimates that are asymptotically incorrect due to approximation error.

These simulations show that the relative divergence between the approximation and the true posterior has an influence on the quality of the bridge sampling estimate of the marginal likelihood. In the typical high-dimensional case where the approximation is chosen to be a multivariate elliptical distribution, moment matching is used to find an approximation which matches the posterior draws. However, moment matching leads to a mismatch of high density regions when the true posterior is skewed in a non-elliptical way. By choosing an approximation that incorrectly assigns mass to low-density regions of the posterior, the bridge sampling estimator can vary widely and, potentially, lead to incorrect inference about the relative quality of multiple models.

One approach which seeks to address these issues is to allow for more flexible methods for choosing a posterior approximation. Choosing an approximation which matches all of the quirks of the posterior will almost certainly lead to a bridge sampling estimate that is very close to the truth. To this end, posterior warping methods that seek to augment the posterior without changing the normalizing constant provide an approach to creating better approximations and, in turn, better estimates of the marginal likelihood (Meng and Schilling, 2002; Wang and Meng, 2016). While this method corrects for skewness, multimodality, and other irregularities in the posterior, it still has a few weaknesses. First, posterior warping relies on automatic differentiation to get estimates of the Jacobian and Hessian associated with the posterior. In high-dimensional posteriors, this leads to significant computational strain that makes the numerical calculation prohibitively difficult for most standard workstations. Second, posterior warping works best when the form of the posterior is close to elliptical. For many models where skewness is expected (e.g. partial identification of parameters, strong prior influence, parameter values close to bounds on the domain, small sample size, etc.) or the sampling scheme does not guarantee draws that accurately represent a smooth posterior (e.g. discrete posterior margins, Metropolis-Hastings draws with poor mixing, etc.), there is no reason to expect that the posterior will have a smooth, elliptical form. Finally, and perhaps most importantly, the method still only produces a single point estimate that does not encode any sense of accuracy or uncertainty.

4.4 An Interval Estimate for the Marginal Likelihood

Addressing distributional complexity in a point estimator is a difficult problem. While a first instinct is to place confidence intervals on the bridge sampling estimator, the previous simulations show that the variance of the estimator quickly shrinks to zero, even when the estimate is incorrect. From the simulations, it seems that the magnitude of the incorrectness grows as a function of the quality of the approximation. Thus, it seems prudent to attempt to build an interval that takes this aspect of the problem into account.

Variational methods for Bayesian inference seek to fit models by creating an approximation of a known form to a model's posterior that minimizes the difference between the true posterior and the approximation. At the core of this approach are identities that bound the marginal likelihood of the model and approximations can be chosen that minimize the divergence between the posterior and approximation. For the purposes of defining an interval estimate for the marginal likelihood, these variational identities provide a framework for creating intervals where poor approximations result in larger interval widths.

I begin with the core identities of variational Bayesian inference: the evidence lower bound, $\mathcal{L}$, and evidence upper bound, $\mathcal{U}$ (Ji et al., 2010; Jordan et al., 1999).

Definition 4.4.1. Let $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ be the true posterior distribution conditional on some data, $\mathcal{X}$. Let $\mathcal{P}(\boldsymbol{\theta}, \mathcal{X})$ be the complete data likelihood of $\mathcal{X}$ given some prior on $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$. As with importance sampling methods, let $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ be any proper density function that shares the same support as the true posterior parameterized by $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, ..., \gamma_k\} \in \boldsymbol{\Gamma}$. Then:

$$\mathcal{L} + KL(\mathcal{Q}||\mathcal{P}) = \log \mathcal{P}(\mathcal{X}) = \mathcal{U} - KL(\mathcal{P}||\mathcal{Q}) \tag{4.21}$$

where:

$$\begin{aligned}
\mathcal{L} &= \int_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta}|\gamma)[\log \mathcal{P}(\boldsymbol{\theta}, \mathcal{X}) - \log Q(\boldsymbol{\theta}|\boldsymbol{\gamma})]d\boldsymbol{\theta} \\
\mathcal{U} &= \int_{\boldsymbol{\Theta}} \mathcal{P}(\boldsymbol{\theta}|\mathcal{X})[\log \mathcal{P}(\boldsymbol{\theta}, \mathcal{X}) - \log Q(\boldsymbol{\theta}|\boldsymbol{\gamma})]d\boldsymbol{\theta} \\
KL(\mathcal{Q}||\mathcal{P}) &= \int_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})[\log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) - \log \mathcal{P}(\boldsymbol{\theta}|\mathcal{X})]d\boldsymbol{\theta} \\
KL(\mathcal{P}||\mathcal{Q}) &= \int_{\boldsymbol{\Theta}} \mathcal{P}(\boldsymbol{\theta}|\mathcal{X})[\log \mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})]d\boldsymbol{\theta}
\end{aligned} \tag{4.22}$$

$KL(\mathcal{P}||\mathcal{Q})$ is the Kullback-Leibler divergence from the approximation $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ to $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$, or the expected density ratio of $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ to $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ taken with respect to the true posterior. As the KL divergence is not a symmetric measure of distance, the reverse divergence, $KL(\mathcal{Q}||\mathcal{P})$, is a separate and unequal measure of divergence.

Directly applying Gibbs' inequality, Definition 4.4.1 shows that $\mathcal{L} \leq \log \mathcal{P}(\mathcal{X}) \leq \mathcal{U}$. This result is key for the usage of variational methods.

These identities are typically used to fit models with respect to one of the conditions. Using the lower bounds, various mean field approximation can be used to derive the approximation that minimizes $KL(\mathcal{Q}||\mathcal{P})$ from the true posterior without ever needing to take draws from the true posterior (Jordan et al., 1999; Jordan, 2004). While this procedure can be much quicker than standard Monte Carlo techniques, there are questions about the inferential properties of variational posteriors and the inherent model selection procedure which is central to variational methods (Zhang et al., 2018).

Addressing some of these concerns, Ji et al. (2010) discuss model fitting using the upper bound and a pseudo-lower bound. While standard variational methods seek to sit an approximation that minimizes $KL(\mathcal{Q}||\mathcal{P})$, Ji et al. (2010) propose first running a Monte Carlo simulation to produce draws from the true posterior, then using these draws to derive an approximation which allows for a sandwiching interval that encompasses the true marginal likelihood. While Ji et al. (2010) use these identities to create schemes for deriving approximate posterior distributions that minimize the evidence upper bound, I use the identities derived in this paper as a starting point for a set of intervals on the model evidence.

Definition 4.4.2. For a specified model with parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_k\} \in \boldsymbol{\Theta}$ and prior density, $\pi(\boldsymbol{\theta})$, assume that N iid draws from $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$, $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, ..., \theta_N^*) \in \boldsymbol{\Theta}$ are taken. Given $\boldsymbol{\theta}^*$, an approximate posterior, $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ is fit and M independent and identically distributed draws are taken, $\boldsymbol{\omega}^* = (\omega_1^*, \omega_1^*, ..., \omega_M^*) \in \boldsymbol{\Theta}$. For all $\boldsymbol{\theta}^*$ and $\boldsymbol{\omega}^*$, assume that $\mathcal{P}(\theta_i^*, \mathcal{X})$ and $\mathcal{P}(\omega_j^*, \mathcal{X})$ can be assessed. Similarly, assume that $\mathcal{Q}(\theta^*|\boldsymbol{\gamma})$ and $\mathcal{Q}(\omega^*|\boldsymbol{\gamma})$ can be assessed.

Define estimators of $\mathcal{U}$ and $\mathcal{L}$ as:

$$\begin{aligned}\mathcal{U}_0 &= \frac{1}{N}\sum_{i=1}^{N}[\log \mathcal{P}(\theta_i^*, \mathcal{X}) - \log \mathcal{Q}(\theta_i^*|\boldsymbol{\gamma})] \\ \mathcal{L}_0 &= \frac{1}{M}\sum_{j=1}^{M}[\log \mathcal{P}(\omega_j^*, \mathcal{X}) - \log \mathcal{Q}(\omega_j^*|\boldsymbol{\gamma})]\end{aligned} \tag{4.23}$$

respectively. By the law of large numbers, these two estimators converge almost surely to the evidence upper bound and evidence lower bound, respectively. Note that these two estimators arise as solutions from Monte Carlo integration of $\mathcal{U}$ and $\mathcal{L}$ - recognizing that $\mathcal{U}$ and $\mathcal{L}$ are expectations with respect to the posterior and approximations, respectively, Monte Carlo integration can performed with $\boldsymbol{\theta}^*$ and $\boldsymbol{\omega}^*$.

Appealing to central limit theorem:

$$\begin{aligned}\mathcal{U}_0 &\xrightarrow{d} \mathcal{N}\left(\mathcal{U}, \frac{\sigma_{\mathcal{U}}^2}{N}\right) \\ \mathcal{L}_0 &\xrightarrow{d} \mathcal{N}\left(\mathcal{L}, \frac{\sigma_{\mathcal{L}}^2}{M}\right)\end{aligned}$$

Assuming flat priors on $\mathcal{U}$ and $\mathcal{L}$:

$$\begin{aligned}\mathbb{P}(\mathcal{U}|\boldsymbol{\theta}^*) &\sim \mathcal{N}\left(\mathcal{U}_0, \sigma_{\mathcal{U}}^2\right) \\ \mathbb{P}(\mathcal{L}|\boldsymbol{\omega}^*) &\sim \mathcal{N}\left(\mathcal{L}_0, \sigma_{\mathcal{L}}^2\right)\end{aligned} \tag{4.24}$$

where $\sigma_{\mathcal{U}}^2 = \frac{\hat{\sigma}_{\mathcal{U}}^2}{N}$, $\sigma_{\mathcal{L}}^2 = \frac{\hat{\sigma}_{\mathcal{L}}^2}{M}$, and $\hat{\sigma}_{\mathcal{U}}^2$ and $\hat{\sigma}_{\mathcal{L}}^2$ are unbiased estimators of the sample variance of $\log \mathcal{P}(\boldsymbol{\theta}^*, \mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\theta}^*|\boldsymbol{\gamma})$ and $\log \mathcal{P}(\boldsymbol{\omega}^*, \mathcal{X}) - \log \mathcal{Q}(\boldsymbol{\omega}^*|\boldsymbol{\gamma})$, respectively.[6]

Definition 4.4.2 shows that evidence upper and lower bounds can be expressed as random variables with moments that can be modeled from samples of the true posterior

[6] $\mathcal{N}(\mu, \sigma^2)$ is the normal density parameterized by mean μ and variance σ^2.

and samples from the chosen approximation. With a light assumption, this can be extended to treat the marginal likelihood as a random variable, itself.

Definition 4.4.3. Let $\Delta = \log \mathcal{P}(\mathcal{X})$. Given values for the evidence upper and lower bounds and assuming that the marginal likelihood between the bounds follows a four-parameter beta distribution, a conditional posterior for the marginal likelihood can be derived:

$$\mathbb{P}(\Delta|a, b, \mathcal{U}, \mathcal{L}) = \frac{(\Delta - \mathcal{L})^{a-1} (\mathcal{U} - \Delta)^{b-1}}{\boldsymbol{B}(a, b) (\mathcal{U} - \mathcal{L})^{a+b-1}} \tag{4.25}$$

where a and b are shape parameters as in the standard beta distribution and $\boldsymbol{B}(a, b)$ is the standard Beta function with inputs a and b.

Integrating over all possible values of $\mathcal{U}$ and $\mathcal{L}$, the marginal posterior takes the form:

$$\mathbb{P}(\Delta|a, b, \boldsymbol{\theta}^*, \boldsymbol{\omega}^*) = \int \int \frac{(\Delta - \mathcal{L})^{a-1} (\mathcal{U} - \Delta)^{b-1}}{\boldsymbol{B}(a, b) (\mathcal{U} - \mathcal{L})^{a+b-1}} \mathbb{P}(\mathcal{U}|\boldsymbol{\theta}^*)\mathbb{P}(\mathcal{L}|\boldsymbol{\omega}^*) d\mathcal{U} d\mathcal{L} \tag{4.26}$$

Definition 4.4.3 establishes a random variable representation of the marginal likelihood. Unlike point estimators which produce a single value for the marginal likelihood, treating the marginal likelihood as a random variable allows for incorporation of approximation error into the computed value and, in turn, credible intervals that should surround the true marginal likelihood. While the posterior for the marginal likelihood does not have a convenient, closed-form representation, asymptotic moments can be examined to give intuition about the role that approximation error plays in the posterior.

Lemma 4.4.1. Let $\alpha \sim \text{Beta}(a, b)$ be a beta distributed random variable that ranges between zero and one. The asymptotic expectation and variance of $\mathbb{P}(\Delta|a, b, \boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$

can be defined as:

$$\begin{aligned} \mathbb{E}[\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*] &= \mathbb{E}[(1-\alpha)]\mathcal{U}_0 + \mathbb{E}[\alpha]\mathcal{L}_0 \\ \mathbb{V}[\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*] &= (\mathcal{U}_0-\mathcal{L}_0)^2\mathbb{V}[\alpha] \end{aligned} \tag{4.27}$$

where $\mathbb{E}[\alpha] = \frac{a}{a+b}$, $\mathbb{E}[1-\alpha] = \frac{b}{a+b}$, and $\mathbb{V}[\alpha] = \frac{ab}{(a+b)^2(a+b+1)}$.

Proof. The proof for this statement can be found in Appendix A.

The posterior variance for the marginal likelihood random variable illuminates how this approach differs from the bridge sampling approach. While the components of the estimators are the same, beginning with Equation 4.22 and deriving bounds incorporates the divergence between the approximation and the posterior in the variance of the resulting distribution. Similarly, this approach leads to a posterior variance which can be easily computed from the sample, which is an important step when deriving intervals. Equation 4.20 shows that the bridge sampling estimator has a variance which shrinks to zero as the number of draws from the true posterior and approximation get large, ignoring the divergence term included in the variance. Since there is evidence that the bridge sampling estimator is asymptotically incorrect, this induces poor behavior of the resulting estimator when the approximation does not perfectly fit the true posterior. The bounds derived from the identities in Ji et al. (2010) (JSW Bounds), move the divergence term outside such that the variance does not approach zero asymptotically. This aspect of JSW bounds allows the interval to hedge against uncertainty due to approximation error.

4.4.1 Estimation and Performance of JSW Bounds

Much like the bridge sampling scheme, estimation of JSW bounds follows a simple computational scheme:

1. Draw N posterior samples from the true posterior distribution, $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$, $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, ..., \theta_N^*)$.

2. Fit an approximation to the posterior distribution.

3. Draws M samples from the approximation, $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$, $\boldsymbol{\omega}^* = (\omega_1^*, \omega_1^*, ..., \omega_M^*)$.

4. Assess $\mathcal{P}(\boldsymbol{\theta}^*, \mathcal{X})$, $\mathcal{P}(\boldsymbol{\omega}^*, \mathcal{X})$, $\mathcal{Q}(\boldsymbol{\theta}^*|\boldsymbol{\gamma})$, and $\mathcal{Q}(\boldsymbol{\omega}^*|\boldsymbol{\gamma})$.

5. Compute $\mathcal{U}_0$, $\mathcal{L}_0$, $\sigma^2_{\mathcal{U}}$, and $\sigma^2_{\mathcal{L}}$.

6. Given choices for the prior hyperparameters, a and b, use a bootstrap scheme to get B Monte Carlo draws from the posterior on the marginal likelihood:

$$\begin{aligned} u_b &\sim \mathcal{N}(\mathcal{U}_0, \sigma^2_{\mathcal{U}}) \\ l_b &\sim \mathcal{N}(\mathcal{L}_0, \sigma^2_{\mathcal{L}}) \\ \alpha_b &\sim \text{Beta}(a, b) \\ \Delta_b^* &= u_b - \alpha_b(u_b - l_b) \end{aligned}$$

7. Compute highest posterior density intervals for Δ^* at the desired width.

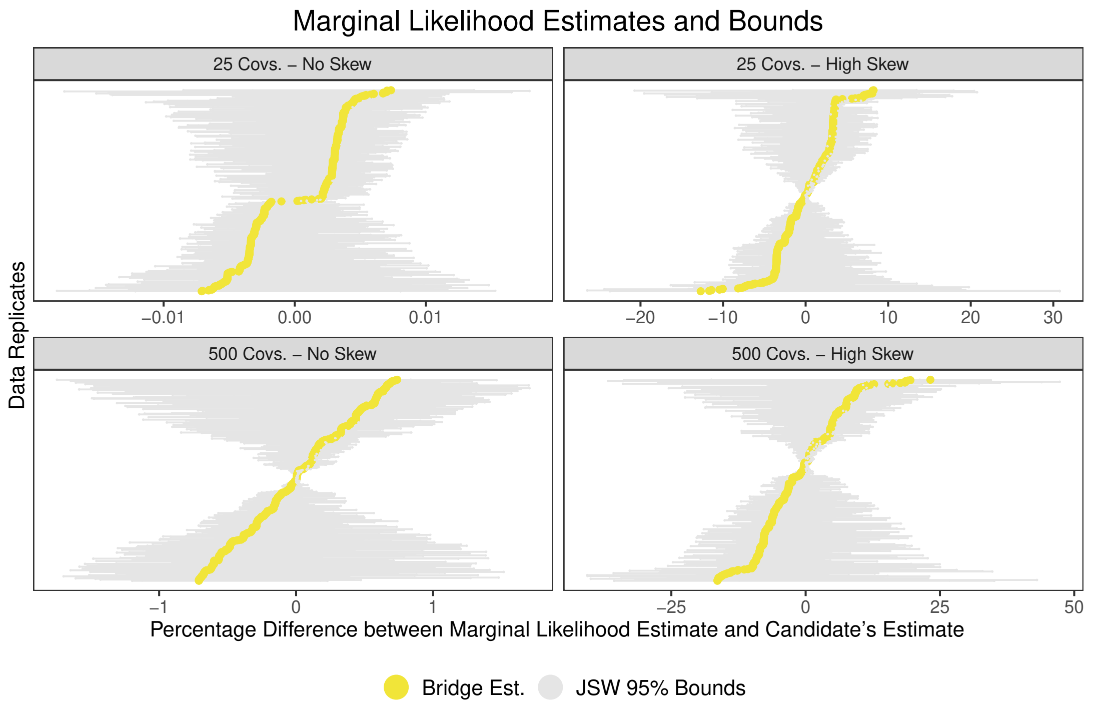
While the posterior variance of the marginal likelihood is known, its form is not. Under extremely restrictive assumptions, symmetry around the optimal point estimator can be imposed. However, initial work shows that the resulting posterior is not generally symmetric. As such, JSW bounds use a Monte Carlo simulation framework to simulate draws from the resulting posterior and compute credible intervals.

One point worth noting is that JSW bounds, like most marginal likelihood estimation approaches, assume that N independent and identically distributed draws are taken from the posterior. While some Monte Carlo approaches explicitly correct draws to ensure that draws are iid, the majority of simulation strategies suffer from autocorrelation across draws, leading to a violation of the assumption. Without correction, it should be expected that any empirical results will underestimate any variances of interest. For this reason, I follow the suggestions of Gronau et al. (2017) and set N equal to the effective sample size of the complete data likelihood of the posterior draws (Geyer, 2011). This approach shrinks the sample size in accordance with the level of autocorrelation among draws and provides a reasonable approximation to an equivalent number of independent and identically distributed draws achieved by the Monte Carlo procedure.

To assess the performance of JSW bounds, I use the same simulations done previously. As before, all approximations are multivariate normal with moments chosen by matching the moments of the joint posterior. For each linear regression data set with differing numbers of covariates and levels of posterior skewness, I compute the true marginal likelihood using the Candidate's estimator, the bridge sampling estimator, and 95% JSW intervals with a $\text{Beta}(0, 0)$ prior (to induce the maximum width on the JSW interval estimates). Specifically, I explore the coverage of the JSW intervals. Coverage explores two of the biggest concerns with JSW intervals. First, coverage demonstrates how frequently JSW intervals include the true marginal likelihood value. Second, coverage demonstrates the specificity of JSW intervals. While the true value may be included in the interval, short intervals are preferred to wide ones and over-encompassing intervals make model comparison extremely difficult. It is also worth exploring how the chosen level for the JSW interval influences the true probability of including the true value.

Figure 4.2 compares the bridge sampling estimator and the corresponding JSW intervals to the true marginal likelihood. A first observation is that the bridge sampling estimator and JSW bounds are generally computing similar values. In fact, over all

FIGURE 4.2. Bridge sampling estimator and JSW intervals for 500 randomly generated linear regression data sets with varying numbers of covariates and amounts of posterior skew.



Take note of the extremely different scales on each x–axis!

Note: Data are generated from a linear regression model with a known error variance. Skewness is induced using posterior warping techniques from Meng and Schilling (2002). 95% HDP intervals are shown for JSW intervals. Percentages are computed with respect to the Candidate's estimator to the marginal likelihood for each model. Bayesian linear regression posterior samples and Candidate's estimator were computed using `MCMCregress` from `MCMCpack` in `R` (Martin et al., 2011).

simulations, the JSW bounds include the bridge sampling estimator approximately 90% of the time. In the case of low skew, the intervals almost always include the bridge sampling estimator while high skew posteriors lead to fewer intervals including the estimate. The most stark conclusion from these simulations is that the 95% JSW bounds always include the true value of the marginal likelihood. While 100% coverage is not always a bad thing, it comes at cost of specificity in JSW Bounds. When there is no skew in the posterior, the width of the JSW bounds are generally small. On the other hand, skewness in the posterior frequently leads JSW bounds to be quite wide, often on the order of 20% above and below the true value. This width leads to problems when comparing models - if the goal is to use posterior intervals to say that one model fits the data better than another, wide bounds make it more difficult to distinguish between models.

4.5 Refinements to the JSW Interval

JSW bounds have widths that grow in accordance with the distance between the approximation and the true posterior. Lemma 4.4.1 shows that the posterior variance that dictates the interval length grows as the distance between the evidence upper bound and evidence lower bound increases. Using the key variational identities, this distance can be shown to be a direct consequence of the forward and reverse KL divergence between the true posterior and approximation:

$$\mathcal{U} - \mathcal{L} = KL(\mathcal{P}||\mathcal{Q}) + KL(\mathcal{Q}||\mathcal{P}) \ ; \ \mathcal{L} \leq \log \mathcal{P}(\mathcal{X}) \leq \mathcal{U} \tag{4.28}$$

While this identity defines the range of possible values, it gives no information about where in the interval the true marginal likelihood exists. To address this concern, JSW bounds treat the marginal likelihood as a random variable. With known $\mathcal{U}$ and $\mathcal{L}$, Definition 4.4.2 formalizes this random variable as:

$$\begin{gathered} \log \mathcal{P}(\mathcal{X}) = \mathcal{U} - \alpha(\mathcal{U} - \mathcal{L}) \\ \alpha \sim \text{Beta}(a, b) \\ \alpha(\mathcal{U} - \mathcal{L}) = KL(\mathcal{P}||\mathcal{Q}) \ ; \ (1-\alpha)(\mathcal{U} - \mathcal{L}) = KL(\mathcal{Q}||\mathcal{P}) \end{gathered} \tag{4.29}$$

$KL(\mathcal{P}||\mathcal{Q}) + KL(\mathcal{Q}||\mathcal{P})$, often referred to as Jeffreys' divergence, is a symmetric measure of distance that encodes the KL divergence in both the forward and reverse directions. As defined above, α encodes the proportion of the Jeffreys' divergence that can be attributed to $KL(\mathcal{P}||\mathcal{Q})$. This construction then admits a set of possible values ranging from $\mathcal{L}$ to $\mathcal{U}$.

While the variational identities dictate that Equation 4.28 is the minimum width bound on the marginal likelihood, the bound edges represent a a combination of forward and reverse KL divergences that cannot, by definition, occur together. To see this, assume that $KL(\mathcal{Q}||\mathcal{P}) = \mathcal{U} - \mathcal{L}$ and $KL(\mathcal{P}||\mathcal{Q}) = 0$. If $KL(\mathcal{P}||\mathcal{Q}) = 0$, then $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) = \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \; \forall \; \theta \; \in \; \boldsymbol{\Theta}$ and $KL(\mathcal{Q}||\mathcal{P}) = 0$, which contradicts the initial conditions. Logically, this notion extends to values close to zero - $KL(\mathcal{P}||\mathcal{Q}) = \epsilon$ and $KL(\mathcal{Q}||\mathcal{P})$ much larger than ϵ occurs occurs when $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ is small for all $\theta \in \boldsymbol{\Theta}$ while there are regions of high density in $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$. If the goal of the approximation is to best match the true posterior, then these cases should arise infrequently. As such, it makes sense that there is a relationship between $KL(\mathcal{P}||\mathcal{Q})$ and $KL(\mathcal{Q}||\mathcal{P})$ - knowing $KL(\mathcal{P}||\mathcal{Q})$ should reasonably bound $KL(\mathcal{Q}||\mathcal{P})$.

This notion can be formalized by exploring the ratio of the two KL divergences, $\frac{KL(\mathcal{P}||\mathcal{Q})}{KL(\mathcal{Q}||\mathcal{P})}$. Specifically, I show that this ratio is bounded due to the fact that it is a ratio of two f-divergences:

Theorem 4.5.1. Let $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ and $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ be two proper density functions that share common support (i.e. $\mathcal{P}(\boldsymbol{\theta} = \theta|\mathcal{X}) = 0 \iff \mathcal{Q}(\boldsymbol{\theta} = \theta|\boldsymbol{\gamma}) = 0$). Further assume that $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ is absolutely continuous with respect to $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ and vice versa, $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \ll\gg \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$. Then the ratio of the forward and reverse KL divergences is bounded such that:

$$\kappa(\beta_2) \leq \frac{KL(\mathcal{P}||\mathcal{Q})}{KL(\mathcal{Q}||\mathcal{P})} \leq \kappa(\beta_1) \tag{4.30}$$

where $\beta_1 = \sup\limits_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$, $\beta_2 = \inf\limits_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$, and $\kappa(t) = \frac{1+t\log t - t}{t - \log t - 1}$.

Proof. The proof for this theorem can be found in Appendix B.

4.5.1 Kappa Bounds on the Marginal Likelihood

Using Theorem 4.5.1, I propose a refined set of bounds, called kappa bounds, on the marginal likelihood:

Theorem 4.5.2. Let $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ be a proper probability density function with marginal likelihood $\mathcal{P}(\mathcal{X})$. Let $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ be an approximation to $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ that shares the same domain. Then:

$$\mathcal{L} + \frac{(\mathcal{U} - \mathcal{L})}{1 + \kappa(\beta_1)} \leq \log P(\mathcal{X}) \leq \mathcal{U} - \frac{\kappa(\beta_2)(\mathcal{U} - \mathcal{L})}{1 + \kappa(\beta_2)} \tag{4.31}$$

Proof. Given $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ and $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$, recall that:

$$\mathcal{U} - KL(\mathcal{P}||\mathcal{Q}) = \mathcal{L} + KL(\mathcal{Q}||\mathcal{P}) = \log \mathcal{P}(\mathcal{X})$$

This can be rearranged such that:

$$\mathcal{U} - \alpha(\mathcal{U} - \mathcal{L}) = \log \mathcal{P}(\mathcal{X}) = \mathcal{L} + (1 - \alpha)(\mathcal{U} - \mathcal{L})$$

where:

$$\alpha = \frac{KL(\mathcal{P}||\mathcal{Q})}{KL(\mathcal{P}||\mathcal{Q}) + KL(\mathcal{Q}||\mathcal{P})}$$

Define ρ as:

$$\rho = \frac{KL(\mathcal{P}||\mathcal{Q})}{KL(\mathcal{Q}||\mathcal{P})}$$

Then α can be rewritten as:

$$\alpha = \frac{\rho}{1 + \rho}$$

By Theorem B.1 and letting $t = \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$, ρ is bounded such that:

$$\kappa(\beta_2) \leq \rho \leq \kappa(\beta_1)$$

Applying to α:

$$\frac{\kappa(\beta_2)}{1+\kappa(\beta_2)} \leq \alpha \leq \frac{\kappa(\beta_1)}{1+\kappa(\beta_1)}$$

Which implies a maximum width bound on $\log \mathcal{P}(\mathcal{X})$:

$$\mathcal{L} + \frac{(\mathcal{U}-\mathcal{L})}{1+\kappa(\beta_1)} \leq \log \mathcal{P}(\mathcal{X}) \leq \mathcal{U} - \frac{\kappa(\beta_2)(\mathcal{U}-\mathcal{L})}{1+\kappa(\beta_2)}$$

since $\beta_2 < 1$ and $\beta_1 > 1$.

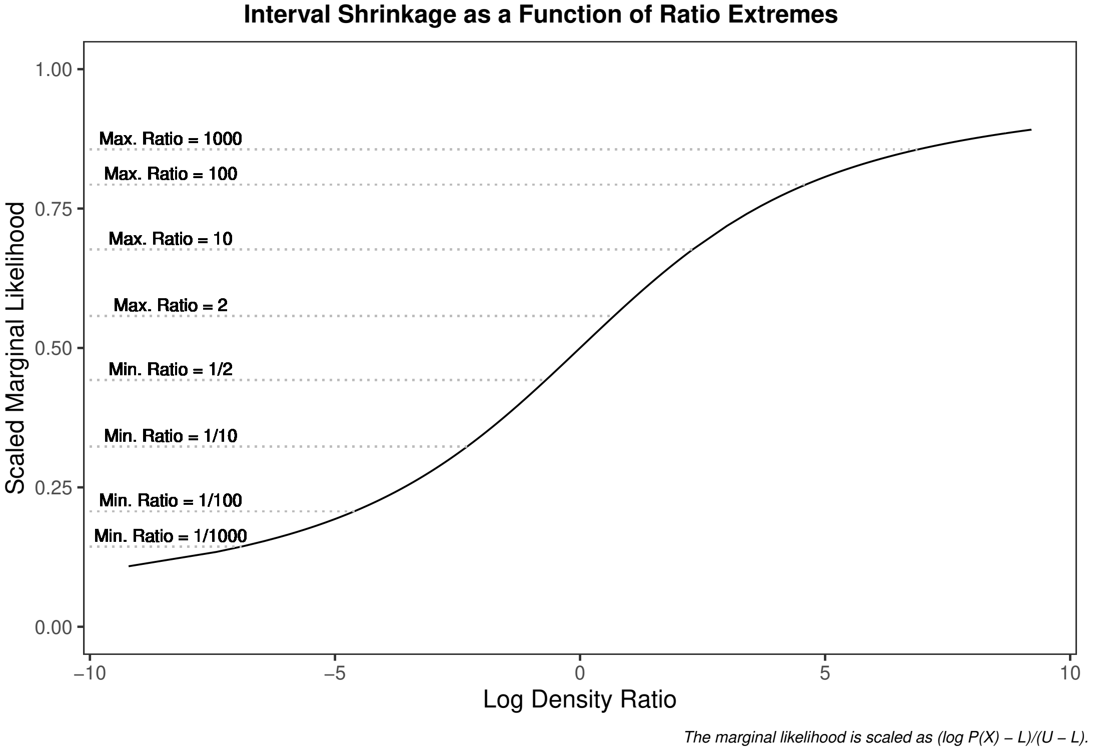
Theorem 4.5.2 demonstrates that JSW bounds often include values that are not possible given the relationship between $KL(\mathcal{P}||\mathcal{Q})$ and $KL(\mathcal{Q}||\mathcal{P})$. While JSW bounds rely wholly on the expectation of the density ratio to create bounds, kappa bounds leverage both the expectation and extrema conditions to further shrink the bounds on the marginal likelihood. In words, kappa bounds leverage the fact that if the true posterior and approximation are never that far apart, then the forward and reverse KL divergences cannot be that far apart. In a case where the approximation is carefully chosen to match moments and areas of high mass in the posterior, the approximation should never have areas of mass that do not match areas of mass in the posterior. Even when the choice of approximation is not perfect, there are significant portions of JSW bounds that can be ruled out as inadmissible.

The improvement in bound width given by kappa bounds compared to JSW bounds is:

$$1 - \frac{\kappa(\beta_2)}{1+\kappa(\beta_2)} - \frac{1}{1+\kappa(\beta_1)} \tag{4.32}$$

Improvement is guaranteed when the approximation and posterior share common support - the width of kappa bounds is equal to JSW bounds only when the maximum density ratio is infinite and the minimum density ratio is zero. Because of the logarithmic scale of $\kappa(t)$, kappa bounds provide a significant improvement over JSW bounds in all but the most extreme cases.

FIGURE 4.3. Shrinkage provided by Kappa Bounds compared to JSW Bounds.



Note: Interval shrinkage is a function of $\min_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$ and $\max_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$. Values closer to 1 yield smaller intervals.

Figure 4.3 shows the proportional improvement over JSW bounds at a number of minimum and maximum density ratios. A first observation is that even at the most extreme density ratios plotted, $\inf_{\boldsymbol{\theta}} \log \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} = \frac{1}{1000}$ and $\sup_{\boldsymbol{\theta}} \log \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} = 1000$, the improvement over JSW bounds is roughly 15%. Taken together, this shows that even if the magnitude of the minimum and maximum density ratios is in the thousands, kappa bounds provide roughly a 30% improvement over JSW bounds. Decreasing the order to the minimum and maximum increases the improvement. On the other side, looking at an approximation where the maximum and minimum density ratios are 2 and $\frac{1}{2}$ shows massive improvement, ruling out around 90% of the JSW interval.

This result directly addresses the large width and 100% coverage of JSW intervals. Under a reasonably diffuse prior on the marginal likelihood, especially under the least favorable prior, some portion of the JSW interval gives values that cannot exist. Since these values are at the end of the interval and are almost surely covered, JSW intervals will always include all of the possible values that the marginal likelihood can viably take. Kappa bounds correct for this problem and presents shorter intervals that only include values that can possibly exist, leading to shorter intervals and more accurate coverage.

4.5.2 Estimation of Kappa Bounds

The inequality in Theorem 4.5.2 implies a range on the log marginal likelihood, but gives no information about where in the range it may be. As in Definition 4.4.3, treating the log marginal likelihood as a random variable allows for specification of a conditional posterior:

$$\mathbb{P}(\Delta|a,b,\mathcal{U},\mathcal{L},\kappa(\beta_1),\kappa(\beta_2),\boldsymbol{\theta}^*,\boldsymbol{\omega}^*) = \frac{\left(\Delta - \mathcal{L} - \frac{(\mathcal{U}-\mathcal{L})}{1+\kappa(\beta_1)}\right)^{a-1}\left(\mathcal{U} - \frac{\kappa(\beta_2)(\mathcal{U}-\mathcal{L})}{1+\kappa(\beta_2)} - \Delta\right)^{b-1}}{\boldsymbol{B}(a,b)\left((\mathcal{U}-\mathcal{L})\left(1 - \frac{\kappa(\beta_2)}{1+\kappa(\beta_2)} - \frac{1}{1+\kappa(\beta_1)}\right)\right)^{a+b-1}} \tag{4.33}$$

Given this prior construction, the distribution of interest is a posterior over the possible values of $\log \mathcal{P}(\mathcal{X})$:

$$\mathbb{P}(\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*) = \int\limits_{-\infty}^{0}\int\limits_{-\infty}^{0}\int\limits_{0}^{1}\int\limits_{1}^{\infty}\mathbb{P}(\Delta|a,b,\mathcal{U},\mathcal{L},\kappa(\beta_1),\kappa(\beta_2),\boldsymbol{\theta}^*,\boldsymbol{\omega}^*) \\ \mathbb{P}(\mathcal{U},\mathcal{L},\kappa(\beta_1),\kappa(\beta_2)|\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)d\kappa(\beta_1)d\kappa(\beta_2)d\mathcal{L}d\mathcal{U} \tag{4.34}$$

where $\boldsymbol{\theta}^*$ and $\boldsymbol{\omega}^*$ are a set of N draws from the true posterior and M draws from the approximation, respectively. Note that a and b are hyperparameters which influence the shape of the final distribution. Further, assuming that $\mathcal{U} \perp\!\!\!\perp \mathcal{L} \perp\!\!\!\perp \beta_1 \perp\!\!\!\perp \beta_2$

conditional on $\boldsymbol{\theta}^*$ and $\boldsymbol{\omega}^*$:

$$\mathbb{P}(\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*) = \int\limits_{-\infty}^{0}\int\limits_{-\infty}^{0}\int\limits_{0}^{1}\int\limits_{1}^{\infty}\mathbb{P}(\Delta|a,b,\mathcal{U},\mathcal{L},\kappa(\beta_1),\kappa(\beta_2),\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)\mathbb{P}(\mathcal{U}|\boldsymbol{\theta}^*)\mathbb{P}(\mathcal{L}|\boldsymbol{\omega}^*)$$
$$\mathbb{P}(\kappa(\beta_1)|\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)\mathbb{P}(\kappa(\beta_2)|\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)d\kappa(\beta_1)d\kappa(\beta_2)d\mathcal{L}d\mathcal{U} \tag{4.35}$$

The posterior distribution for the log marginal likelihood has no convenient form. Like Lemma 4.4.1, an asymptotic posterior expectation and variance can be defined.

Lemma 4.5.1. Let $\alpha \sim \text{Beta}(a,b)$ be a beta distributed random variable that ranges between zero and one. Also, let $K_1 = \frac{1}{1+\kappa(\beta_1)}$ and $K_2 = \frac{\kappa(\beta_2)}{1+\kappa(\beta_2)}$. Assuming that $\mathbb{V}[K_1]$ and $\mathbb{V}[K_2]$ go to zero as N and M get large, the asymptotic expectation and variance of $\mathbb{P}(\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)$ can be defined as:

$$\begin{aligned}\mathbb{E}[\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*] &= \mathbb{E}[(1-\alpha)]\left(\mathcal{U}_0 - (\mathcal{U}_0-\mathcal{L}_0)\mathbb{E}[K_1]\right) + \mathbb{E}[\alpha]\left(\mathcal{L}_0 + (\mathcal{U}_0-\mathcal{L}_0)\mathbb{E}[K_2]\right)\\ \mathbb{V}[\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*] &= (\mathcal{U}_0-\mathcal{L}_0)^2(1-\mathbb{E}[K_1]-\mathbb{E}[K_2])^2\mathbb{V}[\alpha]\end{aligned} \tag{4.36}$$

where $\mathbb{E}[\alpha] = \frac{a}{a+b}$, $\mathbb{E}[1-\alpha] = \frac{b}{a+b}$, and $\mathbb{V}[\alpha] = \frac{ab}{(a+b)^2(a+b+1)}$.

Proof. The proof follows Lemma 4.4.1 and leverages the assumption that $\kappa(\beta_1) \perp\!\!\!\perp \kappa(\beta_2) \perp\!\!\!\perp \mathcal{U} \perp\!\!\!\perp \mathcal{L}$. The result follows this assumption trivially.

Compared to Lemma 4.4.1, it is easy to see that the variance of the posterior shrinks in accordance with the amount of information gained from the extrema of the density ratio. As with JSW bounds, the variance is a function of the evidence upper and lower bounds and a prior belief about the location of the log marginal likelihood. Kappa bounds introduce the extrema conditions and show that the variance of the posterior can shrink in accordance with how good the approximation performs - an approximation that fits well everywhere can reduce the variance to essentially zero even if there is distance between the evidence upper and lower bounds.

Estimating kappa bounds first requires estimating the evidence lower and upper bounds and their corresponding variances. Definition 4.4.2 can be leveraged to get these values. Kappa bounds also require estimation of the relative extrema of the density ratio of the posterior and the approximation. In order to estimate the full posterior, it is necessary to estimate a random variable representation of these values. Recall that:

$$\kappa(\beta_1) = \kappa\left(\sup_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}\right)$$
$$\kappa(\beta_2) = \kappa\left(\inf_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}\right)$$

where $\kappa(t) = \frac{1+t\log t-t}{t-\log t-1}$. Estimating these two quantities requires estimating the density ratio of $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ to $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$. Density ratio estimation can be done in a variety of ways and one accurate and computationally viable approach leverages the duality of classification and density ratios.

Definition 4.5.1. Assume that we have N samples from $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ and M samples from $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$. Let $\boldsymbol{Y}$ be a vector of labels that encodes the distribution from which each sample was drawn and let $\boldsymbol{F}$ be the collection of samples:

$$\boldsymbol{Y} = \{\underbrace{1, 1, 1, ...}_{\text{N}}, \underbrace{0, 0, 0, ...}_{\text{M}}\} = \{y_1, y_2, ..., y_{N+M}\}$$

$$\boldsymbol{F} = \{\underbrace{\theta_1^*, \theta_2^*, ..., \theta_N}_{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}, \underbrace{\omega_{N+1}^*, \omega_{N+2}^*, ..., \omega_{N+M}^*}_{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}\} = \{f_1, f_2, ..., f_{N+M}\}$$

Applying Bayes rule:

$$\frac{\mathcal{P}(f_i|\mathcal{X})}{\mathcal{Q}(f_i|\boldsymbol{\gamma})} = \frac{\mathbb{P}(y_i = 1|f_i)\mathbb{P}(f_i)}{\mathbb{P}(y_i = 1)} \left(\frac{\mathbb{P}(y_i = 0|f_i)\mathbb{P}(f_i)}{\mathbb{P}(y_i = 0)}\right)^{-1} = \frac{\mathbb{P}(y_i = 1|f_i)}{\mathbb{P}(y_i = 0|f_i)} \frac{\mathbb{P}(y_i = 0)}{\mathbb{P}(y_i = 1)}$$

Assuming that the prior probability of classification is proportional to the sample sizes:

$$\frac{\mathcal{P}(f_i|\mathcal{X})}{\mathcal{Q}(f_i|\boldsymbol{\gamma})} = \frac{\mathbb{P}(y_i = 1|f_i)}{\mathbb{P}(y_i = 0|f_i)} \frac{M}{N}$$

Finally, recognizing this as a two class problem:

$$\frac{\mathcal{P}(f_i|\mathcal{X})}{\mathcal{Q}(f_i|\boldsymbol{\gamma})} = \frac{\mathbb{P}(y_i = 1|f_i)}{1 - \mathbb{P}(y_i = 1|f_i)} \frac{M}{N} \tag{4.37}$$

This classification approach provides a viable way to estimate the density ratio for any point $f_i \in \boldsymbol{\Theta}$ from samples. At its core, the classification method is estimating a function with inputs from the set of all possible parameters that outputs the density ratio at that point. A variety of classification approaches exist and provide varying types of classifiers. Any classification method used for estimating density ratios should exhibit a few properties and operate under a few assumptions:

1. The method should provide class probabilities. Further, the class probabilities should be asymptotically consistent to the true class probabilities.

2. $\mathcal{P}(y_i = 1|f_i)$ should be smooth and nonlinear in f. For the vast majority of distributions, there should be no expectation that the likelihood ratio monotonically increases or decreases in the feature space. As such, the classifier should uncover a smooth, but nonlinear, function of f for the probability that the value comes from $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$.

3. Related to nonlinearity, the class probability for values outside of the observed range should approach equality. Specifically, assume that $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ and $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ are probability distributions that share a common domain over all dimensions. Let $\mathcal{R} \subseteq \boldsymbol{\Theta}$ be the convex hull dictated by $\boldsymbol{F}$. For any $\eta \notin \mathcal{R}$, $\frac{\mathcal{P}(\eta|\mathcal{X})}{\mathcal{Q}(\eta|\boldsymbol{\gamma})} = 1$.

Using a classification approach that meets these requirements, the class probabilities can be used to estimate posterior distributions for $\kappa(\beta_1)$ and $\kappa(\beta_2)$ that converge, in expectation, to their true values:

Lemma 4.5.2. Let $\phi(f) = \mathbb{P}(y = 1|f \in \mathcal{R}, \boldsymbol{F})$ be the true probability that f comes from $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$. Given $T = N + M$ samples from the posterior and approximation, $f_i \in (f_1, f_2, ..., f_T)$, let $\hat{\phi}(f_i) = \mathbb{P}(y = 1|f = f_i)$ be the individual class probability estimates. Assuming that $\hat{\phi}(f_i) \xrightarrow{p} \phi(f)$, the individual class probability estimates can be used to derive an estimate for $\kappa(\beta_1)$ and $\kappa(\beta_2)$ such that:

$$\begin{aligned} g\left(\{\hat{\phi}(f_1), \hat{\phi}(f_2), ..., \hat{\phi}(f_T)\}\right) &\xrightarrow{p} \kappa(\beta_1) \\ h\left(\{\hat{\phi}(f_1), \hat{\phi}(f_2), ..., \hat{\phi}(f_T)\}\right) &\xrightarrow{p} \kappa(\beta_2) \end{aligned} \tag{4.38}$$

where $g()$ and $h()$ are functions of estimated class probabilities of the samples. In turn, there exist posterior distributions for $\kappa(\beta_1)$ and $\kappa(\beta_2)$ such that:

$$\begin{aligned} \mathbb{E}[\kappa(\beta_1)|\boldsymbol{F}] &= g\left(\{\hat{\phi}(f_1), \hat{\phi}(f_2), ..., \hat{\phi}(f_T)\}\right) \\ \mathbb{E}[\kappa(\beta_2)|\boldsymbol{F}] &= h\left(\{\hat{\phi}(f_1), \hat{\phi}(f_2), ..., \hat{\phi}(f_T)\}\right) \end{aligned} \tag{4.39}$$

Proof. A proof for this statement can be found in Appendix C.

Lemma 4.5.2 shows that tractably estimating a posterior distribution for $\kappa(\beta_1)$ and $\kappa(\beta_2)$ only requires estimating the class probabilities given by a probabilistic classification approach and accepting a few asymptotic assumptions. As long as the probabilistic classification method used gives consistent estimates of the class probabilities for each of the samples in the training set, the distribution function for the maximum and minimum density ratios can be estimated.

A nonlinear classification approach that meets the necessary conditions is the classification random forest and corresponding nonparametric probability machine (Malley et al., 2012; Wright and Ziegler, 2015). Using this approach for classification, a method for estimating the posterior distributions for $\mathbb{P}(\kappa(\beta_1)|\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$ and $\mathbb{P}(\kappa(\beta_2)|\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$ has the following structure:

1. Consider a $T \times K$ matrix of training data, $\boldsymbol{F}$, that includes N samples from $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ and M samples from $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$.

2. A bootstrap sample with replacement of size T is taken from the training set. The samples left out due to the bootstrap procedure are called the out-of-bag (OOB) samples.

3. A classification tree is grown using the bootstrap data set. The tree is constructed by recursively splitting data into distinct subsets so that one parent node leaves two child nodes. For splitting data, all splits of a random subset of features are considered. The optimal split of a feature is the one minimizing the value of some dichotomous purity measure, such as the Gini index or deviance.

4. The tree is grown to the greatest extent possible but requiring a minimum nodesize of 10% of the sample. No pruning is performed. The final nodes in a tree are called terminal nodes. Note that this step is somewhat different than the standard random forest algorithm.

5. Class probabilities for each OOB sample are computed by dropping the observation down the tree to its terminal node and computing the proportion of in-bag samples belonging to each class at that node.

6. Steps 2-5 are repeated B times to grow B classification trees.

7. Across the B trees, let $\hat{\phi}(f_i)$ be the set of S_i class probabilities dictating $\mathbb{P}(y = 1|f_i, \boldsymbol{F})$ computed when observation $i \in (1, ..., T)$ is OOB. Given that the bootstrap with replacement includes, on average, 63.2% of the data, $\mathbb{E}[S_i] = .368B$. Define the sample means and variances as:

$$\bar{\phi}_{S_i}(f_i) = \frac{1}{S_i}\sum_{s=1}^{S_i}\hat{\phi}(f_{si})$$

$$\hat{\sigma}^2_{\phi(f_{si})} = \frac{1}{S_i - 1}\sum_{s=1}^{S_i}(\hat{\phi}(f_{si}) - \bar{\phi}_{S_i}(f_i))^2$$

and define the sampling distribution for $\mathbb{P}(y=1|f_i,\boldsymbol{F})$ as:

$$\mathcal{N}\left(\bar{\phi}_{S_i}(f_i),\frac{\hat{\sigma}^2_{\phi(f_{si})}}{S_i}\right)$$

8. Use empirical evaluation methods to construct the CDF of the minimum and maximum class probabilities. For a large number of points $g\in(0,1)$:

$$\mathbb{P}(\max\phi(f)<g)=\prod_{i=1}^{T}\int\limits_{0}^{g}\mathcal{N}\left(x\ ;\ \bar{\phi}_{S_i}(f_i),\frac{\hat{\sigma}^2_{\phi(f_{si})}}{S_i}\right)dx$$
$$\mathbb{P}(\min\phi(f)<g)=\prod_{i=1}^{T}\int\limits_{g}^{1}\mathcal{N}\left(x\ ;\ \bar{\phi}_{S_i}(f_i),\frac{\hat{\sigma}^2_{\phi(f_{si})}}{S_i}\right)dx$$

9. Transform g such that:

$$u(y)=\frac{My}{N(1-y)}\ ;\ v(y)=\frac{1+y\log y-y}{y-\log y-1}\ ;\ g^*=v(u(g))$$

10. Given the distribution functions for the minimum and maximum, random draws can be taken from each distribution by passing a uniform random variate to the inverse distribution functions. Draws taken in this way are equivalent to draws taken from $\mathbb{P}(\kappa(\beta_1)|\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)$ and $\mathbb{P}(\kappa(\beta_2)|\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)$.

JSW bounds represent a limiting case of kappa bounds; when samples from the posterior and approximation are associated with class probabilities that are essentially zero or one, kappa bounds are exactly the same as JSW bounds. In cases where the approximation and posterior are truly far apart, this is the desired behavior. However, random forest classification algorithms seek to draw sharp boundaries between classes and unbalanced gaps that exist between the posterior and approximations samples can lead the algorithm to estimate class probabilities that are much closer to zero or one than they truthfully are. This can be prevented by drawing many samples from both the posterior and approximation, so it is of utmost importance that a large number of samples is taken from each distributions. This is even more important for high dimensional posteriors due to the curse of dimensionality. Similarly, it is important that the number of trees that are grown in the random forest algorithm is large enough that each observation is OOB enough times such that the normality assumption is not egregiously violated. A general rule of thumb that leads to good performance is to set N and M equal to at least two times the number of dimensions in the posterior. Similarly, growing at least 3680 trees for the random forest classifier leads to each observation being OOB 1000 times, in expectation. These values show good performance in tests and bootstrapping shows that these values allow the assumptions made above to hold. New implementations of the random forest algorithm allow for quick and efficient training of classifiers, even on large and high-dimensional data sets, so computation time is of little concern.

Using this algorithm to estimate the distribution of $\kappa(\beta_1)$ and $\kappa(\beta_2)$, a full algorithm for estimating kappa bounds has the following scheme:

1. Given a model with parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_k\} \in \boldsymbol{\Theta}$, prior density, $\pi(\boldsymbol{\theta})$, and observed data, $\mathcal{X}$, use simulation methods to get N draws from $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$, $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, ..., \theta_N^*) \in \boldsymbol{\Theta}$.

2. For each $\theta_i^* \in \boldsymbol{\theta}^*$, compute the complete data likelihood, $\mathcal{P}(\theta_i^*, \mathcal{X})$.

3. Given $\boldsymbol{\theta}^*$, choose an approximate posterior, $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$.

4. Assess the density of each $\theta_i^* \in \boldsymbol{\theta}^*$ with respect to $\mathcal{Q}$, $\mathcal{Q}(\theta_i^*|\boldsymbol{\gamma})$.

5. Take M independent and identically distributed draws from $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$, $\boldsymbol{\omega}^* = (\omega_1^*, \omega_2^*, ..., \omega_M^*) \in \boldsymbol{\Theta}$.

6. For each $\omega_j^* \in \boldsymbol{\omega}^*$, compute the complete data likelihood $\mathcal{P}(\omega_j^*, \mathcal{X})$ and assess the density with respect to $\mathcal{Q}$, $\mathcal{Q}(\omega_j^*|\boldsymbol{\gamma})$.

7. Compute the parameters for the posterior distributions of $\mathcal{U}$ and $\mathcal{L}$ in Definition 4.4.2. Similarly, compute the empirical approximation to the posterior distributions of $\kappa(\beta_1)$ and $\kappa(\beta_2)$ in Definition 4.5.2 outlined above.

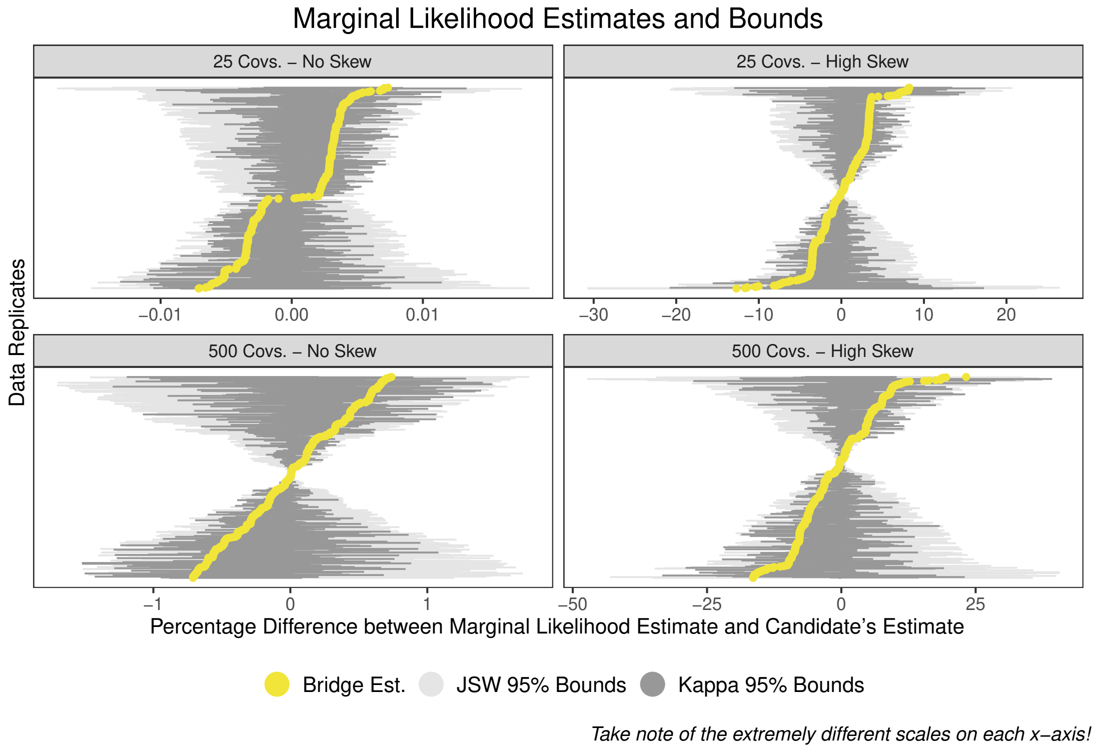
8. Given the set of posteriors, take a large number of draws from $\mathbb{P}(\mathcal{U}|\boldsymbol{\theta}^*)$, $\mathbb{P}(\mathcal{L}|\boldsymbol{\omega}^*)$, $\mathbb{P}(\kappa(\beta_1)|\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$, and $\mathbb{P}(\kappa(\beta_2)|\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$. Given the prior parameters a and b, plug these values into the four-parameter beta distribution and take draws from the posterior. This creates a large number of draws from the posterior for the marginal likelihood. Posterior intervals can be calculated using these Monte Carlo samples.

4.5.3 Performance of Kappa Bounds

Like JSW bounds, empirical estimates of kappa bounds on the marginal likelihood should surround the true marginal likelihood. As with the bridge sampling estimator and JSW intervals, this can be assessed directly in a situation where the true marginal likelihood is known. Unlike JSW intervals, however, kappa bounds should not always include the true value of the marginal likelihood - since kappa bounds limit the width of the interval to only include marginal likelihood values that are possible given the ELBO and EUBO, Again, this is best explored by simulating the coverage of kappa bounds and comparing their width to JSW intervals.

Using the running linear regression example, I compare the bridge sampling estimator, JSW intervals, and kappa intervals to the true log marginal likelihood calculated by the Candidate's estimator. After the true model is estimated under varying numbers of covariates, posterior warping is used to induce skewness. The posteriors are approximated by multivariate normal distributions that are fit using a moment-matching

FIGURE 4.4. Bridge sampling estimator, JSW intervals, and Kappa intervals for 500 randomly generated linear regression data sets with varying numbers of covariates and amounts of posterior skew.



Note: Data are generated from a linear regression model with a known error variance. Skewness is induced using posterior warping techniques from Meng and Schilling (2002). 95% HDP intervals are shown for JSW and Kappa intervals. Percentages are computed with respect to the Candidate's estimator to the marginal likelihood for each model. Bayesian linear regression posterior samples and Candidate's estimator were computed using `MCMCregress` from `MCMCpack` in `R` (Martin et al., 2011).

technique. As with JSW intervals, kappa bounds are estimated using a maximum-width Beta$(0, 0)$ prior. The location of kappa intervals is compared to the location of JSW intervals and the bridge sampling point estimator. Coverage of JSW intervals and kappa bounds are computed and compared. Since 95% intervals are computed, an ideal interval would show that roughly 95% of the 500 intervals estimated in this simulation scheme cover the true log marginal likelihood value.

Figure 4.4 shows the results of these simulations. A first glance shows that the location of kappa bounds and JSW bounds are generally similar and cover the true log marginal likelihood. Compared to the bridge sampling estimator, the interval estimates perform favorably - true values are covered even when the bridge sampling estimator is far away from the true value. There is a strong correlation between the magnitude of the incorrectness of the bridge sampling estimator and the width of the intervals. In many ways, this is a desirable result; since the underlying machinery of the bridge sampling estimator and intervals are similar, it is expected that they share similar locations. However, the fact that the intervals cover the true value make model comparison under the interval estimators superior to the bridge sampling estimator.

JSW intervals are overly wide and have 100% coverage of the true log marginal likelihood value. Figure 4.4 shows that the theoretical improvements of the kappa intervals are carried to estimation and the expected shrinkage of the JSW interval occurs. For the posteriors with no skew, the kappa interval is, on average, 63% smaller. This corresponds to an average maximum log density ratio of the posterior to the approximation of approximately 3 and a minimum of -3. For posteriors with skew, the improvement is roughly 51%, on average, corresponding to maximum and minimum log density ratios of approximately 5 and -5. Under skewness in the posterior, multivariate normality is no longer approximately correct. However, these simulations show that even a rough approximation demonstrates significant improvement.

A second consideration is the interval coverage. In the case of no skewness, across different dimensionalities of the posterior, 95% kappa intervals have roughly 95% coverage (95.2% for 25 covariates and 94.8% for 500 covariates). In the case of skewness, the coverage is lower (91.4% for 25 covariates and 89.6% for 500 covariates), but still reasonable. The decreased coverage is likely explained by violations of the independence assumption, specifically $\mathcal{U} - \mathcal{L} \perp\!\!\!\perp \kappa(\beta_1), \kappa(\beta_2)$. $\mathcal{U} - \mathcal{L}$ encodes one sense of closeness of the posterior and approximation, the average density ratio, while $\kappa(\beta_1)$ and $\kappa(\beta_2)$ encode a notion of absolute closeness similar to the total variation distance. Pinsker's inequality proves these two quantites are related and, for finite sample spaces, the KL divergence is bounded by the total variation distance (Basu and Ho, 2006). As such, it is reasonable to expect that there is some correlation between $\mathcal{U} - \mathcal{L}$ and $\kappa(\beta_1), \kappa(\beta_2)$. While the extent of this effect is not explicitly addressed in this simulation, the coverage is not wildly off from the expected coverage and further simulations show that this effect does not seem to have an impact in realistic model evaluation settings. Thus, corrections for this correlation are left to further work.

Taken as a whole, these simulations show that kappa bounds provide a superior approach to estimating the log marginal likelihood over JSW bounds and the bridge sampling estimator. Even accounting for the low additional computational overhead, kappa bounds provide a fast and flexible approach to estimating the log marginal likelihood that can be applied to a wide variety of Bayesian models.

4.6 Application of Kappa Bounds to Ordered Factor Analysis Models

4.6.1 Kappa Bounds Applied to Simulated Data

The previous simulations demonstrate the accuracy and coverage of kappa bounds in a situation where the log marginal likelihood is known. While this was important for comparison and assessment of quality, kappa bounds are most useful in situations where the marginal likelihood is difficult to calculate and current methods of approximation work poorly. One such model is the ordered factor analysis model.

Let $\mathcal{X}$ be a data set of Q observations of P ordered discrete items. Ordered factor analysis seeks to find sources of shared covariance among the items. Specifically, the

P items associated with observation $q \in (1, ..., Q)$ have the following model:

$$\mathcal{X}_q^* = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\omega}_q + \boldsymbol{\sigma}_q \tag{4.40}$$

where $\mathcal{X}_q^*$ is a P-vector of continuous latent variable representations of the observed discrete outcomes, $\boldsymbol{\alpha}$ is a P-vector of intercepts, $\boldsymbol{\omega}_q$ is a K-vector of factor scores associated with each observation, $\boldsymbol{\Lambda}$ is a $P \times K$ matrix of factor loadings, and $\boldsymbol{\sigma}_q$ is P-vector of idiosyncratic noise that follows a P-dimensional multivariate standard normal distribution with diagonal covariance. In order to effectively model the discrete nature of the outcomes, $\mathcal{X}_j^* \; \forall \; j \in (1, ..., P)$, define a set of $C_j + 1$ ordered cut points for each item, $\gamma_{j,k} \in (\gamma_{j,1}, ..., \gamma_{j,C_j+1})$ where $\gamma_{j,1} = -\infty < \gamma_{j,2} = 0 < ... < \gamma_{j,C} < \gamma_{j,C+1} = \infty$ where C_j is the number of possible values that an observations take on item j. Assuming that $\mathcal{X}^*$ follows a normal distribution, each discrete value can be modeled as:

$$P(\mathcal{Y}_{q,j} = c \in (1, ..., C_j)|-) = \int\limits_{\gamma_{j,c}}^{\gamma_{j,c+1}} \mathcal{N}(x; \mathcal{X}_{qj}^*, 1)dx \tag{4.41}$$

Under a Bayesian specification, model estimation can be done using MCMC methods (see Martin et al. (2011) for details on model fitting). Conjugate priors exist for all parameters except for the cutpoints, which are modeled using a normal distribution.

Ordered factor analysis is often used as a tool for testing theories of relationships between covariates and model comparison between competing models is a key step for using this model. There are two related choices that the researcher must make in order to specify and identify factor analysis models. First, the dimensionality of the latent scores is undefined and must be chosen by the researcher. For identification purposes, K must be less than P, but any other possible specification is a viable choice. Theory may dictate how many dimensions should exist. However, the dimensionality of the latent space is often related to the main question being answered by the ordered factor analysis model and testing is needed to choose between a number of competing theories.

Second, ordered factor analysis is often used to test direct relationships between items and this requires fitting models with a number of constraints in the loadings matrix (see Joreskog (1969) for a more thorough discussion of confirmatory factor analysis methods). While a few constraints in the loadings matrix are required for model identification due to the factor analysis model's inherent rotational invariance (Geweke and Zhou, 1996), most constraints are placed to correspond with a specific measurement construct that matches a theory. These constraints have wide-reaching implications for research questions and can be used to empirically codify different theories. Choosing between competing theories, then, requires comparing numerous models that have different constraints.

As with any factor analysis model (and structural equation models, more generally), model comparison is a key step for theory testing in ordered factor analysis models. A number of approaches exist for comparing factor analysis models, but most of these approaches are theoretically deficient and result in comparisons that are known to unduly favor overly parsimonious models due to strong normality assumptions in the observed data (Forero et al., 2009). This problem is even further exacerbated when using discrete factor analysis models since the observed outcomes are, by definition, not normally distributed.

Many of these problems are addressed for the continuous factor analysis by estimating the model under a Bayesian specification and computing the marginal likelihood for each competing model. Lopes and West (2004) explore a number of approaches for estimating the marginal likelihood of continuous factor analysis models and show that the bridge sampling estimator provides an accurate approximation to the value of interest, performing better than approximations to the Candidate's estimator, Laplace approximation, various factor analysis fitting criteria, and even their own reversible-jump MCMC method when attempting to select the correct model. While this work only seeks to determine dimensionality, the results show that marginal likelihood estimation is an ideal approach to comparing factor analysis models due to its inherently balanced overfitting and underfitting penalties.

Lopes and West (2004) only make comparisons for exploratory factor analysis models with continuous and approximately normal observed outcomes. From previous simulations, it is reasonable to believe that the bridge sampling estimator suffers in less ideal conditions; particularly those that are presented by the posterior from the ordered factor analysis model. First, continuous factor analysis essentially guarantees that the resulting posteriors will be approximately multivariate normal due to its strong conjugate normal-normal structure and this structure can be further ensured by marginalizing over the factor scores. Ordered factor analysis, on the other hand, is known to produce posteriors for the loadings that are skewed when the number of observations associated with each possible outcome are unbalanced and marginalization over the factor scores is computationally infeasible, leading to a high dimensional posterior.[7] Second, Lopes and West (2004) do not consider the ability of various marginal likelihood approximation methods to discriminate between models with various constraints. Dealing with constraints is often more tedious than adding or subtracting whole dimensions from the latent space, and can lead to models where factor loadings are only barely identified and present wide, skewed posteriors (Ghosh and Dunson, 2009). Finally, from a computational standpoint, ordered factor analysis often exhibits high autocorrelation and poor mixing in MCMC chains due to difficulty identifying and estimating cutpoints, especially as the number of dimensions and con-

[7] $PK - P + QK + \sum\limits_{j=1}^{P} C_j$ dimensions in an ordered factor analysis posterior compared to $PK + 2P$ in the reduced form continuous factor analysis posterior, to be exact, minus any fixed values within the factor loadings matrix.

straints increases. While this can be partially addressed by long burnin periods and heavy thinning, the various quirks that exist in the posterior could lead to trouble for the bridge sampling estimator.

Ordered factor analysis is an ideal application for interval estimates of the marginal likelihood, specifically kappa bounds. While the posterior under a well-identified model is expected to be approximately normal, there is little reason to believe that posteriors are approximately symmetric in less well-behaved scenarios. Assuming standard conjugate priors, uniform priors on the cutpoints, and treating $\boldsymbol{\alpha}$ as a standard intercept, the complete data likelihood is easily computed for each posterior draw from an ordered factor analysis model:

$$\mathcal{P}(\mathcal{X}, \boldsymbol{\theta}) = \left(\prod_{q=1}^{Q}\prod_{j=1}^{P}\mathbb{P}(\mathcal{Y}_{q,j} = \mathcal{X}_{q,j}|-)\right)\left(\prod_{j=1}^{P}\prod_{k=1}^{K+1}\mathcal{N}(\lambda_{j,k}; \lambda_0, \sigma^2_{\lambda_0})\right)\left(\prod_{q=1}^{Q}\prod_{k=1}^{K}\mathcal{N}(\omega_{q,k}; 0, 1)\right) \tag{4.42}$$

It is similarly easy to take a set of draws from an approximation to the posterior and compute the associated complete data likelihood values. Thus, the algorithm for kappa bounds is easily implemented on posterior draws from an ordered factor analysis model.

To demonstrate the efficacy of kappa bounds on a more complicated model, I use simulation to show the quality of kappa bounds in estimating reasonable interval estimates of the marginal likelihood of ordered factor analysis models specifically for the purpose of model selection. I compare kappa bounds to the bridge sampling estimator and JSW bounds. Specifically, I generate 500 data sets of 500 observations over 10 discrete, ordered items that have between 3 and 7 unique responses from a known set of loadings, constraints, and factor scores. Using the ordered factor analysis implementation from `MCMCpack` in `R`, `MCMCordfactanal`, I generate samples from the resulting posterior distributions of 5 potential models shown in Table 4.1 (Martin

TABLE 4.1. Factor loading constraints matrix for each model compared on simulated data.

M1	M2	M3	M4	M5
$\begin{bmatrix} + \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{bmatrix}$	$\begin{bmatrix} + & 0 \\ & 0 \\ & 0 \\ & 0 \\ & 0 \\ 0 & + \\ 0 & \\ 0 & \\ 0 & \\ 0 & \end{bmatrix}$	$\begin{bmatrix} + & 0 \\ & 0 \\ & 0 \\ & 0 \\ & 0 \\ & + \\ & \\ & \\ & \\ & \end{bmatrix}$	$\begin{bmatrix} + & 0 \\ & \\ & \\ & \\ & \\ & + \\ & \\ & \\ & \\ & \end{bmatrix}$	$\begin{bmatrix} + & 0 & \\ & 0 & \\ & 0 & \\ & 0 & \\ & 0 & \\ & + & \\ & & \\ & & 0 \\ & & 0 \\ & & + \end{bmatrix}$

Note: 0 indicates that the loading is fixed to zero. + indicates that the loading is constrained to be positive. Note that M3 is the correct model for the simulated data.

et al., 2011). These models range from the minimally constrained one dimensional model to a three dimensional model that is close to the true model. Constraints associated with each model can be seen in Table 4.1 and the simulated data was generated using M3. Diffuse $\mathcal{N}(0, 100)$ priors are used for the factor loadings in each model. For each data set, I ran the MCMC algorithm for 20,000 burnin iterations and kept 10,000 posterior draws that were thinned every 5 iterations. For the cutpoints that are drawn via a Metropolis-Hastings step, the width of the normal proposal distribution was tuned for all parameters such that the acceptance ratio was as close to 25% as possible. Monte Carlo standard errors were used to monitor convergence of each run and there were no indications of convergence issues for the structural parameters. The effective sample size of the complete data likelihood was used as the number of posterior samples.[8]

[8] Effective sample size was monitored for all parameters. However, it is difficult to use this much information to make an informed choice on stopping the chain. As such, I monitored the ESS on the complete data log likelihood for each iteration. Over all 500 replicated data sets, the ESS ranged between roughly 3,500 and 7,200. In the worst case, the ESS was roughly two times the number of parameters being monitored. In practice, this appears to be more than enough draws to establish consistent estimates for the bridge sampling estimator, JSW bounds, and kappa bounds. As a sanity check, I ran multiple chains for a number of the models and these estimates of the log marginal likelihood remained relatively constant. This work leaves little reason to believe that the

For each model, a multivariate normal distribution was used for the approximation to the posterior and maximum likelihood estimates for the mean and covariance matrix was used to parameterize the approximation. Note that the dimensionality of the multivariate normal distribution varied widely depending on the model, ranging 530 dimensions in the smallest model to 1564 dimensions in the largest, with the bulk of the parameters being the factor scores for the 500 observations across 1, 2, or 3 dimensions. For elements of the factor loadings matrix that were constrained to be positive, the approximation was truncated at 0. Posteriors for the cutpoints were also included in the approximation and were truncated at zero.[9] For each model, I took 10,000 i.i.d. draws from the multivariate normal approximation.

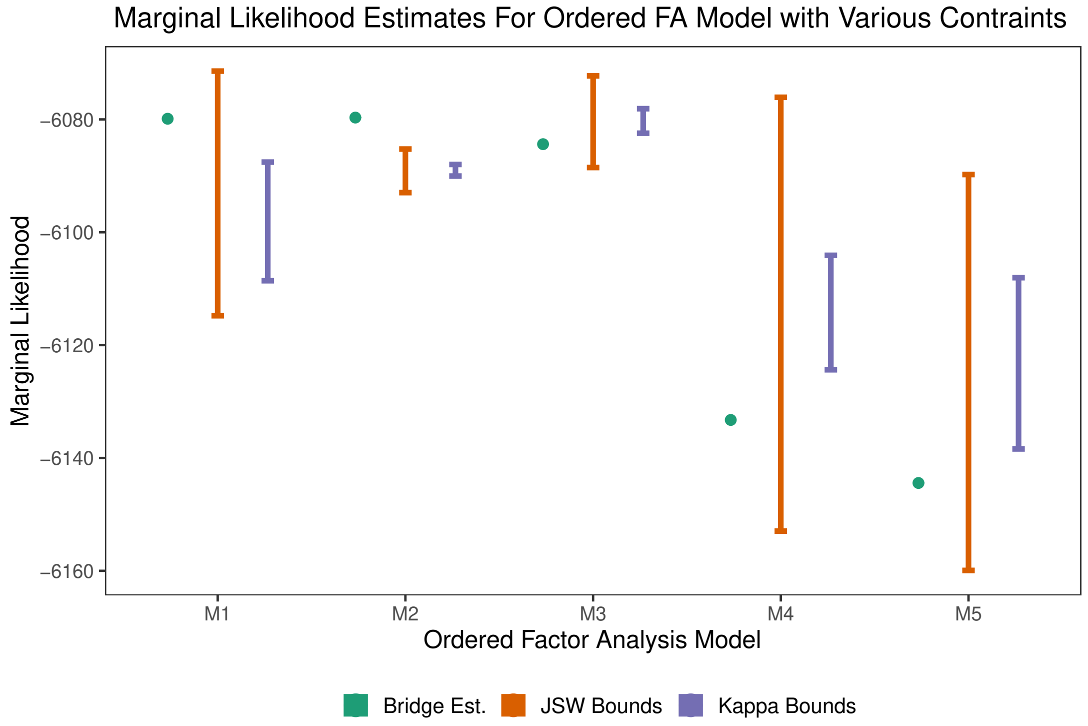
When computing kappa bounds, I used the probability classification random forest algorithm from the `ranger` package in `R`. For each model, I used 6 CPU cores at 3.7 GHz to compute 3,680 trees to learn the classification probabilities and corresponding standard errors. This took 45 seconds per model, on average. For the interval estimators, I used a Beta$(0, 0)$ priors on the log marginal likelihood and took 100,000 Monte Carlo draws from the empirical posteriors of $\mathcal{U}$, $\mathcal{L}$, $\kappa(\beta_1)$, $\kappa(\beta_2)$ which resulted in 100,000 draws from the posterior on the marginal likelihood. 95% highest posterior density intervals were derived for the interval estimates using these draws. For the bridge sampling estimator, the iterative scheme was run until the difference in iterations was less than .0001 on the log scale.[10]

results are a function of having too few draws from the posterior.

[9] The strict ordering constraint of the cutpoints presents a challenge when defining an approximation. Since there is no easy way to guarantee that the approximation draws follow the strict ordering on their own, I ensured that draws taken from the approximation for the cutpoints obeyed the necessary inequalities by ex-post truncating the draws in order from lowest to highest. I also tried to ensure this ordering by only accepting draws from the approximation that had the correct ordering, which proved to be computationally expensive. Both methods yielded similar results, so I went with the less costly approach.

[10] A single run for each model including the MCMC, draws from approximation, computation of the complete data log likelihood for both the posterior and approximation draws, computation of density with respect to the approximation, random forest, and Monte Carlo draws from the various

FIGURE 4.5. Comparison of marginal likelihood estimates for ordered factor analysis models.



Note: Bars denote the 95% HPD implied by 100,000 Monte Carlo samples for JSW and Kappa bounds. Note that M3 is the model from which the data was generated..

TABLE 4.2. Probability each ordered factor analysis model best fits the data.

	M1	M2	M3	M4	M5
Bridge.Est.	0.4472	0.5527	.0001	0	0
JSW Bounds	0.1833 [0,1]	0.0338 [0,0.5128]	0.7364 [0,1]	0.0465 [0,0.9434]	0 [0,0]
Kappa Bounds	0 [0,.0006]	.0004 [0,0.0019]	0.9996 [0.9978,1]	0 [0,0]	0 [0,0]

Note: 95% intervals are presented for JSW Bounds and Kappa Bounds. Probabilities are calculated as the probability each marginal likelihood is greater than all other models' marginal likelihood. 10,000 Monte Carlo samples are used to compute this probability.

Figure 4.5 shows the results for each marginal likelihood estimation method for the first data replicate. Table 4.3 shows the probability that each model is considered the best model under each marginal likelihood estimation technique and presents 95% intervals on this quantity for the interval estimates. Given that M3 is the correct model, it is easy to see that the bridge sampling estimator unduly favors the lower dimensional model (M1) and the overconstrained model (M2). Exploring the posteriors that result from this model demonstrate that the same effects seen in previous simulations hold true: under skewness, the bridge sampling estimator is produces precise but incorrect estimates. In all cases, even the one dimensional model, there is significant skew in the factor loading posteriors and multimodality in the cutpoint posteriors. Beyond favoring the incorrect models, the bridge sampling estimator does a poor job of specifically choosing the best model; M1, M2, and M3 all have essentially the same estimate for the marginal likelihood. In some sense, this is desirable since it is selecting the wrong model, but incorrect is incorrect. JSW intervals fair better as the midpoint of each interval is in an approximately correct order. However, the wide bounds on each interval make choosing a best model impossible since they all include log marginal likelihood estimates around -690. This is reflected in Table 4.3 since the 95% intervals for three separate models essentially cover the entire unit interval.

Kappa bounds provide the correct answer in this specific example. While kappa bounds cannot discriminate between certain models, it does select the correct model under the 95% intervals. Perhaps just as interesting, kappa bounds significantly improve on the JSW intervals in all cases, reducing the width by at least 50% across all models. Even under a truly skewed posterior with incorrect assignment of high

posterior took 2 minutes, on average. All in, this entire simulation took roughly 4 days to complete. This simulation, along with the many other computational strains that were given in pursuit of this topic, also took the life of two sticks of RAM that had served me well for a little over a year. You are gone but not forgotten.

TABLE 4.3. Proportion of times over 500 data replicates that each ordered factor analysis model best fits the data.

	M1	**M2**	**M3**	**M4**	**M5**
Bridge Est.	.384	.364	.252	0	0
JSW Bounds	0	0	.028	0	0
Kappa Bounds	0	.002	.976	0	0

Note: For interval estimates, 95% intervals are estimated and a winner is only declared if the highest marginal likelihood interval does not cross any other interval. This leads to less than 500 unique winners.

density areas in the posterior, the kappa bounds eliminate a large portion of the JSW bounds that represent impossible combinations of $KL(\mathcal{P}||\mathcal{Q})$ and $KL(\mathcal{Q}||\mathcal{P})$. Looking at Table 4.3, it is clear that kappa bounds pick the correct winner almost 100% of the time. When compared to the other two approaches, kappa bounds provide the only marginal likelihood estimates that confidently select the correct model.

To ensure that this result holds over many different data sets, I computed the number of times that each estimation approach distinctly selected each model over the 500 data replicates. For the bridge sampling estimator, this simply requires selecting the model for each data set that has the highest marginal likelihood estimate. For JSW bounds and kappa bounds, this required determining if the lower bound on the highest 95% interval estimate is greater than all other models. This led to a situations where there were not 500 unique winners for the intervals. Table 4.3 shows these results and demonstrates that the single example is generally indicative of the behavior for each method. Bridge sampling presents a tossup between M1, M2, and M3 while JSW bounds almost never pick a winner. However, kappa bounds select the appropriate model 97.6% of the time. This difference makes it clear that kappa bounds are a superior approach for estimating the marginal likelihood of ordered factor analysis models. Generally, this simulation demonstrates the promise of this approach for marginal likelihood estimation in high dimensional non-normal settings.

4.6.2 Kappa Bounds Applied to Racial Resentment in the U.S.

Comparison of ordered factor analysis models is a central task in the social sciences, frequently used when analyzing political beliefs and behavior through survey data. Common survey items ask respondents to agree of disagree, rating their opinion on an ordered scale, with statements related to concepts of interest. Factor analysis methods are then used to model common covariance sources and to map questions to latent constructs of interest (Joreskog, 1967, 1969). While ordered factor analysis implementations are widely available, many scholars still use the standard factor analysis toolkit when analyzing ordered discrete survey data. A large reason that this is the case is that there are few model comparison tools available for the ordered factor analysis toolkit. Kappa bounds provide a reliable option for doing just this.

I choose to demonstrate the practical usage of kappa bounds by exploring the racial resentment scale developed by Kinder et al. (1996). The racial resentment scale uses a carefully designed battery of questions asked in the 1986 American National Election study to explore attitudes of "symbolic racism" in the American public. Kinder et al. (1996) argue that this racial resentment measure acts as a unique combination of traditional racism and old-fashioned American individualism and explains much of the resistance from white Americans to racial policies. Kinder et al. (1996) show evidence for this theory through a series of confirmatory factor analyses that demonstrate that the racial resentment battery explains much of the correlation to attitudes on affirmative action and other racial policies where traditional racism and individualism items do not.

This theory has received numerous criticisms in the political science and sociology literatures since its conception, specifically regarding attitudes towards black Americans. Most of this criticism stems from the fact that the racial resentment measure is confounded with the outcome that it attempts to measure - racial policy attitudes. This has led to numerous studies, Carmines et al. (2011), Feldman and Huddy (2005), and Roos et al. (2019) just to name a few, which have shown that the racial resentment measure explains little variation that standard survey items related to traditional racism and individualism do not.

Since one aspect of this issue related to comparison of factor analysis models, kappa bounds applied to ordered factor analysis provides a new and reliable approach to comparing various models related to racial resentment measure. I use kappa bound estimates for the marginal likelihood to explore the relationship between the racial resentment, individualism, and traditional racism batteries in the 1986 ANES and explore whether there is evidence that the racial resentment battery explains variation in American attitudes that is unexplained by individualism and traditional racist attitudes through simple model comparison.[11]

I used the 1986 ANES data used in Kinder et al. (1996) and Carmines et al. (2011) to perform this exercise. I included three separate survey question batteries: 6 questions related to racial resentment, 5 related to individualism, and 4 related to traditional racist attitudes. Each of these questions asked respondents to rate their agreement or disagreement with a statement on a scale of strongly disagree to strongly agree, coded as a 5-point Likert scale. Of the 2,176 respondents included in the 1986 ANES, half were assigned to the survey group that responded to these questions. I further chose to only keep respondents that chose to respond to at least 12 of the 15 question (i.e.

[11] Note that this exercise is simply empirical. I do not purport that this work confirms or denies any existing theories of how or why attitudes have changed towards racial policies in the U.S. For a more thorough discussion on mechanisms and theories, see Carmines et al. (2011), Feldman and Huddy (2005), Roos et al. (2019), and the many other great works that are cited by these authors.

those that provided an explicit agreement rating to at least 12 out of 15 questions). This led to a final sample size of 1,042 respondents over 15 items.

To explore the relationship between these three survey batteries, I specified a number of confirmatory ordered factor analysis models having between 1 and 3 dimensions, ensuring that each model had enough constraints on the factor loadings matrix to ensure identifiability of the resulting structural estimates.[12] These models tested numerous viable theories related to the racial resentment measure ranging from a simple one dimensional model to more complex factor patterns that imply a range of conditional independence structures between the three survey batteries.

I present results showing model comparison metrics for 8 of these models. A summary of the constraint specifications for each of these eight models can be seen in Table 4.4. Each ordered factor analysis model was run using the same choices for number of draws, thinning intervals, etc. from the previous section. An important note was that for some models, particularly those with poor choices of constraints, model convergence for the structural parameters was more of a problem. This is generally an indication of a poorly fitting model. For the purposes of providing meaningful estimates for each model, any convergence problems were addressed by doubling the number of burnins and the thinning interval.

Figure 4.6 shows kappa bounds and bridge sampling estimates to the marginal likelihood for each of these eight models. The fit of these models varies widely and depends

[12] I followed the identification structure presented by Geweke and Zhou (1996). This amounts to choosing one positivity constraint on each dimensions and ensuring that enough structural zeros are added to the loadings matrix such that each matrix could be expressed in a lower block triangular structure with positive elements on the diagonal and zeros above the main diagonal. Directional constraints were chosen after an initial pilot run for each model without constraints to ensure that the resulting posteriors were sufficiently far away from zero. Zero constraints where added in accordance to numerous constructs being measured or to preserve battery structure in the case of models testing theories with a minimal number of constraints.

TABLE 4.4. Loadings Matrix Constraint Specification for Models of Racial Resentment (RR), Individualism (I), and Traditional Racism (TR) Survey Batteries from 1986 ANES.

Dims. Battery	D1	D2	D3
RR	✓	✗	✗
I	✓	✗	✗
TR	✓	✗	✗

(a) M1

Dims. Battery	D1	D2	D3
RR	✓	✗	✗
I	✗	✓	✗
TR	✗	✓	✗

(b) M2

Dims. Battery	D1	D2	D3
RR	✓	✗	✗
I	✗	✓	✗
TR	✓	✗	✗

(c) M3

Dims. Battery	D1	D2	D3
RR	✓	✗	✗
I	✓	✗	✗
TR	✗	✓	✗

(d) M4

Dims. Battery	D1	D2	D3
RR	✓	✗	✗
I	✗	✓	✗
TR	✗	✗	✓

(e) M5

Dims. Battery	D1	D2	D3
RR	✓	✓	✗
I	✓	✓	✗
TR	✓	✓	✗

(f) M6

Dims. Battery	D1	D2	D3
RR	✓	✓	✓
I	✓	✓	✓
TR	✓	✓	✓

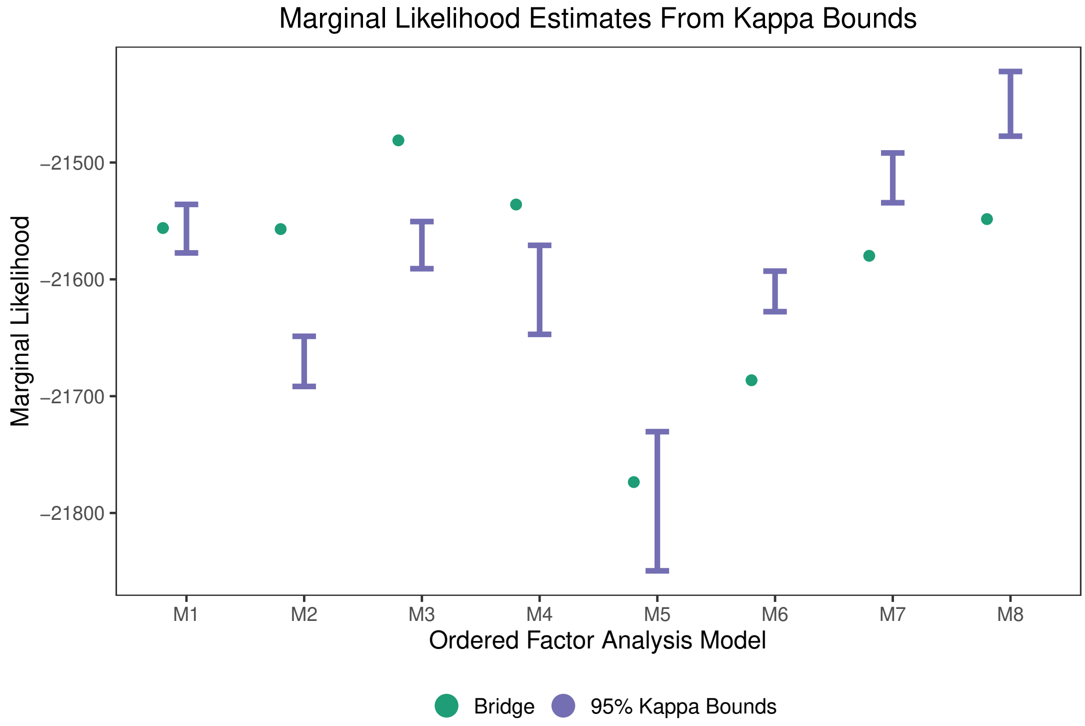
(g) M7

Dims. Battery	D1	D2	D3
RR	✓	✓	✓/✗
I	✓	✓	✗
TR	✓	✗	✓

(h) M8

Note: ✓indicates that the battery was included on a dimension while ✗indicates it was not included (i.e. the loading was restricted to be exactly equal to zero). ✓/✗indicates that the questions were split according to information gained from previous models. Specific details are given for M8 in the text. All specifications and code can be found in this paper's supplemental files and any clarification can be given by the author (i.e. me) on request.

FIGURE 4.6. Comparison of marginal likelihood estimates given by kappa bounds and bridge sampling for an ordered factor analysis model applied to questions related to racial resentment in the 1986 ANES.



Note: Bars denote the 95% HPD implied by 100,000 Monte Carlo samples. Specific model specifications can be found in Table 4.4.

heavily on how constraints are applied to the data. Interestingly, the worst performing model in both bridge sampling and kappa bounds is M5, which assumes that the racial resentment, individualism, and traditional racism batteries are conditionally independent of one another. While both estimates provide marginal likelihood estimates on the same scale as the kappa bounds, using bridge sampling to select a best model leads to a different conclusion that kappa bounds. Bridge sampling selects M3 as the best model, which implies that the racial resentment and traditional racism batteries belong on the same single dimension while the individualism battery is conditionally independent of these concepts. On the other hand, kappa bounds select M8, a more nuanced three dimensional model with specific zero constraints on the racial resentment battery. As simulations show that kappa bounds perform better in the ordered factor analysis scenario, there is reason to trust that M8 is the true best-fitting model. This situation shows how the choice of marginal likelihood approximation technique can have large implications for model selection problems.

The two best performing models according to kappa bounds are the unconstrained dimensional model and a three dimensional model with carefully selected constraints informed by numerous exploratory factor analysis models. This points to a conclusion that there is a statistical relationship between traditional racist attitudes, individualism, and the racial resentment battery beyond a simple one or two dimensional model; rather, the model evidence points to a complex three-way relationship between these concepts that has been noted in numerous studies. However, kappa bounds are able to distinguish between two models that only differ in a few constraints and is able to distinguish that more constraints are needed to improve on the unconstrained three dimensional model. Substantively, the improvement of M8 over M7 leads to two key insights: 1) When controlling for common covariance across all three batteries, there is little evidence that individualism and traditional racism include more shared covariance and 2) three questions from the racial resentment battery have additional sources of covariance that are explained by traditional racist attitudes. These findings echo those found by Carmines et al. (2011) and demonstrate that the survey batteries frequently used to measure racial resentment should be subject to more scrutiny.

4.7 Conclusion

In this paper, I present a new approach for estimating intervals on the marginal likelihood. Where previous methods of marginal likelihood estimation come with no information about the quality of the estimate, the new interval estimation method explicitly bakes quality into the interval width - low quality estimates have larger widths while high quality estimates provide narrow intervals and high certainty about the value of the marginal likelihood. Intervals derived from the key variational inequalities show good behavior, but include a wide set of values that cannot possibly exist because of the relationship between forward and reverse KL divergences. I derive a new bounding inequality on the ratio of these two quantities and use this to derive kappa bounds, a state-of-the-art interval on the marginal likelihood that is narrow and has good coverage properties on the true marginal likelihood. Compared to Monte Carlo estimation approaches, I demonstrate the ease of estimation and superiority of estimates given by kappa bounds in a number of simulated and real-world data settings.

Extensions to kappa bounds are numerous. A first extension uses kappa bounds to handle situations where exact draws from the joint posterior require prohibitively expensive computation. While the derivations in this paper assume that $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ can be known exactly, Grosse et al. (2015) present a novel bidirectional Monte Carlo scheme that combines the computational efficiency of variational approaches with a slightly more expensive sequential Monte Carlo scheme to model departures from the initial variational approximations. Where this approach seeks to sequentially maximize the ELBO and then minimize the EUBO, kappa bounds provide an approach fitting criteria can be conditioned on the sequence of kappa bounds across observations. With light changes, the refined variational inequalities presented in this paper can be used to estimate the marginal likelihood which can then be used as a fitting criteria - the algorithm can find the minimum width set of kappa bounds over all possible joint posteriors. This kappa criteria provides a novel fitting criteria that would seek to improve variational Bayesian estimation procedures.

A second extension adds the advancements in choosing approximations presented by Wang and Meng (2016). Throughout this paper, simple multivariate normal approximations are used. While kappa bounds perform better than other approaches using this approximation, the width of kappa bounds can be further reduced by increasing the quality of the approximation. More computationally efficient implementations of the posterior warping algorithms from Wang and Meng (2016) are one way that approximations can be improved. However, advancements in high dimensional kernel density estimation have made estimation of nonparametric densities in otherwise impossible settings more viable. Combined with advancements in estimation of high dimensional copula to handle complex dependency patterns between parameters in a model, the kernel density estimation approach can make approximation methods more accurate and, in turn, return better estimates of the marginal likelihood. The width of kappa bounds signifies the quality of the approximation, so kappa bounds can be used to assess whether or not these approaches are providing meaningfully better results than the parametric counterparts.

Bibliography

Basu, Mitra and Tin Kam Ho (2006). *Data complexity in pattern recognition.* Springer Science & Business Media.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference.* Ph. D. thesis, Gatsby Computational Neuroscience Unit, University College London.

Besag, Julian (1989). A candidate's formula: A curious result in bayesian prediction. *Biometrika 76*(1), 183–183.

Betancourt, Michael (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Carmines, Edward G , Paul M Sniderman, and Beth C Easter (2011). On the meaning, measurement, and implications of racial resentment. *The Annals of the American Academy of Political and Social Science 634*(1), 98–116.

Chérief-Abdellatif, Badr-Eddine (2019). Consistency of elbo maximization for model selection. In *Symposium on Advances in Approximate Bayesian Inference*, pp. 11–31.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association 96*, 270–281.

Corduneanu, A. and C. Bishop (2001). Variational bayesian model selection for mixture distributions. In J. a. T. (Ed.), *Richardson*, pp. 27–34. Morgan Kaufmann: Proceedings Eighth International Conference on Artificial Intelligence and Statistics.

DiCiccio, T. J. , R. E. Kass, A. Raftery, and L. Wasserman (1997). Computing bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association 92*, 903–915.

Dieng, Adji Bousso , Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei (2017). Variational inference via chi upper bound minimization. In *Advances in Neural Information Processing Systems*, pp. 2732–2741.

Eberly, Lynn E and George Casella (2003). Estimating bayesian credible intervals. *Journal of statistical planning and inference 112*(1-2), 115–132.

Feldman, Stanley and Leonie Huddy (2005). Racial resentment and white opposition to race-conscious programs: Principles or prejudice? *American Journal of Political Science 49*(1), 168–183.

Fong, Edwin and Chris Holmes (2019). On the marginal likelihood and cross-validation. *arXiv preprint arXiv:1905.08737*.

Forero, Carlos G , Alberto Maydeu-Olivares, and David Gallardo-Pujol (2009). Factor analysis with ordinal indicators: A monte carlo study comparing dwls and uls estimation. *Structural Equation Modeling 16*(4), 625–641.

Fruhwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal 7*, 143–167.

Gamerman, D. and H. F. Lopes (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. CRC Press.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society*, 501–514.

Gelman, Andrew , Jessica Hwang, and Aki Vehtari (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing 24*(6), 997–1016.

Gelman, A. and X.-L. Meng (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 163–185.

Geweke, J. F. and G. Zhou (1996). Measuring the pricing error of the arbitrage pricing theory. *Rev. Finan. Stud 9*, 557–587.

Geyer, Charles J (2011). Introduction to markov chain monte carlo. *Handbook of markov chain monte carlo 20116022*, 45.

Ghosh, Joyee and David B Dunson (2009). Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics 18*(2), 306–320.

Gronau, Quentin F , Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S Leslie, Jonathan J Forster, Eric-Jan Wagenmakers, and Helen Steingroever (2017). A tutorial on bridge sampling. *Journal of mathematical psychology 81*, 80–97.

Grosse, Roger B , Zoubin Ghahramani, and Ryan P Adams (2015). Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv:1511.02543*.

Hammersley, J. M. and D. C. Handscomb (1964). *Monte Carlo methods.* London: Methuen.

Humphreys, K. and D. Titterington (2000). *Approximate Bayesian inference for simple mixtures, in: Proceedings in Computational Statistics, COMPSTAT'2000.* Springer-Verlag.

Jaakkola, T. and M. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing 10*, 25–37.

Ji, Chunlin , Haige Shen, and Mike West (2010). Bounded approximations for marginal likelihoods. In *Technical report.* Citeseer.

Jordan, M. (2004). Graphical models. *Statistical Science 19*, 140–15.

Jordan, M. , Z. Ghahramani, T. Jaakkola, and K. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning 37*, 183–233.

Joreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika 32*, 443–482.

Joreskog, Karl G (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika.*

Kass, R. A. and A. E. Raftery (1995). Bayes factor. *J. Amer 90*, 773–795.

Kinder, Donald R , Lynn M Sanders, and Lynn M Sanders (1996). *Divided by color: Racial politics and democratic ideals.* University of Chicago Press.

Lewis, S. M. and A. E. Raftery (1997). Estimating bayes factors via posterior simulation with the laplace-metropolis estimator. *J. Amer 92*, 648–655.

Lopes, H. F. , P. Muller, and G. L. Rosner (2003). Bayesian meta-analysis for longitudinal data models using multivariate mixture priors. *Biometrics 59*, 66–75.

Lopes, Hedibert Freitas and Mike West (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 41–67.

MacKay, D. (1995). Developments in probabilistic modelling with neural networks ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications*, Nijmegen, Netherlands, pp. 191–198. Proceedings of the 3rd Annual Symposium on Neural Networks.

Malley, James D , Jochen Kruppa, Abhijit Dasgupta, Karen G Malley, and Andreas Ziegler (2012). Probability machines. *Methods of information in medicine 51*(01), 74–81.

Martin, Andrew D , Kevin M Quinn, and Jong Hee Park (2011). Mcmcpack: Markov chain monte carlo in r.

Meng, X.-L. and S. Schilling (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics 11*, 552–586.

Meng, X.-L. and W. H. Wong (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Mira, Antonietta and Geoff Nicholls (2004). Bridge estimation of the probability density at a point. *Statistica Sinica*, 603–612.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing 11*, 125–139.

Newton, M. A. and A. E. Raftery (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *J. Roy 56*, 3–48.

Overstall, A. M. and J. J. Forster (2010). Default bayesian model determination methods for generalised linear mixed models. *Computational Statistics and Data Analysis 54*, 3269–3288.

Pajor, A. (2016). Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Analysis*, 1–27.

Pradier, Melanie F , Michael C Hughes, and Finale Doshi-Velez (2019). Challenges in computing and optimizing upper bounds of marginal likelihood based on chi-square divergences.

Raftery, Adrian E (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Raftery, A. E. and J. D. Banfield (1991). Stopping the gibbs sampler, the use of morphology, and other issues in spatial statistics (bayesian image restoration, with two applications in spatial statistics)–(discussion). *Annals of the Institute of Statistical Mathematics 43*, 32–43.

Roos, J Micah , Michael Hughes, and Ashley V Reichelmann (2019). A puzzle of racial attitudes: A measurement analysis of racial attitudes and policy indicators. *Socius 5*, 2378023119842738.

Sason, Igal and Sergio Verdu (2016). f-divergence inequalities. *IEEE Transactions on Information Theory 62*(11), 5973–6006.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer 81*, 82–86.

Ueda, N. and Z. Ghahramani (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks 15*, 1223–1241.

Vandekerckhove, J. , D. Matzke, and E.-J. Wagenmakers (2015). Model comparison and the principle of parsimony. In J. T. Busemeyer, Z. J. Wang, and A. Eidels (Eds.), *J.* Oxford Handbook of Computational and Mathematical Psychology. Oxford: Oxford University Press.

Wang, L. and X.-L. Meng (2016). Warp bridge sampling: The next generation. arxiv. preprint.

Wang, Yixin and David M Blei (2019). Frequentist consistency of variational bayes. *Journal of the American Statistical Association 114*(527), 1147–1161.

Wright, Marvin N and Andreas Ziegler (2015). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

Zhang, Cheng , Judith Bütepage, Hedvig Kjellström, and Stephan Mandt (2018). Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence 41*(8), 2008–2026.

A Proof for Lemma 4.4.1

By the law of iterated expectations and $\alpha \perp\!\!\!\perp \mathcal{U} \perp\!\!\!\perp \mathcal{L}$:

$$\mathbb{E}[\Delta|a, b, \boldsymbol{\theta}^*, \boldsymbol{\omega}^*] = \mathbb{E}_\alpha[(1-\alpha)\mathcal{U} + \alpha\mathcal{L}] = \mathbb{E}[(1-\alpha)]\mathcal{U} + \mathbb{E}[\alpha]\mathcal{L}$$

Leveraging the same independence:

$$\mathbb{E}[\Delta|a, b, \boldsymbol{\theta}^*, \boldsymbol{\omega}^*] = \mathbb{E}[(1-\alpha)]\mathcal{U}_0 + \mathbb{E}[\alpha]\mathcal{L}_0$$

is an unbiased estimator for the posterior mean.

By the law of total variance:

$$\mathbb{V}[\Delta|a, b, \boldsymbol{\theta}^*, \boldsymbol{\omega}^*] = \mathbb{V}_\alpha[(1-\alpha)\mathcal{U} + \alpha\mathcal{L}] + \mathbb{E}_\alpha[(1-\alpha)^2\sigma_u^2 + \alpha^2\sigma_l^2]$$

where:

$$\begin{aligned}\mathbb{E}_\alpha[((1-\alpha)\mathcal{U} + \alpha\mathcal{L})^2] &= \mathcal{U}^2 + (2\mathcal{U}\mathcal{L} - 2\mathcal{U}^2)\mathbb{E}[\alpha] + (\mathcal{U}^2 - 2\mathcal{U}\mathcal{L} + \mathcal{L}^2)\mathbb{E}[\alpha^2]\\ \mathbb{E}_\alpha[(1-\alpha)\mathcal{U} + \alpha\mathcal{L}]^2 &= \mathcal{U}^2 + (2\mathcal{U}\mathcal{L} - 2\mathcal{U}^2)\mathbb{E}[\alpha] + (\mathcal{U}^2 - 2\mathcal{U}\mathcal{L} + \mathcal{L}^2)\mathbb{E}[\alpha]^2\\ \mathbb{E}_\alpha[(1-\alpha)^2\sigma_\mathcal{U}^2 + \alpha^2\sigma_\mathcal{L}^2] &= \mathbb{E}[(1-\alpha)^2]\sigma_\mathcal{U}^2 + \mathbb{E}[\alpha^2]\sigma_\mathcal{L}^2\end{aligned}$$

This gives a posterior variance of:

$$\mathbb{V}[\Delta|a, b, \boldsymbol{\theta}^*, \boldsymbol{\omega}^*] = ((\mathcal{U} - \mathcal{L})^2 + \sigma_\mathcal{U}^2 + \sigma_\mathcal{L}^2)\mathbb{V}[\alpha] + \sigma_\mathcal{U}^2\mathbb{E}[(1-\alpha)]^2 + \sigma_\mathcal{L}^2\mathbb{E}[\alpha]^2$$

In order to work with this value, we need a sample statistic for $(\mathcal{U} - \mathcal{L})^2$. A reasonable option is the asymptotically unbiased and consistent estimator:

$$\sigma_\mathcal{U}^2 + \sigma_\mathcal{L}^2 + (\mathcal{U}_0 - \mathcal{L}_0)^2 \xrightarrow{p} (\mathcal{U} - \mathcal{L})^2$$

Then, the posterior variance is:

$$\mathbb{V}[\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*] = \left(2\sigma_U^2 + 2\sigma_L^2 + (\mathcal{U}_0 - \mathcal{L}_0)^2\right)\mathbb{V}[\alpha] + \sigma_U^2\mathbb{E}[(1-\alpha)]^2 + \sigma_L^2\mathbb{E}[\alpha]^2$$

Assuming that N and M are large, the variance on the expectation of $\mathcal{U}$ and $\mathcal{L}$ can be dropped leading to:

$$\mathbb{V}[\Delta|a,b,\boldsymbol{\theta}^*,\boldsymbol{\omega}^*] = (\mathcal{U}_0 - \mathcal{L}_0)^2\mathbb{V}[\alpha]$$

B Proof for Theorem 4.5.1

I begin by formally defining an f-divergence:

Definition B.1. Let $f(t) : t \in (0,\infty) \to \mathbb{R}^+$ be a convex function such that $f(1) = 0$. Further assume that $f(t)$ is twice differentiable and that $f'(1) = 0$. Let $\mathcal{P}$ and $\mathcal{Q}$ be two proper probability distribution functions defined over the measurable space, $\boldsymbol{\Omega}$. Then, the f-divergence from $\mathcal{P}$ to $\mathcal{Q}$ as:

$$D_f(P||Q) = \int\limits_{\boldsymbol{\Omega}} f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q} \tag{4.43}$$

Forward and reverse KL divergence can be expressed as an f-divergence:

Definition B.2. Let $f_f(t) = 1 + t\log t - t$. Then:

$$\int\limits_{\boldsymbol{\Omega}} f_f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q} = KL(\mathcal{P}||\mathcal{Q}) \tag{4.44}$$

Let $f_r(t) = t - \log t - 1$. Then:

$$\int\limits_{\boldsymbol{\Omega}} f_r\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q} = KL(\mathcal{Q}||\mathcal{P}) \tag{4.45}$$

f-divergences represent a useful class of functions that formalize a notion of difference between two distributions. Most importantly, f-divergences are convex transformations of the density ratio of two distributions and a number of inequalities can be derived due to this convexity. One useful consequence of this convexity is a result of function domination (Sason and Verdu, 2016):

Lemma B.1. Let $f(t)$ and $g(t)$ be two convex functions as in Definition B.1. Let $\mathcal{P}$ and $\mathcal{Q}$ be two proper density functions that are absolutely continuous with respect to a common measurable space. If there exists some $\nu > 0$ such that $f(t) \leq \nu g(t) \forall t > 0$, then:

$$D_f(P||Q) \leq \nu D_g(P||Q) \tag{4.46}$$

Similarly, if there exists some $\nu > 0$ such that $\nu f(t) \leq g(t) \forall t > 0$, then:

$$\nu D_f(P||Q) \leq D_g(P||Q) \tag{4.47}$$

Proof. Since $f(t)$ is a strictly positive convex function, any integral of $f(t)$ is also convex. Therefore, the f-divergence is a convex function in $t = \frac{d\mathcal{P}}{d\mathcal{Q}}$. This also applies to $g(t)$. Since $f(t)$ and $g(t)$ are positive with only $t = 1 \to 0$, $D_f(P||Q)$ and $D_g(P||Q)$ must also be positive.

Let $\nu > 0$ be a value such that:

$$f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) \leq \nu g\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)$$

Because ν is independent of $\mathcal{P}$ and $\mathcal{Q}$ and by the convexity of $f(t)$ and $g(t)$:

$$\int f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q} \leq \nu \int g\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q}$$

The opposite condition is proven in the same way.

This inequality gives rise to an important result:

Lemma B.2. Assume $\mathcal{P}$ is absolutely continuous with respect to $\mathcal{Q}$ and vice versa. Further, to avoid the trivial case of equality, assume $\mathcal{P} \neq \mathcal{Q}$. Let $f(t)$ and $g(t)$ be two convex functions as in Definition B.1. Further, assume that $f(t), g(t) > 0 \; \forall \; t \in (0,1) \cup (1,\infty)$ and $f(1), g(1) = 1$ by left and right limits. Let $\kappa(t) : t \in (0,\infty) \to \mathbb{R}^+$:

$$\kappa(t) = \frac{f(t)}{g(t)}$$

be a continuous function in t. Define $\bar{\kappa}$:

$$\bar{\kappa} = \sup \frac{f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)}{g\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)}$$

Then:

$$D_f(\mathcal{P}||\mathcal{Q}) \leq \bar{\kappa} D_g(\mathcal{P}||\mathcal{Q}) \tag{4.48}$$

Similarly, define $\underline{\kappa}$:

$$\underline{\kappa} = \inf \frac{f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)}{g\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)}$$

Then:

$$\underline{\kappa} D_g(\mathcal{P}||\mathcal{Q}) \leq D_f(\mathcal{P}||\mathcal{Q}) \tag{4.49}$$

This leads to a sandwiching inequality:

$$\underline{\kappa} \leq \frac{D_f(\mathcal{P}||\mathcal{Q})}{D_g(\mathcal{P}||\mathcal{Q})} \leq \bar{\kappa} \tag{4.50}$$

Proof. Let $t = \frac{d\mathcal{P}}{d\mathcal{Q}}$. By definition:

$$\frac{f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)}{g\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)} \leq \bar{\kappa}$$

By Lemma B.1:

$$D_f(\mathcal{P}||\mathcal{Q}) \leq \bar{\kappa} D_g(\mathcal{P}||\mathcal{Q})$$

$\underline{\kappa} D_g(\mathcal{P}||\mathcal{Q}) \leq D_f(\mathcal{P}||\mathcal{Q})$ is shown in a similar way. Since $\mathcal{P} \ll\gg \mathcal{Q}$ and by the continuity of $f(t)$ and $g(t)$, the ratio is defined for all finite, positive values. Thus, dividing each side of the inequality by $D_g(\mathcal{P}||\mathcal{Q})$ is admissible.

Definition B.3. Assume $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \ll\gg \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \; \forall \; \boldsymbol{\theta} \in \boldsymbol{\Theta}$. To avoid the trivial case of equality, further assume that $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \neq \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$. Define the relative extrema for the likelihood ratio of $\mathcal{P}$ to $\mathcal{Q}$ as:

$$\begin{aligned} \beta_1 &= \sup_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} = \left(\inf_{\boldsymbol{\theta}} \frac{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})} \right)^{-1} \\ \beta_2 &= \inf_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} = \left(\sup_{\boldsymbol{\theta}} \frac{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})} \right)^{-1} \end{aligned} \tag{4.51}$$

where the inversion holds due to absolute continuity between $\mathcal{P}$ and $\mathcal{Q}$.

β_1 and β_2 are bounded. Specifically, $0 < \beta_2 < 1$ and $1 < \beta_1 < \infty$. By the shared domain requirement and absolute continuity of $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ and $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$, $\beta_1, \beta_2 > 0$. If $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ and $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$ are proper density functions, then $\inf_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$ must be less than 1 since $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \neq \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})$. The same logic holds to show that $\sup_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} > 1$.

This definition gives rise to a sandwiching inequality for the ratio of two f-divergences:

Lemma B.3. Assume $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \ll\gg \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma}) \; \forall \; \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Assume that β_1 and β_2 are defined. Then:

$$\kappa(\beta_2) \leq \frac{D_f(\mathcal{P}||\mathcal{Q})}{D_g(\mathcal{P}||\mathcal{Q})} \leq \kappa(\beta_1) \tag{4.52}$$

where $\kappa(t) = \frac{f(t)}{g(t)}$ is a continuous and monotonically increasing function in t.

Proof. Define β_1 and β_2 as shown in Definition B.3:

$$\beta_1 = \operatorname*{ess\,sup}_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$$
$$\beta_2 = \operatorname*{ess\,inf}_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$$

By the strict monotonicity of $\kappa(t)$:

$$\kappa(\beta_1) = \operatorname*{ess\,sup}_{\boldsymbol{\theta}}\ \kappa\left(\frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}\right)$$
$$\kappa(\beta_2) = \operatorname*{ess\,inf}_{\boldsymbol{\theta}}\ \kappa\left(\frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}\right)$$

which meets the extremum conditions outlined in Lemma B.2. Therefore, $\kappa(\beta_1)$ and $\kappa(\beta_2)$ sandwich the ratio of corresponding f-divergences.

Finally, show Theorem 4.5.1:

Theorem B.1. Assume $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X}) \ll\gg \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})\ \forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $t = \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})}$ and define:

$$\begin{aligned} f\,(t) &= t\log t + (1-t) \\ g(t) &= -\log t + (t-1) \end{aligned} \tag{4.53}$$

where $f(t)$ and $g(t)$ are f-generators for forward and reverse KL divergence, respectively.

Define $\kappa(t)$ as:

$$\kappa(t) = \frac{1 + t\log t - t}{t - \log t - 1} \tag{4.54}$$

Define the extrema on the ratio of $\mathcal{P}$ to $\mathcal{Q}$ as:

$$\begin{aligned} \beta_1 &= \operatorname*{ess\,sup}_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} \\ \beta_2 &= \operatorname*{ess\,inf}_{\boldsymbol{\theta}} \frac{\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})}{\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\gamma})} \end{aligned} \tag{4.55}$$

Then:

$$\kappa(\beta_2) \leq \frac{KL(\mathcal{P}||\mathcal{Q})}{KL(\mathcal{Q}||\mathcal{P})} \leq \kappa(\beta_1) \tag{4.56}$$

Proof. $f(t)$ and $g(t)$ are convex functions where $f(1) = f'(1) = 0$.

The first derivative of $\kappa(t)$ with respect to t is:

$$\frac{d\kappa(t)}{dt} = \frac{(t-1)^2 - t\log^2(t)}{t(t - \log t - 1)^2}$$

For $t \in (0,1) \cup (1,\infty)$, the derivative is always defined and finite since $t - \log t = 1 \iff t = 1$. The discontinuity at $t = 1$ can be ignored since $\lim\limits_{\epsilon\to 0} \kappa(1-\epsilon) = \lim\limits_{\epsilon\to 0} \kappa(1+\epsilon) = 1$, implying that $\kappa(1)$ is just a pointwise discontinuity that has no real bearing on the overall behavior of the function. Thus, $\kappa(t)$ is continuous for all t.

Second, it can be shown that $\kappa(t)$ is monotonically increasing in t. Here, it is sufficient to show that for all admissible values of t the derivative is always positive. First, note that the derivative approaches zero as t approaches 1 since the numerator equals zero when $t = 1$. Since the numerator and denominator of the derivative are always positive, this sufficiently demonstrates the function is strictly monotonically increasing in t.

Since $\kappa(t)$ meets the conditions of Lemma B.3, the inequality holds and leads to Equation (4.56).

Theorem B.1 is equivalent to the initial statement.

C Proof for Lemma 4.5.2

Let $T = N + M$ and let $T \to \infty$ imply $N \to \infty$ and $M \to \infty$ at equal rates. Let $\phi(f) = \mathbb{P}(y = 1|f \in \mathcal{R}, \boldsymbol{F})$ be the true probability that f comes from $\mathcal{P}(\boldsymbol{\theta}|\mathcal{X})$ given a T observation feature set, $\boldsymbol{F}$. Let $\hat{\phi}(f) = \{\hat{\phi}_1(f), \hat{\phi}_2(f), ..., \hat{\phi}_S(f)\}$ be a collection of S estimates of the classification probability and let $\bar{\phi}_S(f) = \frac{1}{S}\sum\limits_{s=1}^{S} \hat{\phi}_s(f)$. Assume that this mean is consistent to its true value:

$$\lim_{T \to \infty, S \to \infty} \bar{\phi}_S(f) = \phi(f)$$

Appealing to central limit theorem and assuming T is large:

$$\bar{\phi}_S(f) \xrightarrow{d} \mathcal{N}\left(\phi(f), \frac{\sigma^2_{\phi(f)}}{S}\right)$$

where $\sigma^2_{\phi(f)}$ is the variance of the estimate. Plugging in optimal estimators and rearranging:

$$\mathbb{P}(\phi(f)|\boldsymbol{F}) \sim \mathcal{N}\left(\bar{\phi}_S(f), \frac{\hat{\sigma}^2_{\phi(f)}}{S}\right)$$

where $\hat{\sigma}^2_{\phi(f)}$ is an unbiased estimator of the sample variance admitted by the S samples from $\hat{\phi}(f)$.

For any $f \in \mathcal{R}$, let $\Phi_S(f, g)$ be a distribution function in $0 < g < 1$ such that:

$$\Phi_S(f, g) = \int\limits_0^g \mathcal{N}\left(x \; ; \bar{\phi}_S(f), \frac{\hat{\sigma}^2_{\phi(f)}}{S}\right) dx$$

By convention, assume that $\Phi_S(f, 0) = 0$ and $\Phi_S(f, 1) = 1$. Then:

$$\mathbb{P}(\max \phi(f) < g) = \prod_{f \in \mathcal{R}} \Phi_S(f, g)$$

$$\mathbb{P}(\min \phi(f) < g) = \prod_{f \in \mathcal{R}} (1 - \Phi_S(f, g))$$

By the well-established consistency of the extreme order statistics, with large S and T:

$$\frac{\partial}{\partial g}\mathbb{P}(\max\phi(f) < g) \underset{p}{\rightarrow} \sup_f \phi(f)$$
$$\frac{\partial}{\partial g}\mathbb{P}(\min\phi(f) < g) \underset{p}{\rightarrow} \inf_f \phi(f)$$

Define two functions:

$$u(x) = \frac{x}{1-x} \;;\; v(y) = \frac{1 + y\log y - y}{y - \log y - 1}$$

where $0 \leq x \leq 1$ and $0 \leq y$. Define the edge cases such that $u(0) = 0$, $u(1) = \infty$, $v(0) = 0$, $v(1) = 1$, $v(\infty) = \infty$. Note that $u(x)$ and $v(y)$ are continuous, monotonically increasing, one-to-one mappings of x and y, respectively. Let $f^* = v(u(\phi(f)))$ and $g^* = v(u(g))$. By continuous mapping theorem:

$$\frac{\partial}{\partial g^*}\mathbb{P}(\max f^* < g^*) \underset{p}{\rightarrow} \sup_f \, v(u(\phi(f))) = \kappa(\beta_1)$$
$$\frac{\partial}{\partial g^*}\mathbb{P}(\min f^* < g^*) \underset{p}{\rightarrow} \inf_f \, v(u(\phi(f))) = \kappa(\beta_2)$$

meaning that $\mathbf{\Phi}_S(f, g) = \mathbf{\Phi}_S(f^*, g^*)$ and, as a consequence, $v\left(u\left(\sup_f \phi(f)\right)\right) = \sup_f v\left(u\left(\phi(f)\right)\right)$.

Assuming flat priors on $\kappa(\beta_1)$ and $\kappa(\beta_2)$:

$$\begin{aligned}\mathbb{P}(\kappa(\beta_1)|\boldsymbol{\theta}^*, \boldsymbol{\omega}^*) &\sim \frac{\partial}{\partial g^*}\mathbb{P}(\max f^* < g^*)\\ \mathbb{P}(\kappa(\beta_2)|\boldsymbol{\theta}^*, \boldsymbol{\omega}^*) &\sim \frac{\partial}{\partial g^*}\mathbb{P}(\min f^* < g^*)\end{aligned} \tag{4.57}$$