An Introduction to Bayesian Inference Via Variational Approximations*

Justin Grimmer†

July 2, 2010

Abstract

Markov Chain Monte Carlo (MCMC) methods have facilitated an explosion of interest in Bayesian methods. MCMC is an incredibly useful and important tool, but can face difficulties when used to estimate complex posteriors or models applied to large data sets. In this paper I show how a recently developed tool in computer science for fitting Bayesian models, variational approximations, can be used to facilitate the application of Bayesian models to political science data. Variational approximations are often much faster than MCMC for fully Bayesian inference and in some instances facilitates the estimation of models that would be otherwise impossible to estimate. Variational approximations have guaranteed, fast, easily assessed convergence and provide accurate estimates of the expected value of the posterior (given sufficient sample size), but have some limitations. Therefore, variational approximations are best suited to problems when fully Bayesian inference would otherwise be impossible. Through a series of examples, I demonstrate how variational approximations are useful for a variety of political science research. This includes models to describe legislative voting blocs and statistical models for political texts. The code that implements the models in this paper is available in the supplementary material.

---

*I thank the Institute for Quantitative Social Science and the Center for American Political Studies for Financial support. For helpful comments I thank Andrew Coe, Adam Glynn, Gary King, Ben Lauderdale Clayton Nall, Kevin Quinn, and Adam Ramey.

†Assistant Professor, Department of Political Science, Stanford University; Encina Hall West 616 Serra St., Palo Alto, CA, 94305. http://justingrimmer.org, jgrimmer@stanford.edu, 617-710-6803.
1 Introduction

Bayesian models are an increasingly important tool for addressing long standing theoretical questions in political science, including: how nations interact (Hoff and Ward, 2004), the nature of democratic legitimacy (Trier and Jackman, 2008; Gill and Walker, 2005; Western and Jackman, 1994), and the structure and topics of conflict in American politics (Clinton, Jackman and Rivers., 2004; Quinn et al., 2010; Lax and Phillips, 2009). Markov Chain Monte Carlo (MCMC) and related sampling based approaches to Bayesian inference has facilitated the application of Bayesian models to political science data (Geman and Geman, 1984; Gelfand and Smith, 1990). MCMC allows scholars to quickly and accurately obtain estimates from statistical models, are easily programmed in standard software (or even available in prepackaged software, (Martin and Quinn, 2008)) and a large literature describes how to reliably use the sampling based approaches and diagnose problems in estimation (Gelman et al., 1995). Not surprisingly, sampling based approaches to Bayesian inference have become the standard (and often times only) way that political scientists attempt to fit Bayesian models.

MCMC is an extremely important tool for estimating statistical models and is likely to perform well on a wide range of problems. But for some models and data sets, MCMC has serious limitations, limitations that political scientists and methodologists often ignore (Gill, 2004). This is particularly true in the recent application of machine learning methods to political science problems, where complex models applied to large data sets expose the shortcomings of MCMC. When used on this class of problems, MCMC can require massive computing resources, converge too slowly to be useful, and worse yet, might approximate the entirely wrong posterior. In short, MCMC is likely to be useful in many instances, but there are many other instances where MCMC methods might fail to provide accurate posterior approximations in a reasonable amount of time.

With this potential limitation of MCMC in mind, this paper introduces to political scientists a different approach to Bayesian inference that is designed for the approximation of complex posteriors and the estimation models applied to large data sets: variational approximations (Jordan et al., 1999). A variational approximation is a deterministic method for estimating the full posterior distribution that has guaranteed convergence, which is easily assessed using a single scalar. The extremely general variational approximation introduced here is guaranteed to estimate the expected value of the posterior distributions correctly (for a large class of models and sufficient sample size) (Wang and Titterington, 2004), but will understate
the variability in the posterior distribution. This understated variability is a shortcoming of variational approximations, however, it is directly controllable and depends upon a transparent and easily modified set of assumptions.

Through a series of examples I demonstrate how variational approximations make feasible inferences and estimation of models that would be difficult or impossible to estimate using MCMC methods. This includes analysis of substantively interesting legislative behavior that would be difficult using standard sampling approaches and the fast estimation of extremely complicated models applied to large collections of political texts. But variational approximations have applications that stretch far beyond the applications in this paper: they are useful for any model where standard sampling based approaches to posterior approximation are infeasible or severely limiting. This includes many statistical models for political texts (for example, Quinn et al. 2010), the measurement of preferences in both political institutions and the public (for example, Clinton, Jackman and Rivers. 2004), and the estimation of complex models that vary over time and space (for example, Hoff and Ward 2004).

2 Limitations of Standard Approaches to Fitting Bayesian Models

The goal of Bayesian inference is to infer the posterior distribution of a set of parameters, and in many instances these posteriors are intractable: they cannot be used to directly calculate marginal distributions of parameters or other quantities of interest. Given the difficult in directly using the posterior distribution, political scientists have followed a large statistics literature and employed two methods for fitting Bayesian models: Markov Chain Monte Carlo (MCMC) (Geman and Geman, 1984; Gelfand and Smith, 1990) and Expectation-Maximization methods (Dempster, Laird and Rubin, 1977). MCMC and EM methods work well in many substantive problems, but can be difficult to apply or will perform poorly when applied to large data sets or complex models. In these instances, variational approximations will be most useful. In this section I describe a brief overview of MCMC and EM methods and describe instances where the methods may struggle.

The Gibbs sampler is the best known MCMC method for fitting a Bayesian model (Geman and Geman, 1984). To obtain an approximation of the posterior distribution, a Gibbs sampler proceeds in two broad steps. First, a Markov chain is defined with a steady state distribution equal to the posterior distribution (Gelfand and Smith, 1990). Once the Markov chain has likely reached its steady state distribution Monte Carlo is used to approximate the posterior (See Jackman (2000) for a political science introduction). Gibbs
samplers will be useful if the Markov chain converges to the true posterior and if a sufficient number of samples from the posterior have been obtained to accurately characterize the posterior. This motivates the use of burn-in iterations and the derivation of numerous convergence diagnostics, along with the careful analysis of trace-plots and other heuristics to assess whether the Markov chain is mixing (exploring) the posterior after convergence (Gelman and Rubin, 1992; Cowles and Carlin, 1996).

In many cases the careful use of convergence diagnostics and burn-in iterations will be sufficient to ensure that a Gibbs sampler is drawing from the correct distribution, but assessing convergence can be difficult in problems with many parameters. Gill (2004) demonstrates that convergence of a Gibbs sampler (or other MCMC methods) requires that all parameters have converged, not just the parameters of interest for a particular substantive question. The implications of this result are particularly disturbing when assessing the convergence of the Gibbs sampler applied to complex Bayesian models, because this requires checking that thousands of parameters have converged–including nuisance parameters that are often not stored during the sampling (Clinton, Jackman and Rivers., 2004). This problem is magnified as political scientists consider complex models applied to large data sets, particularly because Gibbs samplers may slowly explore some components of the high-dimensional parameter space. Further, convergence to the posterior distribution is insufficient for Gibbs samplers to provide an accurate approximation. The chain may slowly explore (mix) in the posterior distribution, resulting in a poor approximation of the true posterior.

An alternative approach to Bayesian inference is a two-step deterministic method for approximating a posterior. First, the mode of a posterior distribution or the maximum a posteriori (MAP) parameter estimates are obtained, usually using an EM algorithm (Dempster, Laird and Rubin, 1977). Then, a multivariate normal distribution is employed to approximate the posterior around its mode.

While the EM algorithm will prove useful in many instances, in small samples the EM algorithm may perform poorly. Posterior distributions only converge upon the multivariate normal distribution asymptotically and therefore when applied to small samples will provide a poor approximation to the posterior distribution (Gelman et al., 1995). This is a problem for many potential applications of the EM algorithm and normal approximation in political science. For example, posteriors for ideal point estimates based on roll call votes will converge upon the multivariate normal distribution at a slow rate, because of the incidental parameters produced with each new vote and because the number of legislators is fixed (Londregan, 2000). Likewise, the rate of convergence for mixture models, such as the model advanced in Quinn et al.
(2010), is known to be extremely slow, therefore requiring large data sets to justify the normal approxima-
tion (McLachlan and Peel, 2000).

While a posterior will be poorly approximated using a normal distribution in a small sample, actually
applying the normal approximation will be difficult for models with many parameters. Applying the normal
approximation requires the computation and inversion of an often large matrix (a Hessian evaluated at the
mode). For many realistic models, this can be a substantial computational obstacle. For example, Quinn
et al. (2010) introduce a model that would require inverting a $218,694 \times 218,694$ matrix. The substantial
challenges involved in estimating and inverting a matrix of this size often preclude the use of a normal based
approximation to the posterior and result in using only the posterior modes for inferences from a model.

3 Bayesian Inference via Deterministic Approximations: Variational Ap-
proximations

Variational approximations provide a different approach to the estimation of Bayesian models. Like the
EM algorithm, variational approximations are deterministic optimization algorithms that have guaranteed
convergence, easily assessed by examining the change in a scalar. Like MCMC algorithms, variational
approximations estimate the full posterior and do not require an additional step to perform inference. In
this section we describe the basics of the variational approximation. Derivation of the models and further
intuition are available in the supplemental materials.

3.1 The Tractability-Fit Tradeoff in Variational Approximations

The goal of a variational approximation is to approximate a posterior, $p(\beta|Y)$ by making an approximating
distribution, $q(\beta)$, as close as possible to the true posterior (Bishop, 2006). We search over the space of
approximating distributions in order to find the particular distribution with the minimum KL-divergence
with the actual posterior. Formally, we search over the set of approximating distributions $q(\beta)$ to minimize

$$
\text{KL}(q(\beta)||p(\beta|Y)) \equiv \text{KL}(q||p) = -\int q(\beta) \log \left\{ \frac{p(\beta|Y)}{q(\beta)} \right\} d\beta.
$$

If we make no assumptions about the factorized distribution, then Equation 3.1 is minimized when
$q(\beta) = p(\beta|Y)$ (because $\log 1 = 0$). Of course, this is not particularly helpful, because the posterior
is generally intractable. To make manipulation of the approximating distribution possible, we introduce
additional assumptions into the approximating distribution, with the hope that this will make inference
tractable, while also still providing a close approximation to the true posterior.
Following a large literature in computer science and machine learning, we use a very general form of approximating distributions, similar to that employed in Jordan et al. (1999) and Bishop (2006). We focus upon approximating distributions that contain additional independence than in the true posterior, but we make no other assumption about the particular parametric form of the approximating distribution. Rather, the distributional form for the approximating distribution will be estimated. We choose this approximating distribution also because it has been rigorously proven to perform well when applied to a large class of models. Wang and Titterington (2004) demonstrate that, given a sufficient number of observations from the data, this family of approximating distributions will correctly characterize the posterior mean, a guarantee not possible for sampling based approaches to inference.  

Following Bishop (2006), we call this a factorized approximation, because the independence assumption results in the approximating distribution being divided into a set of factors (or blocks of parameters). Similar to the Gibbs sampler, we first partition \( \beta \), into a set of \( K \) blocks, \( \beta = (\beta_1, \beta_2, \ldots, \beta_K) \). Then we will restrict attention to approximating distributions that have the form,  

\[
q(\beta) = \prod_{k=1}^{K} q(\beta_k).
\]

(3.2)

The variational algorithm will identify (rather than assume) the specific parametric families that constitute each component of the factorized distribution.

### 3.2 An Algorithm to Minimize the KL-Divergence

To minimize the KL-divergence we use an iterative algorithm that is analogous to the EM algorithm (Bishop, 2006). We derive the algorithm in the online appendix, but we state the algorithm here. Suppose that we have current estimates for all the factors of the approximating distribution \( q(\beta_1)^{\text{old}}, q(\beta_2)^{\text{old}}, \ldots, q(\beta_K)^{\text{old}} \) and we want to update the \( k^{th} \) factor. To do this, we will define

\[
E_{j \neq k}[\log p(\beta, Y)] = \int \prod_{j \neq k} \log p(\beta, Y) q(\beta_j)^{\text{old}} d\beta_j
\]

(3.3)

or the log posterior, averaged over our current estimates of the approximating distributions for all but the \( k^{th} \) component. Using this value, we then update \( q(\beta_k)^{\text{new}} \) by setting it to \( q(\beta_k)^{\text{new}} = \frac{\exp(E_{j \neq k}[\log p(\beta, Y)])}{\int \exp(E_{j \neq k}[\log p(\beta, Y)]) d\beta_k} \).

In each pass of the algorithm, we update all \( K \) of the factors using the same formula, using our current

---

1This guarantee is very general, but requires sufficient sample size and is proven for exponential family models or mixtures of exponential family models. Blei and Lafferty (2006) applies variational approximations to a non-exponential family model, noting that there are fewer guarantees, but the approximation appeared to perform well in their application.
estimates of the other factors. Therefore, each iteration of the variational approximation will sequentially update each of the factors,

\[
q(\beta_1)^{\text{new}} = \frac{\exp \left( E_{j \neq 1} \log p(\beta, Y) \right)}{\int \exp \left( E_{j \neq 1} \log p(\beta, Y) \right) d\beta_1}
\]

\[
q(\beta_2)^{\text{new}} = \frac{\exp \left( E_{j \neq 2} \log p(\beta, Y) \right)}{\int \exp \left( E_{j \neq 2} \log p(\beta, Y) \right) d\beta_2}
\]

\[\vdots\]

\[
q(\beta_K)^{\text{new}} = \frac{\exp \left( E_{j \neq K} \log p(\beta, Y) \right)}{\int \exp \left( E_{j \neq K} \log p(\beta, Y) \right) d\beta_K}
\]

Convergence of the algorithm is easily assessed using a single scalar, which is described in full in the online appendix.

4 Comparing Approximation Methods Using Bayesian Probit Regression

Variational approximations are best suited for models that are difficult or impossible to approximate using standard sampling based approaches to inference. But to compare variational approximations to Gibbs samplers and EM algorithms I apply a variational approximation to a standard Bayesian probit regression model (Jackman, 2000). This demonstrates the strengths and potential weaknesses of using a variational approximation for Bayesian inference. Sections 5 and 6 demonstrate how variational approximations make possible fully Bayesian inference for more difficult problems.

To introduce the model, suppose that for each observation \(i\), we observe a choice \(Y_i\), which can take on a value of 0 or 1, along with a vector of covariates, \(X_i\). Underlying the dichotomous response \(Y_i\), we suppose that there is a latent propensity for a positive response, \(Y_i^* \sim \text{Normal}(\mu_i, 1)\) with systematic component \(\mu_i = X_i' \beta\) (where \(\beta\) is a vector of coefficients). We suppose the standard observation mechanism for probit models,

\[
Y_i = \begin{cases} 
1 & \text{if } Y_i^* > 0 \\
0 & \text{if } Y_i^* < 0 \end{cases}
\]

This yields a straightforward likelihood that is standard in political science (King, 1998; Jackman, 2000). To complete the specification of the Bayesian model, we assume a set of vague priors on the regression coefficients, \(\beta \sim \text{Multivariate Normal}(0, \sigma^2 I)\), where \(\sigma^2\) is a large value and \(I\) is the appropriately sized identity matrix.

Note the analogy to Gibbs sampling. In Gibbs sampling, we condition on the other parameters and the functional to draw updated parameters. In variational approximations, we average over the other parameters using the approximating distribution.
We now show how to estimate the model using a variational approximation, which we will compare to the estimation from a Gibbs sampler and EM algorithm. (Because the Gibbs sampler and EM algorithm are now standard in political science research (see Jackman (2000)) I have included their derivation in the supplemental appendix.)

4.1 Variational Approximation

To apply the variational approximation, we divide the parameters into two blocks: the latent propensities $Y^*$ and the parameter vector $\beta$. Using these blocks, we will approximate the posterior with a distribution that assumes the latent propensities are independent of the parameter vector $\beta$, $q(Y^*, \beta) = q(Y^*)q(\beta)$. Due to the assumptions made in the model, we have an additional induced factorization $q(Y^*)q(\beta) = \prod_{i=1}^N q(Y^*_i)q(\beta)$.

To make this approximation as close as possible, we apply the iterative algorithm using two steps for each iteration. First, we describe the distributional form for each component of the approximating distribution $q(Y^*_i)$ and $q(\beta)$. Crucially, these functional forms are not assumed, rather are estimated as part of the approximation. Given the distributions for the factors, the variational approximation proceeds by iteratively updating the parameters of the distributions. Derivations of distributions and update steps are technical and are provided in the supplemental notes.

Straightforward derivations show that $q(Y^*_i)$ is a Truncated Normal Distribution, with

$$q(Y^*_i) = \begin{cases} \text{Normal}[0, \infty) (\mu_i, 1) & \text{if } Y_i = 1 \\ \text{Normal}(-\infty, 0) (\mu_i, 1) & \text{if } Y_i = 0, \end{cases}$$

and $q(\beta)$ is a Multivariate Normal distribution, with $q(\beta) = \text{Multivariate Normal}(\beta, \Sigma)$. Now, we iteratively update the parameters of the distributions $\mu_i, \beta, \Sigma$, which will be equivalent to making the approximating distribution as close as possible to the true posterior. Suppose the current value of the regression parameters is $\beta^{old}$. We will set $\mu_i^{new}$ to $\mu_i^{new} = X_i \beta^{old}$. $\Sigma$ is not updated during the algorithm and is given by, $\Sigma = (X'X + \frac{1}{\sigma^2}I)^{-1}$. Finally, we set $\beta^{new}$ to, $\beta^{new} = (X'X + \frac{1}{\sigma^2}I)^{-1}(X'\text{E}[Y^*])$, where $\text{E}[Y^*]$ is the expected value for each observation’s latent-propensity, with,

$$\text{E}[Y^*_i] = \begin{cases} \mu_i^{new} + \frac{\phi_i}{\sigma_i} & \text{if } Y_i = 1 \\ \mu_i^{new} - \frac{\phi_i}{\sigma_i} & \text{if } Y_i = 0, \end{cases},$$

where $\phi_i$ is equal to $\phi(-\mu_i)$ where $\phi$ is the normal density and $\Phi_i = \Phi(-\mu_i)$ where $\Phi$ is the cumulative normal density.
The terms $\mu_i$ and $\beta$, are sequentially updated until the algorithm converges—assessed using either a simple convergence statistic or changes in the parameter vectors (Bishop, 2006). We then use the closed form distributions $q(Y_i^*)$ and $q(\beta)$ to perform inference.

Code implementing variational Bayesian probit regression is available in the supplemental materials.

### 4.2 Comparing the EM, Gibbs, and Variational Approximation

I applied the Gibbs sampler, EM algorithm, and variational approximation to a simple simulated data set to compare the properties of the methods. Specifically, I generated 350 observations using a simple 6 parameter probit model. I then approximated the posterior for this model using the Gibbs sampler, the EM algorithm, and the variational approximation. Figure 1 provides a comparison of the posterior approximations across the three methods. The two left-hand plots compare the expected value of the regression coefficients using the EM (vertical-axis, left-hand plot) and the Gibbs sampler (plot second from left), to the expected value of the regression coefficients using the variational approximation (horizontal axis).

The variational approximation, the EM algorithm, and Gibbs sampler all agree on the same values: the expected values lie along the 45-degree line. This agreement across methods is the result of a theoretical guarantee of variational approximations: Wang and Titterington (2004) show that the factorized distribution used here will provide the correct expected values. It is important to note that this same guarantee cannot be made of Gibbs samplers in general, because it may fail to reach the posterior distribution in finite time. If the sampler is drawing from the wrong posterior, then the expected values of the parameter estimates are likely to be incorrect.

The next two plots presents the ratio of 95 percent credible intervals for the variational approximation to the 95 percent credible interval for the EM (plot second from right) and the Gibbs sampler (far right-hand plot) for the 6 coefficients included in the simulation. If the credible intervals were equal, the points would lie along the horizontal line. But the points are all below the horizontal line, and therefore the variational approximation understates the variability in the posterior, providing credible intervals that are only about 70% of their proper size.

This presents the fundamental shortcoming of the variational approximation: factorized approximations will always understate the variability in the posterior (MacKay, 2003). An active area of research seeks to improve the fit of variational approximations. Recent work has used the variational approximation as a first step and then added an importance sampling stage to provide a better approximation to the full posterior...
This figure compares the posterior estimates from the Gibbs sampler, the EM algorithm, and the variational approximation. The two left-hand plots demonstrate that all three methods agree on the expected value of the coefficients, a theoretical guarantee of variational approximations. But the two right-hand plots demonstrate that variational approximations will understate the variability in the posterior, which is demonstrated here by showing that the 95 percent credible intervals for the variational approximation are too small. Therefore, variational approximations are best suited for instances where other methods for posterior estimation are likely to be unreliable.

(Ghahramani and Beal, 2001). Other work has used “collapsed” variational approximations in order to provide a better estimate of the true posterior (Teh, Newman and Welling, 2007). But, without one of these additional modifications, variational approximations are best suited to problems where properties of the posterior make application of a Gibbs sampler difficult or models where many thousands of parameters and complicated sampling steps make convergence of an MCMC methods slow and difficult to assess.

I now demonstrate how variational approximations make estimation of extremely complex posteriors feasible and facilitate model selection in both parametric and non-parametric Bayesian models.

5 A Model of Legislative Voting Blocs

In this section, I use a variational approximation to estimate a modified version of the Quinn-Spirling voting bloc model to identify voting blocs in the Senate during the 110th Congress (Quinn and Spirling, 2010). Using the observed roll call matrix, the Quinn-Spirling voting bloc model groups legislators together by identifying groups of senators–blocs–who regularly vote together. The Quinn-Spirling voting bloc model is an important tool for describing how members of a legislature group together on amendments, particularly for legislatures where standard item–response theory methods for ideal point estimation are inapplicable (for example, the House of Commons in England) (Quinn and Spirling, 2010). As I will demonstrate, the model also allows identification of votes that distinguish voting blocs, allowing inferences about the issues on the
agenda that create cleavages between the blocs. Unfortunately, sampling based approaches to estimating the posterior from the Quinn-Spirling model severely limits the inferences we can make using the model and may prevent MCMC methods for estimating the correct posterior. A variational approximation avoids these problems, facilitating fully Bayesian inference about all parameters of the model.

Before describing the difficulties and limitations of estimating the model using MCMC, we introduce the model. Suppose that each senator $i$ ($i = 1, \ldots, N$) is a member of one-of-$K$ ($k = 1, \ldots, K$) voting blocs. Represent legislator $i$’s voting bloc with $\tau_i$, a $K \times 1$ indicator vector. Each legislator’s voting bloc is modeled as a draw from a multinomial distribution, $\tau_i \mid \pi \sim \text{Multinomial}(1, \pi)$, where $\pi$ is a vector that describes the prior probability of a senator belonging to each voting bloc.

We observe the legislator’s votes on a set of $J$ roll calls. We assume that each voting bloc is characterized by a $J \times 1$ vector, $\theta_k = (\theta_{k1}, \theta_{k2}, \ldots, \theta_{kJ})$, which describes a voting bloc’s propensity to support each particular proposal. Conditional on senator $i$’s voting bloc, we model a vote on the $j$th roll call, $V_{ij}$, as a draw from a Bernoulli distribution, $V_{ij} \mid \tau_{ik} = 1, \theta_k \sim \text{Bernoulli}(\theta_{kj})$. We assume that $\pi \sim \text{Dirichlet}(\alpha)$ and that, for all $k$ and $j$, $\theta_{kj} \sim \text{Beta}(\gamma_1, \gamma_2)$. The full posterior is presented in the online supplemental material.

An invariance in the posterior of the voting-bloc model complicates the application of Gibbs samplers. This invariance can cause Gibbs samplers to take draws from the wrong posterior distribution, resulting in incorrect inferences about the voting blocs in a legislature and their characteristics (McLachlan and Peel, 2000). The problem arises because a relabeling of the parameters provides the same posterior height, because information is only available about which observations are grouped together. But the posterior height does not change if the labels for each group of observations are permuted—in a two-component mixture model switching the cluster labels of the first and second component does not change the posterior height. More generally, if there are $K$ components in the mixture, then there are $K$ possible labels for the first component, $K - 1$ for the second, and so on. Therefore, mixture posteriors are characterized by $K!$ equivalent modes.

The invariance is problematic because the component labels are easily permuted during a run of a Gibbs sampler. Attempts to identify the components of the mixture through additional structure causes the sampler to draw from the wrong posterior and therefore is an unattractive option (McLachlan and Peel, 2000). Current recommendations are to run a chain without constraints and then post-process using a variety of methods (Jasra, Holmes and Stephens, 2005). This is often a useful solution, but some problems with sam-
pling can remain. First, the $K!$ modes provide a challenge to MCMC, which can sometimes be stuck in local modes, preventing the algorithm from exploring the entire posterior (Celeux, Hurn and Robert, 2000; Jasra, Holmes and Stephens, 2005). There are methods to ensure MCMC algorithms avoid local modes (Kirkpatrick, Gelatt and Vecchi, 1983; Gill and Casella, 2004), but these methods increase the computation time needed, particularly for the large mixture models used in political science applications (Quinn et al., 2010). Other scholars recommend using a “collapsed” Gibbs sampler, which integrates over some parameters, avoiding the invariance problem (indeed, this is essentially the sampler used in Quinn and Spirling (2010)). Collapsed samplers, however, only allow inferences about which pairs of senators belong to the same bloc and does not provide information about the votes that created cleavages among senators, limiting the usefulness of the model.

Given these problems, we use a variational approximation to estimate the Quinn-Spirling voting bloc model.

5.1 Variational Approximation

We divide the parameters into three blocks $\theta$, $\pi$, and $\tau$ and use the following approximating distribution

$$q(\theta, \pi, \tau) = q(\theta)q(\pi)q(\tau) = q(\pi) \prod_{k=1}^{K} q(\theta_{kj}) \prod_{i=1}^{N} q(\tau_{i})$$

(5.1)

where the second line follows from the assumptions of the model. To describe the iterative algorithm used to make this approximation as close as possible we first provide the distributional forms for, $q(\theta)$, $q(\tau_{i})$ and $q(\pi)$, and then describe the specific updates of the parameters of these distributions that constitute the update steps.

Calculating the distributional forms for each component of the approximating distributions (with details in the supplemental notes) we find that, $q(\tau_{i}) = \text{Multinomial}(r_{i})$, where $r_{i} = (r_{i1}, \ldots, r_{iK})$ represents the probability of legislator $i$ belonging to a given bloc. $q(\pi) = \text{Dirichlet}(\lambda)$, where $\lambda = (\lambda_{1}, \ldots, \lambda_{K})$ and that, $q(\theta_{kj}) = \text{Beta}(\eta_{kj1}, \eta_{kj2})$, where $\eta_{kj1}$, $\eta_{kj2}$ are the shape parameters for the Beta distribution.

Therefore, an iteration of the variational approximating algorithm will proceed by updating $r_{i}$, $\lambda_{k}$, and $\theta_{kj}$. Call $\lambda_{k}^{\text{old}}$, and $\theta_{kj}^{\text{old}}$ the values from the previous iteration. We then set $r_{ik}^{\text{new}}$ to

$$r_{ik}^{\text{new}} \propto \exp \left[ E[\log \pi_{k}] + \sum_{j=1}^{J} \{ V_{ij} [E[\log \theta_{kj1}]] + (1 - V_{ij}) [E[\log \theta_{kj1}]] \} \right].$$
where $E[\log \pi_k] = \Psi(\lambda_{old}^k) - \Psi(\sum_{z=1}^K \lambda_z^{old})$, $E[\log \theta_{kj1}] = \Psi(\eta_{kj1}^{old}) - \Psi(\eta_{kj1}^{old} + \eta_{kj2}^{old})$, and $E[\log \theta_{kj2}] = \Psi(\eta_{kj2}^{old}) - \Psi(\eta_{kj2}^{old})$. $\Psi(\cdot)$ is the Digamma function, the derivative of the Gamma function. (Intuition about $\Psi(\cdot)$ is not essential for the remainder of the argument). Next, we set $\lambda_{new}^k$ to $\lambda_{new}^k = \alpha_k + \sum_{i=1}^N r_{ik}^{new}$. And finally we set $\theta_{kj1}^{new}$ to, $\theta_{kj1}^{new} = \gamma_1 + \sum_{i=1}^N r_{ik}^{new} V_{ij}$ and $\theta_{kj2}^{new} = \gamma_2 + \sum_{i=1}^N r_{ik}^{new} (1 - V_{ij})$. Code to run the variational approximation for the legislative voting bloc model is available in the supplemental notes.

Figure 2: Convergence to Best Approximation Occurs Quickly and the Lower-Bound Allows for Model Selection

This figure demonstrates that the variational approximation to the Quinn-Spirling voting bloc model converges quickly, which is easily assessed using the lower-bound on the log-probability of the data. Further, the right-hand plot shows that this lower-bound facilitates Bayesian model selection. Given the roll call voting data and the modeling assumptions in the voting bloc model, four voting blocs are most probable in the US Senate.

5.2 Model Selection

A difficult problem in mixture models, like the Quinn-Spirling voting bloc model, is setting the number of voting blocs to include for the model. Fully Bayesian methods are useful for model selection, because they include an implicit penalization term for model complexity. This provides one data-driven method for ensuring that our model does not overfit the data (Kass and Raftery, 1995). To make explicit each model’s dependence on the assumed number of blocs, we will represent a $k$ component voting bloc model with $M_k$. Our goal when selecting the model is to use Bayes’ rule to determine the probability of each model, given the data (Bishop, 2006). Applying Bayes’ rule to make this intuition formal, $p(M_k|V) \propto p(M_k)p(V|M_k)$. Therefore to calculate the posterior for a particular model, we first need to know the prior probabilities for
each model $p(M_k)$, which we will assume is equal across the comparison models. The model selection will depend upon the evidence or the probability of the data, given a particular model’s assumption about the number of voting blocs, $p(V|M_k)$. Direct computation of the evidence is infeasible, because we are unable to manipulate the posterior distribution directly. But the variational approximation has optimized a lower-bound for the log-evidence, $\log p(V|M_k) \geq \mathcal{L}(q|\mathcal{M}_k)$. For technical reasons, we can not use this lower bound directly and need to add a correction term due to make the lower-bounds comparable (Bishop, 2006), so we will use $\tilde{\mathcal{L}}(q|\mathcal{M}_k) = \mathcal{L}(q|\mathcal{M}_k) + \log k!$.

The right-hand plot in Figure 2 carries out the comparison for the Quinn-Spirling voting bloc model, varying the number of blocs from 2 to 7. The maximum of the lower-bound occurs with four components. Because the bounds are on the log-scale, almost all the posterior mass would be located upon the four-component voting bloc model: given the modeling assumptions of the voting-bloc model and the roll-call voting data, the four-component voting model is the most likely model. So we analyze the voting bloc model using four blocs.

### 5.3 Extreme/ Moderate Cleavages and Divisive Votes

The four voting bloc model recovered two voting blocs within the Democratic and Republican parties: a moderate and an extreme voting bloc within each party, with no senators from different parties grouped into the same voting bloc. Table 1 presents representative members of each bloc (second column), a label to describe the bloc (left-hand column), votes that distinguish the voting blocs (third column), and the proportion of senators that fall within each voting bloc.

<table>
<thead>
<tr>
<th>Label</th>
<th>Examp. Senators</th>
<th>Distinctive Vote</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cons. Rep</td>
<td>Coburn, DeMint, Inhofe, Sessions</td>
<td>Amend. 521: Reduce Federal Debt</td>
<td>37.7%</td>
</tr>
<tr>
<td>Mod. Rep</td>
<td>Coleman, Hagel, Lugar, Murkowski</td>
<td>Amend. 2662: Prohibit Canyon Funds</td>
<td>12.2%</td>
</tr>
<tr>
<td>Mod. Dem</td>
<td>Bayh, McCaskill, Lieberman, Ben Nelson</td>
<td>Cloture on S. 2633: Iraq Redeployment</td>
<td>17.0 %</td>
</tr>
<tr>
<td>Lib. Dem</td>
<td>Clinton, Kennedy, Obama, Sanders</td>
<td>Table Amend. 4388: Mortgages</td>
<td>33.0 %</td>
</tr>
</tbody>
</table>

The variational approximation approximates the entire posterior, which facilitates the identification of votes that separated Republicans from Democrats and votes that created intra-party cleavages. We identify the votes that best distinguish each voting bloc in the third column in Table 1. To identify these votes, we first
obtained the expected probability of a given bloc $k$ voting in favor of a proposal $j$ $\Pr(V_{ij} = 1 | \tau_i = k) = \frac{\eta_{kj1}}{\eta_{kj1} + \eta_{kj2}}$. We then compared each voting bloc’s propensity to vote in favor of a proposal to the average propensity to support among the other blocs, $|\frac{\eta_{kj1}}{\eta_{kj1} + \eta_{kj2}} - \frac{1}{3} \sum_{m \neq k} \frac{\eta_{mj1}}{\eta_{mj1} + \eta_{mj2}}|$. The most divisive roll call vote is then placed in Column 3. A cursory glance at the votes demonstrates that Republican voting blocs were distinctive in their votes on fiscal issues, while the moderate Democrat voting blocs were distinguished by their opposition to Russ Feingold’s (D-WI) Iraq troop redeployment plan and liberal Democrats were separated by their support for an amendment to a mortgage bill offered by Dick Durbin (D-IL).

We can also use the voting bloc model to identify votes that created the intra-party cleavages. For each of the non-unanimous votes, we used the posterior approximation from the variational approximation to identify the votes that best distinguished the voting blocs within each party. This reveals that the major cleavages among Democrats were votes about national defense policy. The most divisive issues among Democrats were votes on the Iraq war, amendments related to immigration reform, and proposals regarding the Foreign Intelligence Surveillance Act (FISA). Divisions within the Republican party formed around votes about government spending and controlling the size of the federal bureaucracy. The most divisive votes among Republicans were votes about how members of Congress use the appropriations process to secure pork for their states and several votes related to the government provision of health care.\(^3\)

The identification of divisive votes between blocs exhibits the usefulness of variational approximations when applied to complex models. Standard sampling based approaches to inference would be difficult to apply to the Quinn-Spirling voting bloc model. And even if they are successful, Gibbs samplers are only able to characterize the pairs of senators who tend to vote together. In contrast, using a variational approximation allows fully Bayesian inference about all parameters, facilitating an important inference about the issues that divide groups in Congress.

### 6 Nonparametric Bayesian Methods and the Dirichlet Process Prior

Finite mixture models, like the voting bloc model, are useful tools for describing substantively interesting behavior across many different data sets. These models can be rendered more flexible (and often times more useful) through the application of nonparametric Bayesian priors to create infinite mixture models (Teh, 2010). We focus upon one particular nonparametric prior, the Dirichlet process prior (Ferguson, 1973; An-

---

\(^3\)This distinction is not just found among the twenty most divisive votes: the correlation between the Democrat and Republican divisiveness measure is -0.21, strong evidence that different votes were controversial for each party.
toniak, 1974; Blei and Jordan, 2006). Heuristically, Dirichlet process priors group together observations with similar characteristics into a countably infinite set of groups (or clusters). In any one sample, however, the prior uses both the observed data and modeling assumptions to select a finite number of clusters to include in the model. Therefore, the Dirichlet process prior provides one method for generating groups of observations from the data (or clusters).

Dirichlet process priors are well known in both the statistical and machine learning literature (Quinn and Spirling, 2010; Gill and Casella, 2009) and a wide-range of studies across many fields including– statistics (for example, Escobar and West 1995), computer science (for example, Teh et al. 2006), biology (for example Medvedovic and Sivaganesan 2002 and Kottas, Branco and Gelfand 2002), and even political science (for example, Quinn and Spirling 2010)–have shown that nonparametric Bayesian methods are useful for identifying groups of observations with similar characteristics. But their use has been limited because sampling based approaches to estimate Bayesian models can be extremely cumbersome. In this section we describe how a variational approximation developed in Blei and Lafferty (2006) can be used to facilitate the application of Dirichlet process priors to the statistical analysis of texts.

**Dirichlet Process Prior: No Free Lunch** While demonstrated to be useful for clustering across many applications, some care must be employed when using infinite mixture models for political science applications. Infinite mixture models based on the Dirichlet process prior are guaranteed to provide groupings of observations, but it is not guaranteed that these methods provide substantively interesting clusterings. Infinite mixture models (like other clustering algorithms) group observations together based upon each observation’s measured characteristics and assumptions built into the clustering procedure. As a result, the output of infinite mixture models may diverge from the theoretically driven clusters researchers would create when provided the same data set. For researchers to establish the theoretical utility of the clusterings from infinite mixture models, they must perform model checks that demonstrate the substantive utility of the clusterings. This is similar to the post-model checks required to establish the utility of finite mixture models (such as in Quinn et al. 2010 and Grimmer 2010) and factor analytic models (such as Clinton, Jackman and Rivers, 2004 and Ansolabehere, Rodden and Snyder 2008).

With this caveat in mind, we apply the Dirichlet process prior to identify clusters of press releases discussing the same issue, or topics, in a collection of over 64,000 press releases: every press release from each Senate office, from 2005-2007 (Grimmer, 2010). The Dirichlet process, like other priors, is over-
whelmed by the data in large samples. However, applying the model to this large collection of press releases demonstrates how variational approximations can substantially reduce the computation time necessary for complicated models applied to very large data sets. When mixture models are used to identify groups of documents that discuss the same basic issue (or topic) it is commonly called topic-modeling (Quinn et al., 2010; Blei and Lafferty, 2009; Grimmer, 2010). The data in topic models are a vector of word counts, describing the relative rate words (or stems) are used in the collection of documents. In this application, we will use a nonparametric topic model to identify groups of press releases that discuss similar topics and we will demonstrate that these groupings of documents are substantively interesting for scholars of Congressional home style. Identifying the topics of press releases provides information about how members of Congress present their work in Washington to constituents (Fenno, 1978).

To apply a statistical model to the press releases, I apply a set of well established procedures to translate the press releases into count vectors (Manning et al., 2008), described in the online supplemental material. The result of the steps is that each document is represented as a count-vector. For each document $i$, we observe the number of times word $j$ occurs, $y_{i,j}$. We then collect this into the $w \times 1$ count vector, $y_i$, where $w = 2,796$ for this example, or the number of unique words included in the corpora.

6.1 Nonparametric Topic Model for Senate Press Releases

The Dirichlet process is a distribution over distributions rather than parameters. Therefore, a draw from a Dirichlet process is a distribution rather than a parameter vector. Dirichlet processes are parameterized with a concentration parameter $\alpha$ and a base distribution $G_0$. We will write the Dirichlet process distribution as $\text{DP}(\alpha, G_0)$. $G_0$ is the expected distribution from the Dirichlet process, analogous to the average or first moment of a distribution over parameters (Teh, 2010). The concentration parameter $\alpha$ determines how close the draws from the distribution are to the base measure: the larger value of $\alpha$, the closer the draws will be to the base measure. The number of components in the mixture will depend strongly on our selection of $\alpha$.

To define an infinite mixture model with a Dirichlet process prior, we suppose that a measure $G$ is drawn from the Dirichlet process, $G|\alpha, G_0 \sim \text{DP}(\alpha, G_0)$. Then, conditional on this distribution, each observation has a parameter vector drawn: $\theta_i|G \sim G$. Finally, we draw the observed data from an appropriate distribution $y_i|\theta_i \sim F(\theta_i)$. The draws from the Dirichlet process prior are discrete with probability one (Teh, 2010). Therefore, we can partition observations according to the value of the parameter vector that is drawn (Teh, 2010), providing the groups of press releases based on their topics.
It will be useful to employ a second representation of the Dirichlet process to derive the variational approximation: the stick-breaking representation, which also clearly shows that the distribution drawn from the Dirichlet Process prior, \( G \), is discrete (Sethuraman, 1994; Blei and Jordan, 2006). To define this representation, suppose that an infinite number of draws are taken from a Beta distribution, \( v_k \sim \text{Beta}(1, \alpha) \) for \( k = 1, \ldots, \infty \) (where we have intentionally reused \( \alpha \)). Collect these draws into the infinite length vector \( \mathbf{v} = (v_1, v_2, \ldots) \). Next, an infinite number of parameters are drawn from a base distribution \( \theta_k \sim G_0 \), for \( k = 1, \ldots, \infty \). To model the press releases, we will suppose that \( G_0 \) is a Dirichlet distribution, with vector of shape parameters given by \( \lambda \). Conditional on \( \mathbf{v} \) define \( \pi(\mathbf{v})_k = v_k \prod_{j=1}^{k-1} (1 - v_j) \) and call \( \pi(\mathbf{v}) = (\pi(\mathbf{v})_1, \ldots) \). Define \( \delta(\cdot) \) as the Dirac delta function, which is a distribution that places all its mass on its argument. We can define \( G \), a draw from a Dirichlet process as (Blei and Jordan, 2006)

\[
G = \sum_{k=1}^{\infty} \pi(\mathbf{v})_k \delta(\theta_k). \tag{6.1}
\]

Equation 6.1 simultaneously shows that \( G \) is discrete and an infinite mixture over parameters. We construct an arbitrary distribution (measure) \( G \) by mixing together a set of discrete points (indicated by using the dirac delta function). The dirac delta function shows that \( G \) places all of its mass on a countably infinite set of points and therefore we can “cluster” the observations by observing the parameters it is assigned. The probability of each value of \( \theta_k \) occurring is governed by the value of \( \pi(\mathbf{v})_k \), which is the stick breaking portion of the model. \( \pi(\mathbf{v})_k \) breaks off a portion of the “probability-stick” for the \( k^{th} \) component.

Because the number of components used in anyone application of the Dirichlet process prior depends strongly on \( \alpha \), we place a prior on \( \alpha \) and obtain a posterior estimate of the concentration parameter. Blei and Jordan (2006) suggest placing a Gamma distribution as the prior on \( \alpha \). Define the sampling distribution for the Gamma\((s_1, s_2)\) distribution as, \( p(\alpha|s_1, s_2) = \alpha^{s_1-1} \exp(-s_2\alpha) \frac{s_2^{s_1}}{\Gamma(s_1)} \). This distribution is conjugate to a Beta\((1, \alpha)\) distribution, simplifying the update steps.

We suppose that the topic of each press release, \( \tau_i \), is a draw from a Multinomial distribution, \( \tau_i | \pi(\mathbf{v}) \sim \text{Multinomial}(1, \pi(\mathbf{v})) \). Conditional on press release’s topic, we suppose that \( y_i | \tau_{i,k} = 1, \theta \sim \text{Multinomial}(n_i, \theta_k) \), where \( n_i \) are the total number of words used in the \( i^{th} \) press release.

Therefore, we will model the press releases using the following posterior,
\[\alpha \mid s_1, s_2 \sim \text{Gamma}(s_1, s_2)\]
\[v_k \mid \alpha \sim \text{Beta}(1, \alpha) \text{ for } k = 1, \ldots, \infty\]
\[\theta_k \mid G_0, \lambda \sim \text{Dirichlet}(\lambda) \text{ for } k = 1, \ldots, \infty\]
\[\tau_i \mid \pi(v) \sim \text{Multinomial}(1, \pi(v)) \text{ for } i = 1, \ldots, N\]
\[y_i \mid \tau_{i,k} = 1, \theta_k \sim \text{Multinomial}(n_i, \theta_k) \text{ for } i = 1, \ldots, N\]

### 6.2 Variational Approximation for Dirichlet Process Prior

Inference for Dirichlet process priors (and other nonparametric Bayesian methods) is complicated. EM algorithms cannot be used and sampling based approaches often only provide information about a subset of parameters (in this case, the topic of press releases) (Neal, 2000). Further, sampling based approaches are often difficult to apply because they explore the posterior slowly and often require several restarts before capturing the posterior. These problems are magnified when applied to very large data sets, like the collection of press releases here.

A variational approximation for the Dirichlet process prior, developed in Blei and Jordan (2006), avoids these problems. We approximate the infinite mixture model with a truncated approximating distribution with a finite number of components in the model. Critically, this does not limit the number of components that will be used in the posterior estimate, which we ensure by setting the truncation to be much higher than the likely number of components used in the posterior.\(^4\)

After assuming this truncation, we divide the approximating distribution into four blocks, \(v, \tau, \theta\) and \(\alpha\), assuming that the approximating distribution has the form, \(q(v, \tau, \theta, \alpha) = q(v)q(\tau)q(\theta)q(\alpha)\). The modeling assumptions imply that the approximating distribution can be written as

\[
q(v, \tau, \theta, \alpha) = \prod_{k=1}^{K-1} q(v_k) \prod_{i=1}^{N} q(\tau_i) \prod_{k=1}^{K} q(\theta_k)q(\alpha)
\]

Again, this adds no additional assumptions and is a direct consequence of the assumptions in the model.

To define the algorithm used to make this approximation as close as possible to the true posterior, we first provide the functional forms for each factor of the approximating distribution and then we will describe the specific update steps on the parameters of the distributions. The distributional forms for the approximating distribution are given by, \(q(\tau_i) = \text{Multinomial}(1, \pi_i), q(v_k) = \text{Beta}(\gamma_{k,1}, \gamma_{k,2}), q(\theta_k) =\)

---

\(^4\)Blei and Jordan (2006) show that the approximation improves very quickly as the number of components included in the approximating distribution increase.
Dirichlet($\eta_k$), and $q(\alpha) = \text{Gamma}(w_1, w_2)$. The variational approximation will proceed by sequentially updating $r_i$, $\gamma_{k,1}$, $\gamma_{k,2}$, $\eta_k$, $w_1$, $w_2$. Suppose that the current values of the parameters are given by $\gamma_{k,1}^{\text{old}}$, $\gamma_{k,2}^{\text{old}}$, $\eta_k^{\text{old}}$, $w_1^{\text{old}}$, $w_2^{\text{old}}$. We then set $r_{i,k}^{\text{new}} \propto \exp\{E[\log v_k] + E[\log(1 - v_k)] + y_{i,k}E[\log \theta_k]\}$, where $E[\log v_k] = \Psi(\gamma_{1,k}^{\text{old}}) - \Psi(\gamma_{1,k}^{\text{old}} + \gamma_{2,k}^{\text{old}})$, $E[\log(1 - v_k)] = \Psi(\gamma_{2,k}^{\text{old}}) - \Psi(\gamma_{1,k}^{\text{old}} + \gamma_{2,k}^{\text{old}})$ and $E[\log \theta_k] = \sum_{j=1}^w \Psi(\eta_{k,j}) - \Psi(\sum_{m=1}^w \eta_{k,m})$. Next, we set $\eta_{k}^{\text{new}}$ to, $\eta_{k}^{\text{new}} = \lambda + \sum_{i=1}^N r_{i,k}^{\text{new}} y_i$. We set $\gamma_{k,1}^{\text{new}}$ and $\gamma_{k,2}^{\text{new}}$ to, $\gamma_{k,1}^{\text{new}} = 1 + \sum_{i=1}^N r_{i,k}^{\text{new}}$ and $\gamma_{k,2}^{\text{new}} = w_2^{\text{old}} + \sum_{i=1}^N \sum_{j=k+1}^N r_{i,j}^{\text{new}}$. Finally, we set $w_1^{\text{new}} = s + K - 1$ and $w_2^{\text{new}} = s_2 - \sum_{k=1}^{K-1} \left[ \Psi(\gamma_{k,2}^{\text{new}}) - \Psi(\gamma_{k,1}^{\text{new}} + \gamma_{k,1}^{\text{new}}) \right]$. 

### 6.3 Political Attention in Senate Press Releases

I applied the algorithm to the full collection of 64,033 press releases. The variational approximation took approximately 45 minutes to converge, implemented in R and run on a standard desktop computer. In infinite mixture models the number of components used by the model is inferred from the model using a combination of data and modeling assumptions (note that the number of components employed by the model need not correspond to the “true” number of clusters in the population (Petrone and Raftery, 1997)). The left-hand plot in Figure 3 shows the approximated posterior distribution on the number of topics. To obtain this posterior distribution, I used the variational approximation to generate a posterior distribution on the stick-breaking proportions, $\pi(\nu)$ and then drew topic labels, conditional $\pi(\nu)$. Using this simulation approach, we find that the 95 percent credible interval on the number of topic stretches from 66 topics to 79.

While the model groups together the press releases into 72 topics (or clusters), many of those topics contain only a few press releases, which is shown in the right-hand plot of Figure 3. Here, we present the expected number of documents for each topic $k$ or $\sum_{i=1}^N r_{i,k}$. Only 30 topics have an expected number of documents greater than 100 and 45 topics are expected to have more than 10 documents. This an interesting property of nonparametric topic models: a large number of topics will receive only a few documents and a few topics will have a large number of documents assigned (Teh, 2010). Therefore, we will focus initially on the largest components that the model identified and note that a substantively interesting problem is combining the components with fewer documents with the components with many more documents (Gill and Casella, 2009).

The most important quantity of interest from the infinite mixture model are the identified *topics* in the press releases and the proportion of press releases allocated to each of those topics. This provides one measure of how senators divide their attention when communicating with constituents (Grimmer, 2010).
Figure 3: Number of Topics Employed and the Distribution of Documents Over Topics

This figure presents the distribution on the number of topics and documents per topic, as estimated by the Dirichlet process prior. The left-hand plot shows that the model identifies about 78 topics in the collection of press releases, but the right-hand plot shows that only about 45 have more than just a few press releases per topic. This reflects the assumption that the distribution of topics is assumed to grow according to a power law distribution (Teh, 2010) and demonstrates how the number of components obtained using a Dirichlet process depend on the modeling characteristics.

Table 2 presents the ten largest topics (as measured by the expected number of documents per topic). The left-hand column contains an identifying label for each topic (generated by hand after reading a sample of 15 press releases assigned to the topic), the center column contains ten stems that accurately label the press releases in a topic (using a method developed in Grimmer (2010)), and the right-hand column provides the percentage of press releases in each topic. The three largest topics demonstrate how senators balance between credit-claiming for particularistic goods (the Appropriations/Grants topic); symbolic activities, such as honoring constituents and memorializing major national holiday (the Honorary topic); and the discussion of major substantive issues (the Iraq war topic). The presence of all three demonstrates the need to have coding schemes that go beyond the standard focus on policy-oriented speech to understand how legislators express their priorities, while also showing that the model is able to identify substantively interesting topics of press releases.

This section demonstrated how variational approximations make possible the application of nonparametric Bayesian methods to large data sets. This is difficult using MCMC because algorithms applied to topic models applied to large data sets tend to converge slowly (Blei and Lafferty, 2006). But variational approximations provide a fast approximation to the true posterior, providing useful insights into what members
Table 2: Ten Most Discussed Topics

<table>
<thead>
<tr>
<th>Label</th>
<th>Identifying Stems</th>
<th>% Press Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriations/Grants</td>
<td>fund, project, 000, million, water, transport, develop, improv, airport, citi</td>
<td>8.6</td>
</tr>
<tr>
<td>Honorary</td>
<td>honor, servic, school, serv, american, veteran, academ, famili, student, world</td>
<td>8.2</td>
</tr>
<tr>
<td>Iraq War</td>
<td>iraq, troop, war, iraq, american, militari, polit, secur, support, countri</td>
<td>6.6</td>
</tr>
<tr>
<td>Health Grants</td>
<td>health, program, educ, children, school, fund, student, care, servic, 000</td>
<td>6.3</td>
</tr>
<tr>
<td>Homeland Security</td>
<td>secur, homeland, port, border, depart, fund, guard, air, servic, transport</td>
<td>5.3</td>
</tr>
<tr>
<td>Judicial Nominations</td>
<td>court, vote, justic, american, judg, case, hous, congress, constitut, protect</td>
<td>4.8</td>
</tr>
<tr>
<td>Hurricanes/Disasters</td>
<td>disast, assist, hurrican, fema, flood, damag, fund, katrina, storm, declar</td>
<td>4.5</td>
</tr>
<tr>
<td>Taxes</td>
<td>tax, american, budget, social, secur, wage, famili, worker, increas, benefit</td>
<td>4.4</td>
</tr>
<tr>
<td>Defense Projects</td>
<td>million, defens, fund, air, militari, base, facil, guard, armi, project</td>
<td>4.2</td>
</tr>
<tr>
<td>Health Policy</td>
<td>health, care, drug, medicar, senior, prescript, plan, medic, program, cost</td>
<td>3.8</td>
</tr>
</tbody>
</table>

of Congress communicate with their constituents.

7 Conclusion

In this paper I have demonstrated how variational approximations allow political scientists to estimate complex Bayesian models applied to large data sets, even when standard approaches to Bayesian inference fail. The result is that variational approximations facilitate inferences that are otherwise impossible to make. Variational approximations made possible the fast estimation of a voting bloc model that revealed intra-party cleavages in the US Senate and a nonparametric topic-model that provides a flexible method for describing the content of senators’ press releases, an important quantity of interest for studying home style.

This paper presents only an introduction to variational approximations, leaving undiscussed details from a large literature in machine learning and computer science (Jordan et al., 1999; Bishop, 2006). The extensions of variational approximations makes them useful for a potentially large number of social scientific problems. Political scientists are now attempting to fit increasingly complex models to describe the contents of very large data sets. From models of social networks to dynamic models of public opinion, variational approximations can make feasible inferences that were previously impossible. And as a result, political scientists using variational approximations will be to make inference about politics that would otherwise be impossible.

References


22


