Bayesian Semiparametric Structured Additive Regression Models for Quantitative Historical Social Scientific Inquiry

Jonah Gabry

April 7, 2015

Contents

Chapter 1

Introduction

Wawro and Katznelson (2013) highlight a divide between historians and quantitative political scientists in regards to the historical study of political institutions. To account for concerns voiced by qualitative researchers, Wawro and Katznelson (henceforth W&K) propose a novel quantitative approach, "rejecting [. . .] the idea that one must choose between historical depth and quantitative rigor" (W&K, 2013).

W&K emphasize several areas that the commonly employed quantitative methods fail to adequately address: temporality, periodicity, context and specificity.[1] To tackle these issues, they advocate the implementation of Bayesian semi-parametric mixed models.[2] Their claim is that, compared to the typical models used in political science, these models can not only better account for parameter variation and unobserved heterogeneity but also limit the instability of estimation when the ratio of parameters to data is large (W&K, 2013). Furthermore, the Bayesian

[1] Note that here periodicity is not used in the mathematical sense but rather refers how relationships between variables are influenced differently by break points in history at particular locations and times.

[2] Briefly, Bayesian semi-parametric mixed models can be thought of as extending Bayesian generalized linear models (GLMs) to allow the additive predictor to incorporate metric or spatially correlated independent variables with unknown nonlinear effects (Fahrmeir and Lang, 2000).

framework requires the specification of prior distributions that can be tailored to reflect important assumptions or results from previous findings. In particular, priors like Gaussian Markov random fields can be used to model temporal and spatial relationships between parameters.

W&K apply their method in two replications of studies from the subfield of American Political Development (APD). While the results are promising, there is still much work to be done to support their larger claims regarding the value of such models for historical social scientific research also outside of APD. Working with Wawro – I?m currently his research assistant – I hope to use my thesis to help complete some of this work.

There is also the practical problem of computation, a topic W&K omit from their discussion. W&K use software called BayesX, which is appropriate and very efficient for the particular replications they choose, but offers limited flexibility in the choice of priors and likelihoods. To make the recommendations of W&K applicable in the broader context of social scientific inquiry will require software with greater flexibility. Goodrich et al. (2012) advocates the Bayesian software Stan (under development at Columbia), and I aim to use Stan to develop a framework for applying W&K's approach to a broader range of models than is currently possible with BayesX. Without such a framework it will remain unclear whether the recommendations of W&K can be extended beyond the confines of their replications.

In Cox and Katz as well as the other replications I intend to work on, accounting for parameter variation over time (and often space) entails estimating a large number of parameters. In many cases this results in a high parameter to data ratio, which presents greater challenges for maximum likelihood estimation – e.g. under-identification – than for the Bayesian approach of W&K. But the point is not only that the Bayesian models described here can overcome this problem, but moreover that they allow us to reframe the problems as an opportunity. Bayesian

semi-parametric mixed models with historically relevant smoothing priors can allow for greater parameter heterogeneity and let the data play a larger role in guiding our inferences.

Chapter 2

Literature review

2.1 Hierarchical Bayesian models

A hierarchical Bayesian model is simply a model written in hierarchical form and estimated via Bayesian methods. Essentially, such a model can be viewed as the combination of a within-unit model – e.g., describing feature of individuals over time – and an across-unit model – e.g., describing heterogeneity across individuals. An alternative yet mathematically equivalent way of conceiving of these models is as an infrastructure linking two or more layers of unknowns: the quantities we are interested in estimating (*process variables*) and the quantities introduced in developing our model (*model parameters*)

DATA | PROCESS, PARAMETERS

PROCESS | PARAMETERS

PARAMETERS

where the notation $A|B$ means A conditional on B (wikle_bried_2004). The link between the levels of the hierarchy is Bayes' rule, which states that for data y with likelihood $p(y|\theta)$ and parameter(s) θ with prior density $p(\theta)$ the posterior distribution of θ given the data y is proportional to the product of the prior θ and likelihood:

$$p(\theta|y) \propto p(\theta)p(y|\theta).$$

Bayes' rule provides the framework for updating beliefs about parameters θ based on the observed data y. With observations at the base level of the hierarchy it is possible to use Bayes' rule to find the conditional distributions for the quantities/parameters of interest at the upper levels of the hierarchy (Gelman et al., 2013).

For example, imagine a series of J experiments with outcomes $\{y_j : j = 1,\ldots,J\}$. We want to make inferences about the parameter θ_j underlying the data generating process for y_j. Conditional on the value of the unknown parameter θ_j we have a likelihood $p(y_j|\theta_j)$ for y_j parameterized by θ_j. For expositional simplicity, we can assume that parameters $\theta = (\theta_1,\ldots,\theta_J)'$ share a common prior distribution with hyperparameter ϕ. Now, we can either fix ϕ, or we can estimate ϕ by giving it its own prior distribution (a *hyperprior*) $p(\phi|\alpha)$, where α is the parameter (or parameters). If we make the choice to estimate ϕ then we have a hierarchical model in that we have added another level of variability to the model that is absent if ϕ is assumed to be known. We can extend the hierarchy to an arbitrary number of levels by assigning a prior distribution to α and the parameters of that distribution, etc. If α is assumed to be fixed then model is then

$$\begin{aligned} y_j &\sim p(y_j|\theta_j) \\ \theta_j &\sim p(\theta_j|\phi) \\ \phi &\sim p(\phi|\alpha). \end{aligned}$$

A thorough introduction to Bayesian hierarchical models can be found in Gelman et al. (2013).

2.2 Gaussian Markov random fields

The prior distributions of particular interest in this thesis are known as Gaussian Markov random field (GMRF) priors. The literature on GMRFs is vast, as they are frequently used in image processing and spatial statistics, however GMRFs appear only rarely in quantitative social science and Wawro and Katznelson (2014) is the first example extending the applications of GMRFs to historical social scientific inquiry. Before defining GMRFs it is first necessary to introduce some basic concepts from graph theory, particular the idea of an undirected graph.

2.2.1 Undirected graphs

An undirected graph G is simply an ordered pair of sets (V,E) containing the vertices (also called nodes) and edges of the graph, respectively. In what follows assume that V and E are finite sets and let $e_{ij} \in E$ denote the element in E that corresponds to the edge connecting the vertices i and j in V. The term *undirected* refers to the fact that the edges lack orientation.[1] The *neighbors* of a node $j \in V$ are all nodes $i \neq j$ such that $e_{ij} \in E$. We will denote the set of neighbors of node j by ∂_j. For a more formal and thorough discussion of undirected graphs see Rue and Held (2005).

[1] Intuitively, it is helpful to think of the edges of an *undirected* graph as line segments with no implied direction, whereas the edges in a *directed* graph can be thought of as arrows.

2.2.2 GMRFs

A normally distributed random vector $\theta = (\theta_1, \ldots, \theta_k) \sim \mathcal{N}_k(\mu, \Sigma)$ is said to be a GMRF with respect to a graph $G = (V, E)$ if each element in θ has a corresponding node in V and there is no edge between a pair of nodes i and j *if and only if* θ_i and θ_j are conditionally independent given all other elements of θ.

The relationship between G and θ – the information they provide about each other – is fully contained in the covariance matrix Σ. This must be the case, since from the mean vector μ nothing can be inferred about conditional independencies among the elements of θ, but it is not obvious from the individual elements of Σ. It is more useful to instead consider the precision matrix $Q = \Sigma^{-1}$, for which it can be shown that

$$\forall i \neq j,\ v_i \in \partial_j \iff q_{ij} = 0,$$

which is to say that conditional independence between θ_i and θ_j (no edge between nodes i and j) always corresponds to a zero in cell ij of the precision matrix, and vice-versa. For a simple proof of this result Rue and Held (2005).

Additionally, the following interpretations of the elements of a precision matrix will be helpful to keep in mind:

Off-diagonal For $i \neq j$, $q_{ij} = -Cov(\theta_i, \theta_j | \theta_{-ij})$, the negative (flipped sign) covariance between θ_i and θ_j conditional on all θs except i and j.

Diagonal For $i = j$, $q_{ij} = Var(\theta_i | \theta_{-i})$, the variance of θ_i conditional all the variables except θ_i.

2.3 Bayesian STAR models

An extension of generalized linear models (GLMs), semiparametric structured additive regression (STAR) models replace the linear predictor with a structured additive predictor of the form

$$\eta = f_1(x_1) + \ldots + f_j(x_j) + \ldots + f_J(x_J) + u'\gamma,$$

which can include nonlinear unknown functions of covariates as well as linear components (Fahrmeir & Lang, 2001). From the Bayesian perspective, any fixed effects parameters γ as well as the unknown functions $f_1, \ldots, f_J$ are treated as random variables distributed according to priors that we must specify. More specifically, if we have N observations, then for an unknown function f_j of a covariate x_j we treat the vector of evaluations $\mathbf{f}_j^{eval} = \big(f_j(x_{1j}), \ldots f_j(x_{Nj})\big)'$ as a random vector. Following Brezger and Lang (2006), we can conviently express $\mathbf{f}_j^{eval}$ as the matrix product $\mathbf{M}_j\theta_j$ of a design matrix $\mathbf{M}_j$ and a parameter vector θ_j. GMRF priors for each parameter vector θ_j take the general form

$$p(\theta_j|\tau_j^2) \propto \frac{1}{\left(\tau_j^2\right)^{\text{rank}(\mathbf{P}_j)/2}} \exp\left\{-\frac{1}{2\tau_j^2}\theta_j'\mathbf{P}\theta_j\right\}$$

where $\mathbf{P}$ is a penalty matrix whereby we operationalize our prior assumptions about the smoothness of the unknown function f_j.

To provide some intuition, suppose we are interested in learning about an effect with assumed variation over geographic regions. A simple map with R distinct regions $r_1, r_2, \ldots r_R$ can be represented as undirected graph G with vertices $V = \{1, 2, \ldots, R\}$ and edges connecting vertices corresponding to neighboring regions. To express prior beliefs about the spatial depen-

dence of the quantity of interest between regions we can construct a prior on a parameter vector $\theta = (\theta_1, \ldots, \theta_r)$ such that the matrix $\mathbf{P}$ has a zero in the ijth cell if and only if θ_i and θ_j are assumed to be conditionally independent given θ_{-ij}. Then θ is a GMRF with respect to G with $\Sigma^{-1} = \mathbf{Q} = \mathbf{P}/\tau^2$.

2.3.1 The penalty matrix

More specifically, the penalty matrix $\mathbf{P}$ can be constructed as $\mathbf{P} = \mathbf{D} - \mathbf{A}$, where $\mathbf{A}$ is a symmetric matrix with $a_{ij} = 1$ if temporal or spatial units i and j are considered neighbors (the associated graph G has an edge between nodes i and j) and 0 otherwise, and $\mathbf{D}$ is a diagonal matrix such that $\forall i = j$, $d_{ij} = \sum_j a_{ij}$. The matrices $\mathbf{A}$ and $\mathbf{D}$ are commonly referred to as the adjacency and degree matrices because encoded in $\mathbf{A}$ are all neighbor relationships (temporal or spatial) and the diagonal elements of $\mathbf{D}$ are the number of neighbors (the degree) of each vertex in the graph G.

To illustrate why this form of $\mathbf{P}$ captures these particular assumptions, consider N measurements of a variable x, with each measurement of x made at one of T evenly spaced points in time. For simplicity, but without loss of generality, assume that x is a unit of time and there is exactly one measurement per time period. The sequence $(x^{[t]})_{t=1}^{T}$ then corresponds to a grid of points on a line. Fahrmeir and Lang (2001) suggest several possible choices for a prior on a smooth function $f(x)$, the simplest of which is a first order random walk (RW_1) prior. Under the RW_1 prior, the first differences

$$\Delta_t = f(x^{[t]}) - f(x^{[t-1]})$$

are treated as independent and identically distributed standard normal random variables.

While this formulation of the RW_1 prior is *directed*, conditioning also on $f(x^{[t+1]})$ – one step into the future – forms an undirected RW_1, where the neighbors of time t are both $t-1$ and $t+1$. The associated graph G therefore has vertices $V = \{v_t : t = 1, \ldots, T\}$, each of which has two neighbors, with the exception of v_1 and v_T, which have one neighbor. The penalty matrix $\mathbf{P}$ corresponding to the RW_1 prior with equally spaced observations is the tridiagonal matrix

$$\mathbf{P} = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}$$

which can be derived by computing the difference of the appropriate degree and adjacency matrices. An extension of the RW_1, the RW_2 prior used later in this thesis also considers the measurements $x^{[t-2]}$ and $x^{[t+2]}$ to be neighbors of $x^{[t]}$.

The construction of $\mathbf{P}$ as the matrix difference $\mathbf{D} - \mathbf{A}$ also allows the estimation of an optional scalar parameter $\omega \in [0, 1]$, which, as a coefficient on $\mathbf{A}$, can be interpreted as representing the strength of spatial or temporal dependence (Rue & Held, 2005). The interpretation of ω and its role in estimation will be discussed further in the later sections.

2.3.2 Hyperpriors

To complete the hierarchical model we must also specify hyperpriors $p(\tau_j^2)$. Assigning a prior distribution to the variance hyperparameter τ_j^2 allows for the simultaneous estimation of a smoothing function f_j and the amount of smoothness, which is governed by τ_j. Fahrmeir and Lang (2001) and Brezger, Kneib, and Lang (2005) recommend a weakly informative but proper inverse-gamma prior on τ_j^2. However, in light of concerns about this type of inverse-gamma prior

raised by Gelman (2006), in this thesis several priors for τ_j are used and compared to check the degree to which the results are sensitive to the choice of prior.

Chapter 3

Empirical example: Cox and Katz (2007)

To demonstrate the benefits of the methods proposed by Wawro and Katznelson (2014), we conduct a reanalysis of Cox and Katz (2007), a study of bias and responsiveness in congressional roll-call votes in the 46th through 106th US Congresses. Cox and Katz's findings suggest persistent bias towards the majority party with particularly elevated levels during the periods from 1889-1910 and 1961-2000, known as the period of "czar rule" and the post-packing era, respectively. Analyzing the data using Bayesian STAR models we find much weaker evidence for bias towards the majority over the time periods in question.

3.1 Definitions

Bias In the legislative context of Cox and Katz's analysis, bias is defined as an advantage for one party in the efficiency with which its votes translate into legislative victories. For example, consider a majority party with only a small advantage in the number of seats its members occupy. If the majority is well-organized and unified it may win a much larger proportion of votes than its

slim seat advantage would typically suggest.

Responsiveness MISSING DEFINITION

3.2 Data

The data include – with a few exceptions detailed below – all roll-call votes in the U.S. House of Representatives during the 46th through 106th Congresses, corresponding to the period from 1879 to 2001. Following Cox and Katz, excluded from the analysis are any records that meet at least one of the following three conditions:

- The majority of both parties voted for the same position
- The purpose of the vote was electing the Speaker of the House
- The vote required a two-thirds majority for passage.

The resulting data consist only of votes that required a simple majority for passage and on which the Republican and Democratic parties were in clear opposition.

Let RC_{it} denote the result of roll-call vote i in Congress t such that

$$RC_{it} = \begin{cases} 1, & \text{if Democratic position wins vote } i \text{ in Congress } t \\ 0, & \text{if Republican position wins vote } i \text{ in Congress } t. \end{cases}$$

Here, the Democratic position refers to the outcome preferred by the majority of Democrats. The definition is analogous for the Republican position.

The outcomes of interest are the number of wins (w_t^{DEM}) and proportion of wins (p_t^{DEM}) for the Democratic position in each Congress t

$$w_t^{DEM} = \sum_{i=1}^{n_t} RC_{it}, \qquad p_t^{DEM} = w_t^{DEM}/n_t.^{1}$$

Here n_t is the number of roll-call votes in Congress t, which ranges from a minimum of 33 votes in the 70th Congress (1927-1929) to a maximum of 836 votes in the 104th Congress (1995-1997). The median number is 143 votes.

The sole predictor of interest is v^{RATIO}, the ratio of the average vote share earned by the Democratic position to the average vote share earned by the Republican position in each Congress. For each Congress the average is taken over all roll-call votes. For vote i in Congress t, let π_{it}^{DEM} be the proportion of representatives who cast their vote in favor of the Democratic position. Then the mean vote shares for the Democratic and Republican positions in Congress t is

$$v_t^{DEM} = \frac{1}{n_t}\sum_{i=1}^{n_t} \pi_{it}^{DEM}, \qquad v_t^{REP} = 1 - v_t^{DEM}$$

and the ratio of Democratic to Republican vote shares is simply $v_t^{RATIO} = v_t^{DEM}/v_t^{REP}$. Figure 3.1 shows $\log(v^{RATIO})$ plotted against p^{DEM}. The shape of the curve is similar to the standard seats-votes curve used in analyses of bias and responsiveness in electoral contexts. The curve is analogous in the legislative context of Cox and Katz's example, although we are not concerned with seat shares in a given Congress but rather roll-call vote shares (p_t^{DEM} and p_t^{REP}). This is discussed further in the Methods section.

The trends of $\log(v^{RATIO})$ and p^{DEM} over time are shown in the visual summary of the data set in Figure 3.2.

[1] The superscript *MAJ* (e.g. w_t^{MAJ}) will be used later in the thesis as a general way of referring to the majority party in a specific time period. While the data is organized and more naturally described in terms of w_t^{DEM} and w_t^{REP} (which is just $n_t - w_t^{DEM}$), the proposed statistical model is more straightforward to implement in terms of $w_t^{MAJ.}$

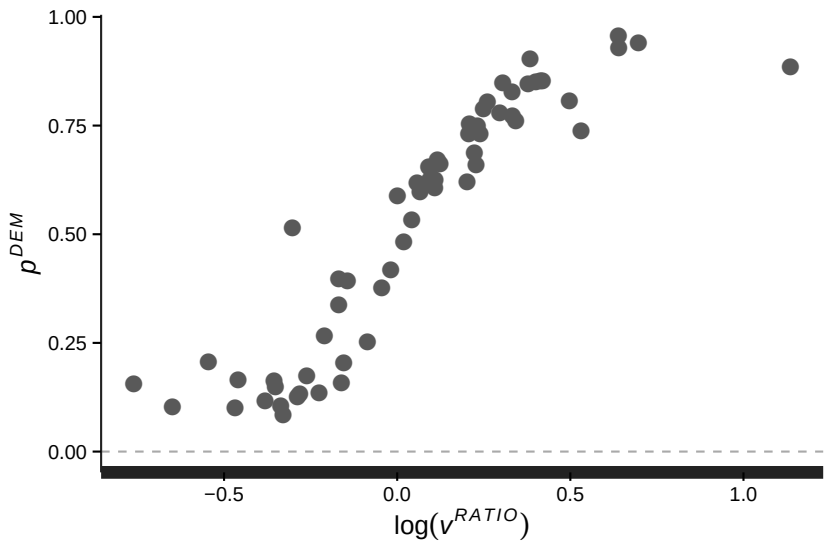

Figure 3.1: $\log(v^{RATIO})$ vs. p^{DEM}

One method of identifying bias toward or against the majority party is to estimate $E[p_t^{MAJ}|v_t^{MAJ} = 0.5]$, the proportion of majority party victories conditional on equal vote share, and compare the estimate to 0.5, the expected proportion of majority party victories in the absence of bias. There are many close votes in the data – that is, votes where $v_t^{MAJ} \approx 0$ – which can be used to estimate $E[p_t^{MAJ}|v_t^{MAJ} = 0.5]$. Figure 2, below, shows the proportion of majority party victories under four different definitions of a close votes corresponding to margins of victory of 0.125%, 0.25%, 0.5%, and 1.0%.

FIGURE

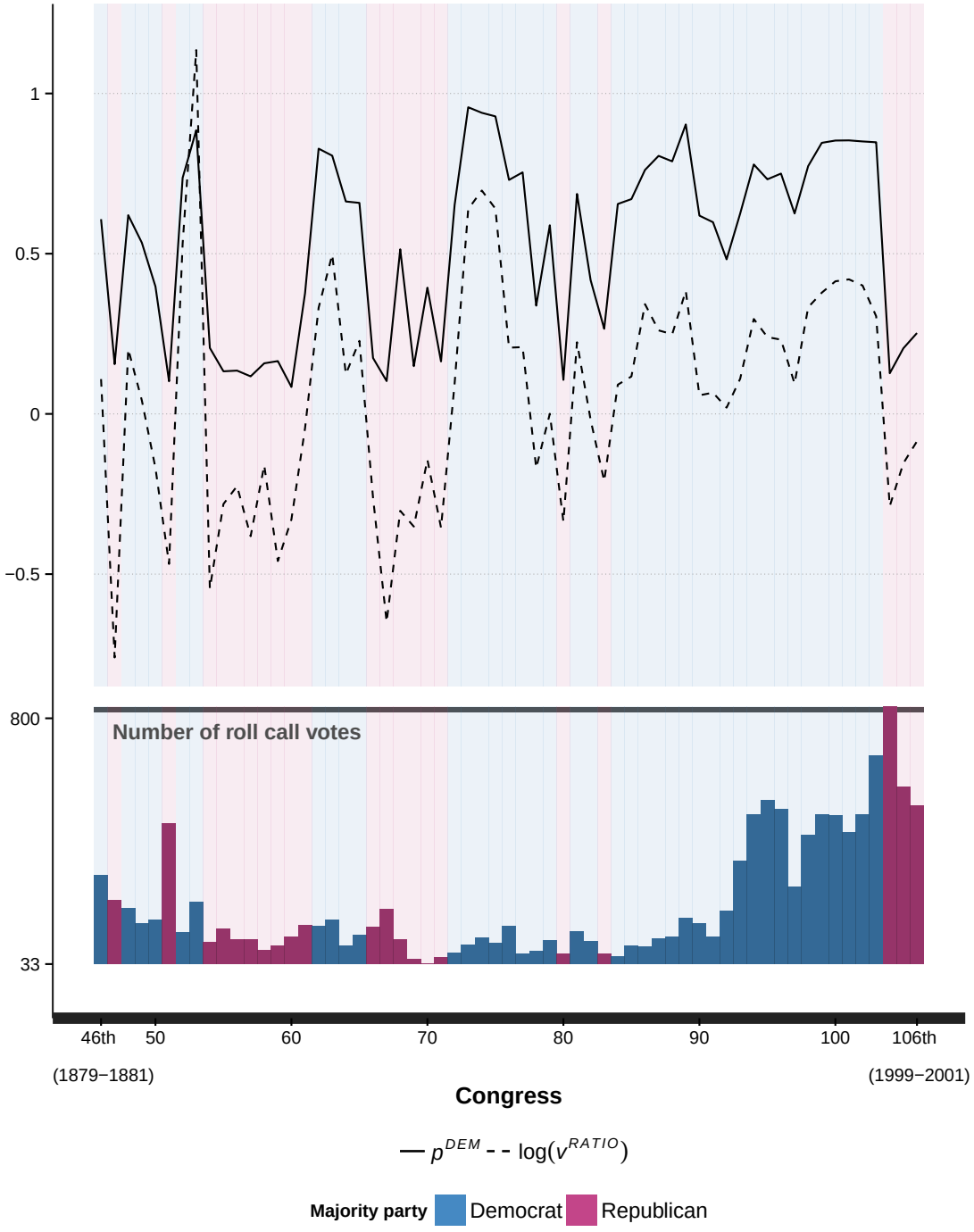

Figure 3.2: Visual summary of the data

3.3 Methods of Analysis

3.3.1 Methods

One of the analyses conducted by Cox and Katz concerns the estimation of parameters $\lambda = (\lambda_t : t = 46, \ldots, 106)$ representing bias (on the logit scale) toward the majority party in each C_t (Congress t), which is done by maximum likelihood estimation of grouped logit models with linear predictor

$$\lambda_t + \rho_t \log\left(v_t^{RATIO}\right).$$

Here the parameter ρ represents responsiveness, as defined above. The estimation of λ and ρ by grouped logit models follows naturally from solving the seats-votes equation for the average seat share, which in the legislative context of Cox and Katz's example is p^{DEM}, the expected roll-call win share for the Democrats

$$E(p_t^{DEM}) = \left(1 + \exp\left\{-\lambda_t - \rho_t \log\left(v_t^{RATIO}\right)\right\}\right)^{-1},$$

which is the familiar logistic function.

Cox and Katz's strategy is to estimate what they call "a sort of running average" of bias across time (p. 116). To do this they take as their estimate of λ_t the average estimate of λ over the seven congresses centered at t, the set $\{C_\tau, t-3 \leq \tau \leq t+3\}$. The authors attempt to account for temporal dependence in λ but their approach requires reusing the data to estimate models for each Congress and the observations for each Congress are used up to seven times. Goodrich, Katznelson, and Wawro (2012) point out that recycling data in this fashion can lead to overly precise parameter estimates.

Although Cox and Katz do not acknowledge this potential for exaggerated precision, they do call attention to another important concern. To obtain reasonable estimates, their method requires a nontrivial amount of variation in average vote share (v^{RATIO}) between the seven Congresses that comprise each set C_τ.

The analysis proposed in thesis overcomes both of these concerns by employing a hierarchical Bayesian framework with partial pooling. Following Cox and Katz, a BetaBinomial likelihood is used, however the linear predictor is replaced by the structured additive predictor η

$$w_t^{MAJ} | n_t, \alpha_t, \beta_t \sim \text{BetaBinomial}(n_t, \alpha_t, \beta_t),$$

$$\alpha_t = \theta_t \phi, \qquad \beta_t = \theta_t (1 - \phi),$$

$$\log\left(\frac{\theta_t}{1-\theta_t}\right) = \eta_t = f_\lambda(\lambda_t) + f_\rho\left(\log(v_t^{RATIO})\right).$$

The BetaBinomial(α, β) distribution can be thought of as a compound distribution resulting from a binomial distribution where the probability parameter follows a Beta(α, β) distribution. In other words, rather than assuming that the Congress-by-Congress probabilities of a majority party roll-call victory are independent and identically distributed – in which case the binomial distribution would suffice – the BetaBinomial allows for direct modeling of the variation in the probability of victory through the Beta distribution. The parameters α and β govern the shape of the Beta distribution, however it is both more intuitive and computationally attractive to reparameterize in terms of the mean θ, which requires introducing a parameter ϕ as the sum of α and β. This parameterization allows for logit(θ) to be estimated by the semi-parametric structured additive predictor η of the STAR model.

Instead of Cox and Katz's "sort of running average" approach, the hierarchical Bayesian STAR model entails estimating the unknown functions f_λ and f_ρ. As we've seen, the vector of evaluations of the unknown functions can be conveniently expressed as the products

$$\mathbf{f}_\lambda^{eval} = \lambda \mathbf{M}_\lambda, \qquad \mathbf{f}_\rho^{eval} = \rho \mathbf{M}_\rho,$$

of the parameter vectors λ and ρ of length 61 (the number of Congresses in the data) and $N \times 61$ design matrices $\mathbf{M}_\lambda$ and $\mathbf{M}_\rho$ (where N is the total number of observations in the data).

The the structured additive predictor η can be written compactly in vector notation as

$$\eta = f_\lambda(\lambda) + f_\rho(\log(v^{RATIO})) = \lambda \mathbf{M}_\lambda + \rho \mathbf{M}_\rho.$$

The matrices $\mathbf{M}_\lambda$ and $\mathbf{M}_\rho$ are identical in structure but not content. Since λ plays the role of an intercept – that is, the parameters in λ are not coefficients – the elements of $\mathbf{M}_\lambda$ are zeros and ones indicating the Congress to which each observation pertains. For $\mathbf{M}_\rho$ each element is either a zero or the appropriate value of $\log(v^{RATIO})$.

The GMRF priors for f_λ and f_ρ are expressed as prior distributions for the parameter vectors λ and ρ as

$$p(\lambda|\tau_\lambda^2) \propto \exp\left(-\frac{1}{2\tau_\lambda^2}\lambda' \mathbf{P} \lambda\right), \qquad p(\rho|\tau_\rho^2) \propto \exp\left(-\frac{1}{2\tau_\rho^2}\rho' \mathbf{P} \rho\right)$$

where the penalty matrix $\mathbf{P}$ encodes assumptions about the temporal dependence between Congresses.[2] To model temporal dependence such that the set of Congresses $\partial_t = \{C_{t-2}, C_{t-1}, C_{t+1}, C_{t+2}\}$ provides information about C_t, the undirected RW_2 prior discussed earlier is used to penalize deviations from this hypothesized trend. This corresponds to constructing $\mathbf{P}$ from the adjacency

[2]The impropriety of this prior stems from the fact that the matrix $\mathbf{P}$ is not full rank. The computational challenges this presents can be overcome by coding the model in a statistically equivalent but less intuitive form.

matrix $\mathbf{A}$ with $a_{ij} = 1$ if $C_j \in \partial_i$ and 0 otherwise.

3.3.2 Estimation

For notational convenience, let y denote the outcome w^{MAJ} and X denote the observed data n and v^{RATIO}. We require the joint posterior distribution

$$\begin{aligned} p(\lambda, \rho, \tau_\lambda, \tau_\rho, \phi | y, X) &\propto p(\lambda, \rho, \tau_\lambda, \tau_\rho, \phi) p(y | \lambda, \rho, \tau_\lambda, \tau_\rho, \phi, X) \\ &= p(\phi) p(\tau_\lambda) p(\tau_\rho) p(\lambda | \tau_\lambda) p(\rho | \tau_\rho) \prod_i p(y_i | \eta_i, X_i) \end{aligned}$$

where the second line follows from assumptions conditional independence.[3,4]

Estimation is performed numerically by Markov chain Monte Carlo (MCMC) methods and implemented via RStan, the R interface to the probabilistic programming language and C++ library Stan (Stan Development Team, 2015a).[5] Estimating the parameters of the STAR model with GMRF priors is not easy. Since this is the first attempt to estimate these models in Stan that we are aware of, the next section is a description of the estimation strategy and the code required to carry it out in Stan.

3.3.3 Stan

Stan is a probabilistic modeling language, MCMC sampler, and optimizer. The particular MCMC algorithm implemented in Stan is a variant of Hamiltonian Monte Carlo (HMC) called the no-U-turn sampler (NUTS) (Hoffman & Gelman, 2012). Borrowing from physics the con-

[3] In particular we assume that hyperparameters are mutually independent in their priors, $p(\lambda|\tau_\lambda)$ and $p(\rho|\tau_\rho)$ are conditionally independent, and the observations y_i are independent conditional on parameters and predictors.

[4] The absence of the previously mentioned parameters α, β, and θ from the expression for the posterior distribution is due to the fact that their values are determined by the other parameters.

[5] All relevant R and Stan code will be made publicly available in a repository on GitHub.

cepts and mathematics behind Hamiltonian dynamics, HMC treats the vector of unknown parameters as the position of a particle. In Hamiltonian dynamics, momentum and position change continuously over time, with the gradient of the particle's potential energy function – which corresponds to the negative log posterior – responsible for changes in momentum and momentum governing changes in position. Stan works by simulating a discretization of this process, making necessary corrections to preserve detailed balance (i.e., to ensure that the resulting Markov chains are reversible).

Several characteristics of HMC, and in particular Stan's implementation of HMC, make it a more appealing choice than traditional Metropolis-Hastings (M-H) and Gibbs samplers in many cases. Both M-H and Gibbs samplers suffer from random walk behavior that leads to inefficient exploration of the parameter space. Using gradient information, HMC can find posterior modes much more efficiently, greatly reducing the number of iterations required to obtain a sufficient number of effective draws from the posterior.

M-H samplers in particular also require a great deal of tuning from the user. Although HMC itself does not overcome this problem, Stan's automatically takes care of the tuning during a warmup period before sampling. Gibbs sampling has an advantage over MH in that it does not require tuning, but a serious drawback is that Gibbs sampling requires the full conditional distributions of all parameters. Except in a limited number of cases, full conditionals are difficult or impossible to derive, which results in a small number of prior distributions that are feasible to use with Gibbs samplers. In particular, conjugate priors are often used even when more believable priors are available. On the other hand, there is no advantage to conjugacy when using Stan. Users are free to specify priors that more accurately reflect their prior knowledge.

For a more thorough introduction to Stan see Stan Development Team (2015b) and Gelman

et al. (2013).

Data

In the `data` block of a Stan model we declare the data that will be passed to Stan, in this case from R. The `transformed data` block contains transformations of the variables declared in the `data` block. The declarations below are all straightforward, except for the matrix `P_inv`. This is the inverse of the difference of the degree and adjacency matrices, which is precomputed and passed to Stan as data. In `transformed data` we compute the Cholesky decomposition of `Pinverse`, which will be used for a more efficient implementation of the multivariate normal distributions required for the GMRF priors. Values for the location and scale parameters of the Cauchy distribution are also set in `transformed data`.

```
data {
  // dimensions
  int<lower=1>              N ; # number of observations
  int<lower=1>              C ; # number of congresses (time periods)

  // variables
  int<lower=1,upper=C>      cong[N] ;   # maps between observations & congresses
  int<lower=1,upper=56>     nVotes[N] ; # numerber of votes
  int<lower=0,upper=55>     nWins[N] ;  # number of majority party victories
  real                      vRatio[N] ; # vRatio

  // inverse of penalty matrix for GMRF prior
  matrix[C,C]               Pinverse ;
}
transformed data {
  real<lower=0> tau_scale ;   # scale for Cauchy priors on taus
  real<lower=0> tau_loc ;     # location for Cauchy priors on taus
  matrix[C,C] cholPinverse ;  # Cholesky decomposition of Pinverse

  tau_loc <- 0.0 ;
  tau_scale <- 2.5 ;
  cholPinverse <- cholesky_decompose(Pinverse) ;
}
```

Parameters

In the `parameters` and `transformed parameters` blocks we declare model parameters and deterministic transformations of the declared parameters.

```
parameters {
  real<lower=0>         phi ;
  vector[C]             bias_noise ;
  vector[C]             resp_noise ;
  real<lower=0>         tau_bias_noise ;
  real<lower=0>         tau_resp_noise ;
}
```

Parameter names with the suffix `_noise` denote variables to be given standard normal priors. Stan tends to work best if the target posterior distribution is marginally standard normal and uncorrelated and parameters on a similar scale. This means that it can help improve efficiency and convergence if variables defined in `parameters` are given standard normal priors when possible and then transformed in `transformed parameters` to have the desired distribution.[6]

In the `transformed parameters` block below both `tau_noise` parameters are transformed using the inverse-CDF method such that `tau_bias` and `tau_resp` have half-Cauchy distributions.[7] The transformations of `bias_noise` and `resp_noise` lead to the multivariate normal distributions required for the GMRF priors.

```
transformed parameters {
  vector[C]             b_bias ;
  vector[C]             b_resp ;
  real<lower=0>         tau_bias ;
  real<lower=0>         tau_resp ;

  tau_bias <- tau_loc + tau_scale * tan(pi() * (Phi_approx(tau_bias_noise) - 0.5)) ;
```

[6] This is an extremely oversimplified description and a thorough examination of this issue is beyond the scope of this thesis. See Betancourt and Girolami (2013).

[7] The inverse-CDF of the Cauchy distribution with location ℓ and scale s is $F^{-1}(p,\ell,s) = \ell + s\tan\{\pi(p-0.5)\}$. Thus if $z \sim \mathcal{N}(0,1)$ with CDF Φ then $\ell + s\tan\{\pi(\Phi(z)-0.5)\}$ is distributed Cauchy(ℓ,s). The transformations above therefore result in half-Cauchy(`tau_loc`, `tau_scale`) distributions due to the constraints that `tau_b_noise` and `tau_r_noise` be positive.

```
  tau_resp <- tau_loc + tau_scale * tan(pi() * (Phi_approx(tau_resp_noise) - 0.5)) ;

  b_bias  <- (tau_bias * cholPinverse) * bias_noise ;
  b_resp  <- (tau_resp * cholPinverse) * resp_noise ;
}
```

The lines

```
b_bias  <- (tau_bias * cholPinverse) * bias_noise ;
b_resp  <- (tau_resp * cholPinverse) * resp_noise ;
```

come from the fact that if z is a K-vector of iid standard normal random variables $z_k \sim \mathcal{N}(0,1)$ and $\theta = \mu + Lz$, where $LL' = \Sigma$, then $\theta \sim \mathcal{N}_K(\mu, \Sigma)$.[8,9] The advantage to this approach is twofold. For one, directly specifying a multivariate normal distribution for θ would require repeated computation of the inverse matrix Σ^{-1}. Additionally, unlike the elements of θ, the elements of z are independent which can lead to large gains in efficiency for MCMC algorithms in terms of effective sample size (Stan Development Team, 2015b).

Model

In the `model` block the likelihood and priors are specified. More specifically, Stan uses the log-likelihood and log-priors, the sum of which equals the log-posterior up to an additive constant.[10] The priors are the standard normals required for the transformations described above as well as a gamma distribution for `phi`.[11] Two vectors `alpha` and `beta` are also declared and will be filled in while looping over observations. This allows for the BetaBinomial likelihood to be vectorized.

[8] This is analogous to the univariate case where $z \sim \mathcal{N}(0,1)$ and $\theta = \mu + \sigma z$ together imply that $\theta \sim \mathcal{N}(\mu, \sigma^2)$.

[9] Here no mean vector μ is added since we are assuming means of zero for the multivariate normals.

[10] Stan works on the log scale for the same reason as most other statistical software. Logarithms help avoid the loss of the numerical precision in addition to greatly simplifying expressions for complicated probability distributions.

[11] Various priors for ϕ were tested. Differences in the resulting estimates were negligible.

```
model {
  // local variables
  real                logLik ;    # log likelihood
  real                logPrior ;  # log prior
  vector<lower=0>[N]  alphas ;
  vector<lower=0>[N]  betas ;

  logPrior <- (
    gamma_log(phi, 0.0001, 0.0001) +
    normal_log(tau_bias_noise, 0, 1) +
    normal_log(tau_resp_noise, 0, 1) +
    normal_log(bias_noise, 0, 1) +   // vectorized
    normal_log(resp_noise, 0, 1)     // vectorized
  ) ;

  // likelihood
  for (n in 1:N) {
    real theta_n ;
    theta_n   <- inv_logit(b_bias[cong[n]] + b_resp[cong[n]] * vRatio[n]) ;
    alphas[n] <- theta_n * phi ;
    betas[n]  <- (1 - theta_n) * phi ;
  }

  logLik <- beta_binomial_log(nWins, nVotes, alphas, betas) ; // vectorized

  increment_log_prob(logPrior + logLik) ; # increment log-posterior
}
```

Expressing the likelihood in this way is perhaps less intuitive than incrementing the likelihood within the loop

```
  for (n in 1:N) {
   ...
   logLik <- logLik + beta_binomial_log(nWins[n], nVotes[n], alpha[n], beta[n]) ;
  }
```

but constructing `alpha` and `beta` in the loop and then using the vectorized version of the BetaBinomial is much faster.

Generated quantities

The `generated quantities` block allows for the computation of distributions of quantities of interest without affecting the log posterior specified in the model block. In this case we compute bias towards the majority party in each congress, which can be calculated as $-1/2$ plus the inverse-logit of the `b_bias` parameter. For model checking we also simulate data from the posterior predictive distribution, which is discussed in further detail in the Results and Model Checking section.

```
generated quantities {
  real        bias[C] ; # bias toward majority
  int         nWins_rep[N] ; # posterior predictive simulations

  for (c in 1:C)
    bias[c] <- inv_logit(b_bias[c]) - 0.5 ;

  for (n in 1:N) {
    real theta_n ;
    real alpha_n ;
    real beta_n ;
    theta_n   <- inv_logit(b_bias[cong[n]] + b_resp[cong[n]] * vRatio[n]) ;
    alpha_n   <- phi * theta_n ;
    beta_n    <- phi * (1 - theta_n) ;
    nWins_rep[n]  <- beta_binomial_rng(nVotes[n], alpha_n, beta_n) ;
  }
}
```

3.4 Results and Model Checking

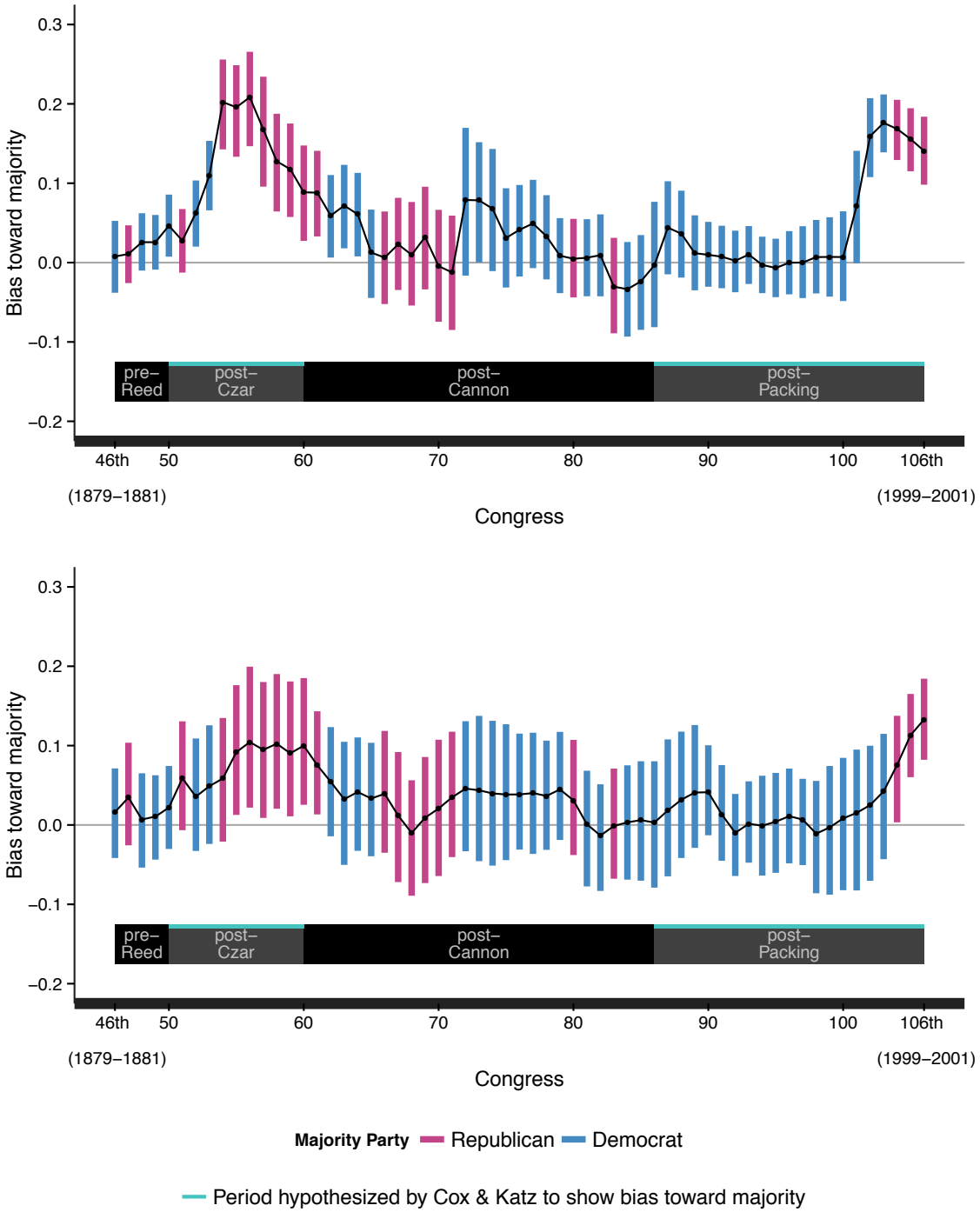

Figure 3.3: Estimated bias by Congress in original analysis (top) and reanalysis (bottom). Vertical bars represent 95% intervals, with the black line connecting medians.

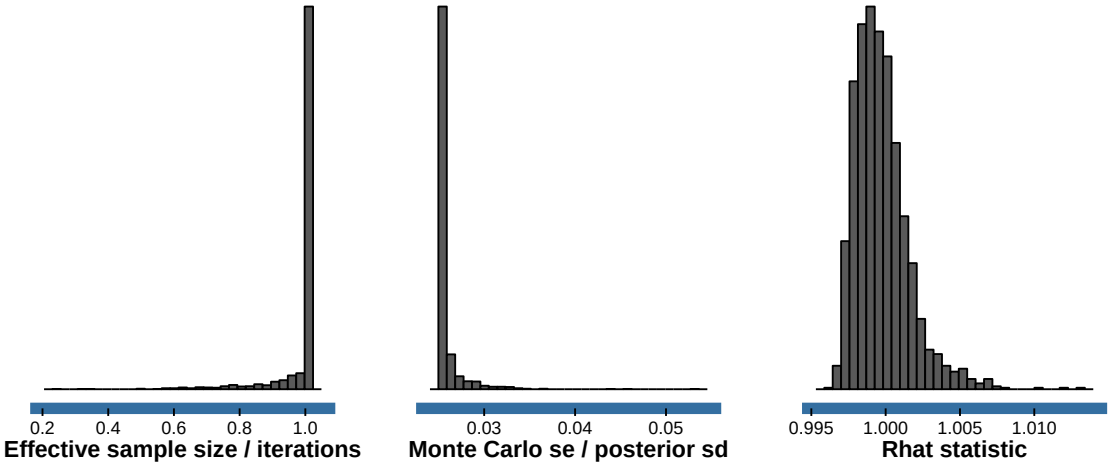

Figure 3.4: Diagnostics

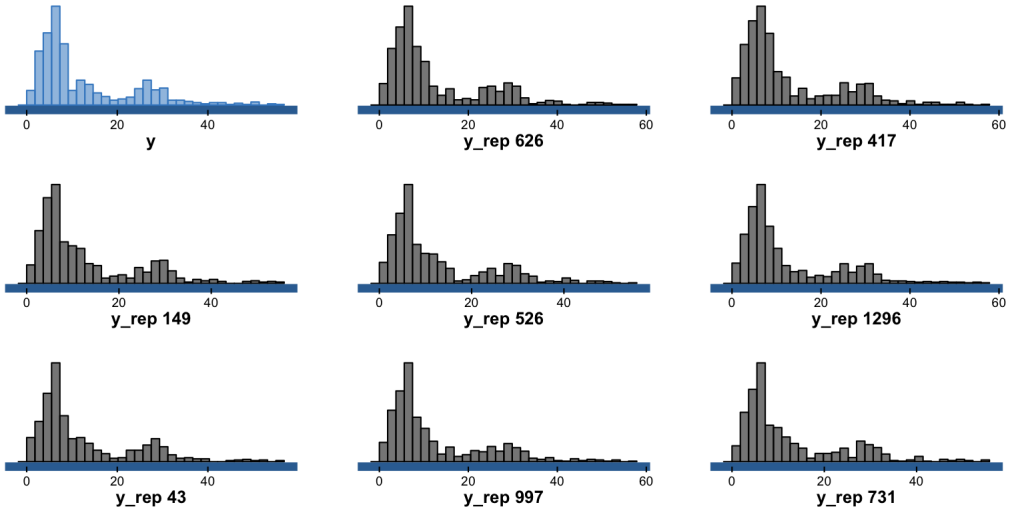

Figure 3.5: Replications from posterior predictive distribution vs. observed data

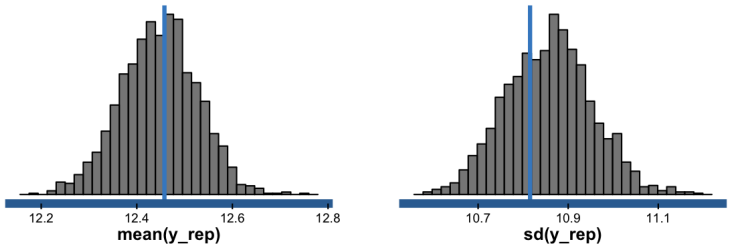

Figure 3.6: Distributions of test statistics $T(y_{rep})$ vs observed values $T(y)$

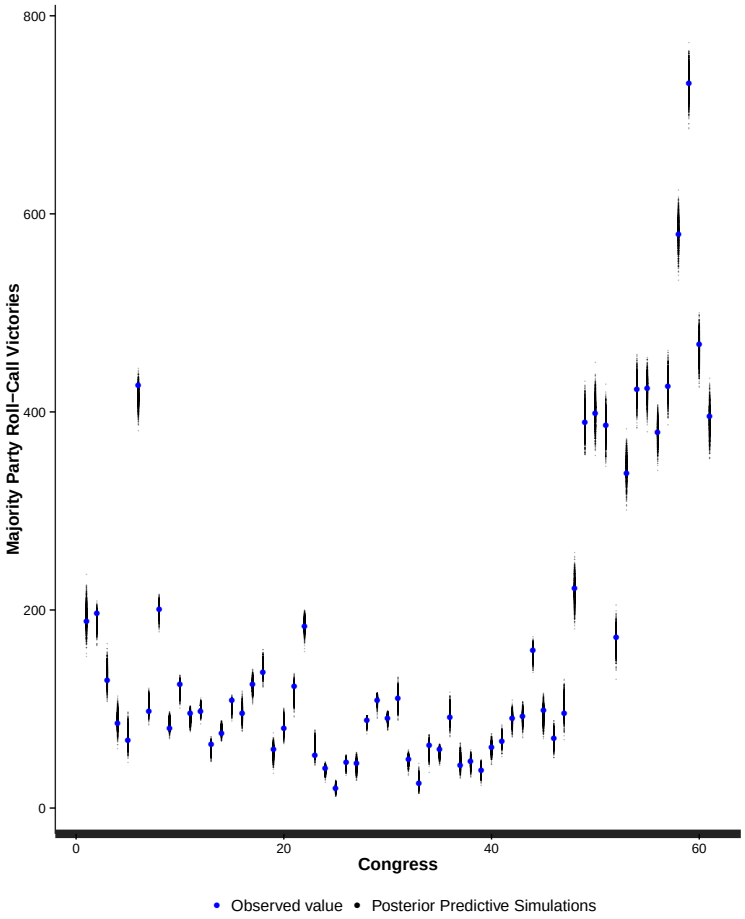

Figure 3.7: Replications from posterior predictive distribution vs. observed data

References

Betancourt, M. J., & Girolami, M. (2013). Hamiltonian Monte Carlo for Hierarchical Models. *arXiv preprint arXiv:1312.0906*.

Brezger, A., Kneib, T., & Lang, S. (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of statistical software*, *14*(11), 1–22.

Brezger, A., & Lang, S. (2006, February). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, *50*(4), 967–991.

Cox, G. W., & Katz, J. N. (2007, January). Gerrymandering Roll Calls in Congress, 1879–2000. *American Journal of Political Science*, *51*(1), 108–119.

Fahrmeir, L., & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *50*(2), 201–220.

Gelman, A. (2006, September). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., D., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third ed.). Chapman & Hall/CRC.

Goodrich, B., Katznelson, I., & Wawro, G. J. (2012). *Designing Quantitative Historical Social Inquiry: An Introduction to Stan.*

Hoffman, M. D., & Gelman, A. (2012). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org`

Rue, H., & Held, L. (2005). *Gaussian Markov random fields : theory and applications*. Boca Raton: CRC Press.

Stan Development Team. (2015a). *Rstan: the r interface to stan, version 2.6.0.* Retrieved from `http://mc-stan.org/rstan.html`

Stan Development Team. (2015b). Stan modeling language users guide and reference manual, version 2.6.0 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Wawro, G. J., & Katznelson, I. (2014). Designing Historical Social Scientific Inquiry: How Parameter Heterogeneity Can Bridge the Methodological Divide between Quantitative and Qualitative Approaches. *American Journal of Political Science*, *58*(2), 526–546.