Advance Access publication November 11, 2010

Political Analysis (2011) 19:32–47
doi:10.1093/pan/mpq027

An Introduction to Bayesian Inference via Variational Approximations

Justin Grimmer
Department of Political Science, Stanford University, 616 Serra St., Encina Hall West, Room 100, Stanford, CA 94305
e-mail: jgrimmer@stanford.edu

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, we 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 facilitate the estimation of models that would be otherwise impossible to estimate. As a deterministic posterior approximation method, variational approximations are guaranteed to converge and convergence is easily assessed. But variational approximations do have some limitations, which we detail below. Therefore, variational approximations are best suited to problems when fully Bayesian inference would otherwise be impossible. Through a series of examples, we 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.

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 (Western and Jackman 1994; Gill and Walker 2005; Trier and Jackman 2008), and the structure and topics of conflict in American politics (Clinton, Jackman, and Rivers 2004; Lax and Phillips 2009; Quinn et al. 2010). 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, is easily programmed in standard software (or even available in prepackaged software, Martin, Quinn, and Park, Forthcoming), 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 other instances where MCMC methods might fail to provide accurate posterior approximations in a reasonable amount of time.

Author's note: For helpful comments I thank Andrew Coe, Adam Glynn, Gary King, Ben Lauderdale Clayton Nall, Kevin Quinn, Adam Ramey, the helpful comments of the editors, and the anonymous reviewers. Supplementary materials for this article are available on the *Political Analysis* Web site.

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, we 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 (e.g., Quinn et al. 2010), the measurement of preferences in both the political institutions and the public (e.g., Clinton, Jackman, and Rivers 2004), and the estimation of complex models that vary over time and space (e.g., 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 given observed data. 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 difficulty in directly using the posterior distribution, political scientists have followed a large statistics literature and employed two methods for fitting Bayesian models: MCMC (Geman and Geman 1984; Gelfand and Smith 1990) and expectation-maximization (EM) methods (Dempster, Laird, and Rubin 1977). MCMC and EM methods work well in many substantive problems but can perform poorly when applied to large data sets or complex models. In these instances, variational approximations will be most useful. In this section, we describe MCMC and EM methods and discuss 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 reached its steady-state distribution, Monte Carlo is used to approximate the posterior. 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. Furthermore, 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 estimating a posterior. First, the mode of a posterior distribution or the *maximum a posteriori* 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.

Although the EM algorithm and multivariate normal approximation will prove useful in many instances, in small samples, the multivariate normal distribution will provide a poor substitute for the true posterior. Posterior distributions only converge upon the multivariate normal distribution asymptotically, so the application of the normal approximation is not justified in small data sets (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 approximation (McLachlan and Peel 2000).

Although 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 Approximations

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.

3.1 *The Tractability-Fit Tradeoff in Variational Approximations*

The goal of a variational approximation is to approximate a posterior, $p(\boldsymbol{\beta}|\boldsymbol{Y})$ with a second distribution, called the *approximating* distribution, $q(\boldsymbol{\beta})$ (Bishop 2006). To make this approximation as close as possible, we search over the space of approximating distributions to find the particular distribution with the minimum Kullback–Leibler (KL) divergence with the actual posterior. Formally, we search over the set of approximating distributions $q(\boldsymbol{\beta})$ to minimize

$$\mathrm{KL}(q(\boldsymbol{\beta})||p(\boldsymbol{\beta}|\boldsymbol{Y})) \equiv \mathrm{KL}(q||p) = -\int q(\boldsymbol{\beta}) \log\left\{\frac{p(\boldsymbol{\beta}|\boldsymbol{Y})}{q(\boldsymbol{\beta})}\right\}\mathrm{d}\boldsymbol{\beta}. \quad (1)$$

If we make no assumptions about the factorized distribution, then equation (1) is minimized when $q(\boldsymbol{\beta}) = p(\boldsymbol{\beta}|\boldsymbol{Y})$ (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. The goal of the additional assumptions is to make inference tractable while also 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 assume independences between parameters that may not be present 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 proven to perform well

when applied to a large class of models. Wang and Titterington (2004) demonstrate that, given a sufficient number of observations, this family of approximating distributions will correctly characterize the posterior mean, a guarantee not possible for sampling-based approaches to inference.[1]

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 $\boldsymbol{\beta}$, into a set of K blocks, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$. Then, we restrict attention to approximating distributions that have the form,

$$q(\boldsymbol{\beta}) = \prod_{k=1}^{K} q(\boldsymbol{\beta}_k). \tag{2}$$

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). Suppose that we have current estimates for all the factors of the approximating distribution

$$q(\boldsymbol{\beta}_1)^{\text{old}}, q(\boldsymbol{\beta}_2)^{\text{old}}, \ldots, q(\boldsymbol{\beta}_K)^{\text{old}} \tag{3}$$

and we want to update the kth factor. To do this, we define

$$\mathrm{E}_{j\neq k}[\log p(\boldsymbol{\beta}, \boldsymbol{Y})] = \int \prod_{j\neq k} \log p(\boldsymbol{\beta}, \boldsymbol{Y}) q(\boldsymbol{\beta}_j)^{\text{old}} \mathrm{d}\boldsymbol{\beta}_j \tag{4}$$

or the log posterior, averaged over our current estimates of the approximating distributions for all but the kth component. Using this value, we then update $q(\boldsymbol{\beta}_k)^{\text{new}}$ by setting it to $q(\boldsymbol{\beta}_k)^{\text{new}} = \frac{\exp\left(\mathrm{E}_{j\neq k}\left[\log p(\boldsymbol{\beta},\boldsymbol{Y})\right]\right)}{\int \exp\left(\mathrm{E}_{j\neq k}\left[\log p(\boldsymbol{\beta},\boldsymbol{Y})\right]\right)\mathrm{d}\boldsymbol{\beta}_k}$. In each pass of the algorithm, we update all K of the factors using the same formula, using our current estimates of the other factors.[2] Therefore, each iteration of the variational approximation will sequentially update each of the factors,

$$\begin{aligned}
q(\boldsymbol{\beta}_1)^{\text{new}} &= \frac{\exp\left(\mathrm{E}_{j\neq 1}\left[\log p(\boldsymbol{\beta}, \boldsymbol{Y})\right]\right)}{\int \exp\left(\mathrm{E}_{j\neq 1}\left[\log p(\boldsymbol{\beta}, \boldsymbol{Y})\right]\right)\mathrm{d}\boldsymbol{\beta}_1} \\
q(\boldsymbol{\beta}_2)^{\text{new}} &= \frac{\exp\left(\mathrm{E}_{j\neq 2}\left[\log p(\boldsymbol{\beta}, \boldsymbol{Y})\right]\right)}{\int \exp\left(\mathrm{E}_{j\neq 2}\left[\log p(\boldsymbol{\beta}, \boldsymbol{Y})\right]\right)\mathrm{d}\boldsymbol{\beta}_2} \\
&\vdots \quad \vdots \quad \vdots \\
q(\boldsymbol{\beta}_K)^{\text{new}} &= \frac{\exp\left(\mathrm{E}_{j\neq K}\left[\log p(\boldsymbol{\beta}, \boldsymbol{Y})\right]\right)}{\int \exp\left(\mathrm{E}_{j\neq K}\left[\log p(\boldsymbol{\beta}, \boldsymbol{Y})\right]\right)\mathrm{d}\boldsymbol{\beta}_K}.
\end{aligned} \tag{5}$$

Convergence of the algorithm is easily assessed using a single scalar.

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, we apply a variational approximation to a standard Bayesian probit

[1] This 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 nonexponential family model, noting that there are fewer guarantees, but the approximation appeared to perform well in their application.

[2] 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.

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, $\boldsymbol{X}_i$. 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 $\mu_i = X_i^{'}\boldsymbol{\beta}$ (where $\boldsymbol{\beta}$ is a vector of coefficients). We suppose the standard observation mechanism for probit models,

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0, \\ 0 & \text{if } Y_i^* < 0. \end{cases} \tag{6}$$

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, $\boldsymbol{\beta} \sim$ Multivariate Normal $(\mathbf{0}, \sigma^2\boldsymbol{I})$, where σ^2 is a large value and $\boldsymbol{I}$ is the appropriately sized identity matrix.

We now show how to estimate the model using a variational approximation, which we compare to the estimation from a Gibbs sampler and EM algorithm.

4.1 *Variational Approximation*

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

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 $q(\boldsymbol{Y}^*)$ and $q(\boldsymbol{\beta})$. 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.

First, we provide the functional form for the components of the factorized distributions, which are *estimated* as part of the approximation. $q(Y_i^*)$ is a truncated normal distribution, with

$$q(Y_i^*) = \begin{cases} \text{Normal}_{[0,\infty)}(\mu_i, 1) & \text{if } Y_i = 1, \\ \text{Normal}_{(-\infty,0]}(\mu_i, 1) & \text{if } Y_i = 0, \end{cases} \tag{7}$$

and $q(\boldsymbol{\beta})$ is a multivariate normal distribution, with $q(\boldsymbol{\beta}) =$ Multivariate Normal$(\boldsymbol{\beta}, \sum)$. Now, we iteratively update the parameters of the distributions μ_i, $\boldsymbol{\beta}$, $\sum$, 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 $\boldsymbol{\beta}^{\text{old}}$. We set μ_i^{new} to $\mu_i^{\text{new}} = \boldsymbol{X}_i\boldsymbol{\beta}^{\text{old}}$. $\sum$ is not updated during the algorithm and is set to $\boldsymbol{\Sigma} = (\boldsymbol{X}^{'}\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I})^{-1}$. Finally, we set $\boldsymbol{\beta}^{\text{new}} = (\boldsymbol{X}^{'}\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I})^{-1}(\boldsymbol{X}^{'}E[\boldsymbol{Y}^*])$, where E[$\boldsymbol{Y}^*$] is the expected value for each observation's latent propensity, with,

$$\text{E}[Y_i^*] = \begin{cases} \mu_i^{\text{new}} + \frac{\phi_i}{1-\Phi_i} & \text{if } Y_i = 1, \\ \mu_i^{\text{new}} - \frac{\phi_i}{\Phi_i} & \text{if } Y_i = 0, \end{cases} \tag{8}$$

where ϕ_i is equal to $\phi(-\mu_i)$, where ϕ is the normal density and $\Phi_i = \Phi(-\mu_i)$, where Φ is the cumulative normal density.

The terms μ_i and $\boldsymbol{\beta}$ 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 $q(\boldsymbol{\beta})$ to perform inference.

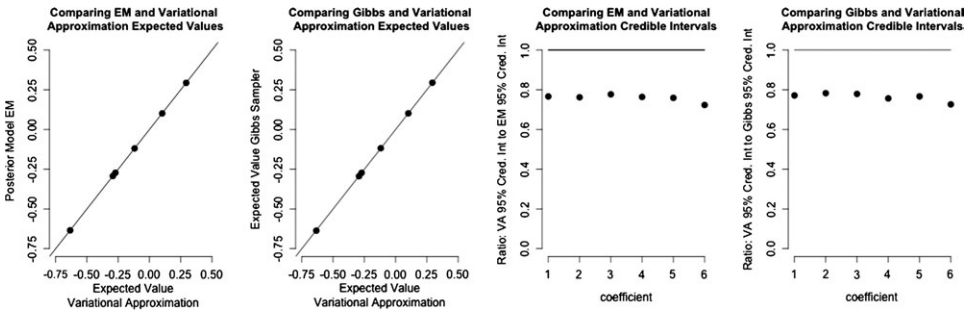

Fig. 1 Comparing variational approximations to Gibbs wampling and the EM qlgorithm. This figure compares the posterior estimates from the Gibbs sampler, the EM algorithm, and the variational approximation. The two left-hand plots demonstrate that all three methods agree on the expected value of the coefficients, 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% 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.

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

We applied the Gibbs sampler, EM algorithm, and variational approximation to a simple simulated data set to compare the properties of the methods. Specifically, we generated 350 observations using a simple six parameter probit model. We 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° 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 they 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 two right-hand plots compare the variational approximation's estimate of uncertainty to the EM and Gibbs sampler's uncertainty estimates. The second plot from the right presents the ratio of the variational approximation's 95% credible interval to the 95% credible interval from the EM algorithm and the right-hand plot compares the 95% credible intervals from the variational approximation and the Gibbs sampler. 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 demonstrates 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 (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.

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

5 A Model of Legislative Voting Blocs

In this section, we 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 who regularly vote together or blocs. The Quinn–Spirling voting-bloc model is an important tool for describing how members of a legislature group together on votes, particularly for legislatures where standard item-response theory methods for ideal point estimation are inapplicable (e.g., the House of Commons in England) (Quinn and Spirling 2010). As we 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 limit 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 $\boldsymbol{\tau}_i$, a $K \times 1$ indicator vector. Each legislator's voting bloc is modeled as a draw from a multinomial distribution, $\boldsymbol{\tau}_i|\boldsymbol{\pi} \sim \text{Multinominal}(1,\boldsymbol{\pi})$, where $\boldsymbol{\pi}$ 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, $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k1}, \boldsymbol{\theta}_{k2}, \ldots, \boldsymbol{\theta}_{kJ})$, 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 Jth roll call, V_{ij} as a draw from a Bernoulli distribution, $V_{ij}|\boldsymbol{\tau}_{ik} = 1,\boldsymbol{\theta}_k \sim \text{Bernoulli}(\boldsymbol{\theta}_{kj})$. We assume that $\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ and that, for all k and j, $\boldsymbol{\theta}_{kj} \sim \text{Beta}(\gamma_1,\gamma_2)$.

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 components provides the same posterior height because information is only available about which observations are grouped together and not the component labels. For example, consider a two-component mixture model. Suppose we arbitrarily label one component as ''component 1'' and the other component as ''component 2'' and that we evaluate the posterior for a set of parameters. Now, suppose that we *switch* the labels: the component previously labeled ''component 1'' is now labeled ''component 2'' and the previous ''component 2'' is now ''component 1''. This relabeling does not change the posterior because the evaluation of the posterior did not depend at the component labels. More generally, if there are K components in a 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 cause the sampler to draw from the wrong posterior and therefore are an unattractive option (McLachlan and Peel 2000). Current recommendations are to run a chain without constraints and then postprocess using a variety of methods (Jasra, Holmes, and Stephens 2005). This is often a useful solution, but some problems with sampling 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 that 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 do 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 $\boldsymbol{\theta}$, $\boldsymbol{\pi}$, and $\boldsymbol{\tau}$ and use the following approximating distribution:

$$q(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\tau}) = q(\boldsymbol{\theta})q(\boldsymbol{\pi})q(\boldsymbol{\tau}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} \prod_{j=1}^{J} q(\boldsymbol{\theta}_{kj}) \prod_{i=1}^{N} q(\boldsymbol{\tau}_i), \tag{9}$$

where the additional independences follow from the assumptions of the model. Our derivation of the algorithm proceeds in two steps. We first provide the distributional forms for $q(\boldsymbol{\theta})$, $q(\boldsymbol{\tau}_i)$, and $q(\boldsymbol{\pi})$. Then, we describe the specific updates of the parameters of these distributions that constitute the update steps.

The distributional form for the components are given by

- $q(\boldsymbol{\tau}_i) = \text{Multinominal}(\boldsymbol{r}_i)$, where $\boldsymbol{r}_i = (\boldsymbol{r}_{i1}, \ldots, \boldsymbol{r}_{iK})$ represents the probability of legislator i belonging to a given bloc,
- $q(\boldsymbol{\pi}) = \text{Dirichlet}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_K$,
- $q(\boldsymbol{\theta}_{kj}) = \text{Beta}(\eta_{kj1}, \eta_{kj2})$, where η_{kj1} and η_{kj2} are the shape parameters for the Beta distribution.

Therefore, an iteration of the variational approximating algorithm will proceed by updating $\boldsymbol{r}_i$, $\boldsymbol{\lambda}_k$, and $\boldsymbol{\theta}_{kj}$. Call $\boldsymbol{\lambda}_k^{\text{old}}$ and $\boldsymbol{\theta}_{kj}^{\text{old}}$, the values from the previous iteration. We then set r_{ik}^{new} to

$$r_{ik}^{\text{new}} \propto \exp\left[\text{E}[\log\, \pi_k] + \sum_{j=1}^{J} \left\{V_{ij}[\text{E}[\log\, \boldsymbol{\theta}_{kj1}]] + \left(1 - V_{ij}\right)[\text{E}[\log\, \theta_{kj1}]]\right\}\right], \tag{10}$$

where $\text{E}[\log\, \pi_k] = \Psi(\lambda_k^{\text{old}}) - \Psi(\sum_{z=1}^{K} \lambda_z^{\text{old}})$, $\text{E}[\log\, \theta_{kj1}] = \Psi(\eta_{kj1}^{\text{old}}) - \Psi(\eta_{kj1}^{\text{old}} + \eta_{kj2}^{\text{old}})$, and $\text{E}[\log\, \boldsymbol{\theta}_{kj2}] = \Psi(\eta_{kj2}^{\text{old}}) - \Psi(\eta_{kj1}^{\text{old}} + \eta_{kj2}^{\text{old}})$. $\Psi(\cdot)$ is the *Digamma* function, the derivative of the *Gamma* function. Next, we set λ_k^{new} to $\lambda_k^{\text{new}} = \alpha_k + \sum_{i=1}^{N} r_{ik}^{\text{new}}$. And finally, we set $\boldsymbol{\theta}_{kj}^{\text{new}}$ to, $\theta_{kj1}^{\text{new}} = \gamma_1 + \sum_{i=1}^{N} r_{ik}^{\text{new}} V_{ij}$ and $\theta_{kj2}^{\text{new}} = \gamma_2 + \sum_{i=1}^{N} r_{ik}^{\text{new}} \left(1 - V_{ij}\right)$.

5.2 *Model Selection*

A difficult problem in mixture models is determining the number of components to include in the mixture. 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 *over fit* the data (Kass and Raftery 1995). To make explicit each model's dependence on the assumed number of blocs, we represent a k component voting-bloc model with M_k. Our goal when selecting the model is to use Bayes's rule to determine the probability of each model, given the data (Bishop 2006). Applying Bayes's rule formalizes this intuition, $p(M_k|\boldsymbol{V}) \propto p(M_k)p(\boldsymbol{V}|M_k)$. Therefore, to calculate the posterior for a particular model, we first need to know the prior probabilities for each model $p(M_k)$. We will assume that each model has the same prior probability, and therefore, 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, $p(\boldsymbol{V}|M_k)$. Direct computation of the evidence is infeasible because we are unable to manipulate the posterior distribution directly. But, as detailed in the supplementary material, the variational approximation has optimized a lower bound for the log evidence, $\log p(\boldsymbol{V}|M_k) \geqslant \mathscr{L}(p)M_k$. For technical reasons, we cannot use this lower bound directly and need to add a correction term due to make the lower bounds comparable (Bishop 2006), so we use $\mathscr{L}\ (q)_{\text{M}_k} = \mathscr{L}(q)_{\text{M}_k} + \log k!$.

The right-hand plot in Fig. 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.

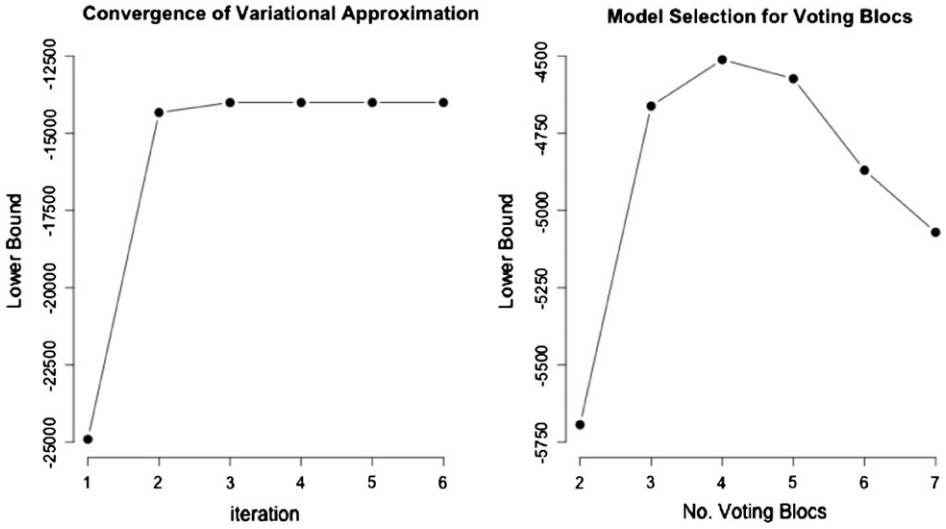

Fig. 2 Convergence to best approximation occurs quickly and the lower bound allows for model selection. This figure demonstrates that the variational approximation to the Quinn-Spirling voting bloc model converges quickly, which is easily assessed using the lower bound on the log probability of the data. Furthermore, 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 U.S. Senate.

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.

The variational approximation approximates the entire posterior, which facilitates the identification of votes that separated Republicans from Democrats and votes that created intraparty 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, whereas 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).

Table 1 Voting blocs in U.S. Senate

Label	*Example senators*	*Distinctive Vote*	*%*
Cons. Rep	Coburn, DeMint, Inhofe, Sessions	Amendment 521: Reduce Federal Debt	37.7
Mod. Rep	Coleman, Hagel, Lugar, Murkowski	Amendment 2662: Prohibit Canyon Funds	12.2
Mod. Dem	Bayh, McCaskill, Lieberman, Ben Nelson	Cloture on S. 2633: Iraq Redeployment	17.0
Lib. Dem	Clinton, Kennedy, Obama, Sanders	Table Amendment 4388: Mortgages	33.0

We can also use the voting-bloc model to identify votes that created the intraparty cleavages. For each of the nonunanimous 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. 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. 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; Antoniak 1974; Blei and Jordan 2006). Heuristically, Dirichlet process priors group together observations with similar characteristics into a countably infinite set of groups. In any one sample, however, the prior uses both the observed data and the 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.

Dirichlet process priors are well known in both the statistical and the machine-learning literature (Gill and Casella 2009; Quinn and Spirling 2010). Furthermore, a wide range of studies across many fields have made use of Dirichlet processes to identify groups of observations with similar characteristics. This includes applications in statistics (e.g., Escobar and West 1995), computer science (e.g., Teh et al. 2006), biology (e.g., Kottas, Branco, and Gelfand 2002; Medvedovic and Sivaganesan 2002), and political science (e.g., Quinn and Spirling 2010). 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

Although 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 are not guaranteed to provide substantively interesting clusterings. Infinite mixture models (like other clustering algorithms) group observations together based an 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 motivated 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 postmodel checks required to establish the utility of finite mixture models (such as in Quinn et al. 2010 and Grimmer 2010)

[3]This distinction is not just found among the 20 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.

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 to 2007 (Grimmer 2010). The Dirichlet process, like other priors, is overwhelmed 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* (Blei and Lafferty 2009; Grimmer 2010; Quinn et al. 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 use a nonparametric topic model to identify groups of press releases that discuss similar topics and we 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, we apply a set of well-established procedures to translate the press releases into count vectors (Manning et al. 2008). 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, $\boldsymbol{y}_i$, where $w = 2796$ 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 α and a base distribution G_0. We write the Dirichlet process distribution as DP (α,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 α determines how close the draws from the distribution are to the base measure: the larger value of α, the closer the draws will be to the base measure. The number of components in the mixture will depend strongly on our selection of α.

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: $\boldsymbol{\theta}_i|G \sim G$. Finally, we draw the observed data from an appropriate distribution $\boldsymbol{y}_i|\boldsymbol{\theta}_i \sim F(\boldsymbol{\theta}_i)$. The draws from the Dirichlet process prior are discrete with probability 1 (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 α). Collect these draws into the infinite length vector $\boldsymbol{v} = (v_1, v_2, \ldots)$. Next, an infinite number of parameters are drawn from a base distribution $\boldsymbol{\theta}_k \sim G_0$, for $k = 1, \ldots, \infty$. To model the press releases, we suppose that G_0 is a Dirichlet distribution, with vector of shape parameters given by $\boldsymbol{\lambda}$. Conditional on $\boldsymbol{v}$ define $\pi(\boldsymbol{v})_k = v_k \prod_{j=1}^{k-1} (1-v_j)$ and call $\boldsymbol{\pi}(\boldsymbol{v}) = (\boldsymbol{\pi}(\boldsymbol{v})_1, \ldots)$. 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)

$$G = \sum_{k=1}^{\infty} \pi(\boldsymbol{v})_k \delta(\boldsymbol{\theta}_k). \tag{11}$$

Equation (11) 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 its mass on a countably infinite set of points, and therefore, we can ''cluster'' the observations by observing the parameters an observation is assigned. The probability of each value of $\boldsymbol{\theta}_k$ occurring is governed by the value of $\boldsymbol{\pi}(\boldsymbol{v})_k$, which is the *stick-breaking* portion of the model. $\boldsymbol{\pi}(\boldsymbol{v})_k$ breaks off a portion of the ''probability stick'' for the kth component.

Because the number of components used in any one application of the Dirichlet process prior depends strongly on α, we place a prior on α and obtain a posterior estimate of the concentration parameter. Blei and Jordan (2006) suggest placing a Gamma distribution as the prior on α. 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,α) distribution, simplifying the update steps.

We suppose that the *topic* of each press release, $\boldsymbol{\tau}_i$, is a draw from a multinomial distribution, $\boldsymbol{\tau}_i|\boldsymbol{\pi}(\boldsymbol{v}) \sim$ Multinominal(1,$\boldsymbol{\pi}(\boldsymbol{v})$). Conditional on press release's topic, we suppose that $\boldsymbol{y}_i|\boldsymbol{\tau}_{ik} = 1,\boldsymbol{\theta} \sim$ Multinominal(n_i, $\boldsymbol{\theta}_k$), where n_i are the total number of words used in the ith press release.

Therefore, we model the press releases using the following posterior

$$\begin{aligned}
\alpha|s_1,s_2 &\sim \text{Gamma}(s_1,s_2),\\
v_k|\alpha &\sim \text{Beta}(1,\alpha) \quad \text{for } k = 1,\ldots,\infty,\\
\theta_k|G_0,\boldsymbol{\lambda} &\sim \text{Dirichlet}(\boldsymbol{\lambda}) \quad \text{for } k = 1,\ldots,\infty,\\
\boldsymbol{\tau}_i|\boldsymbol{\pi}(\boldsymbol{v}) &\sim \text{Multinomial}(1,\boldsymbol{\pi}(\boldsymbol{v})) \quad \text{for } i = 1,\ldots,N,\\
\boldsymbol{y}_i|\tau_{i,k} = 1,\boldsymbol{\theta}_k &\sim \text{Multinomial}(n_i,\boldsymbol{\theta}_k) \quad \text{for } i = 1,\ldots,N.
\end{aligned} \tag{12}$$

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). Furthermore, 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, $\boldsymbol{v}$, $\boldsymbol{\tau}$, $\boldsymbol{\theta}$, and α, assuming that the approximating distribution has the form, $q(\boldsymbol{v},\boldsymbol{\tau},\boldsymbol{\theta},\alpha) = q(\boldsymbol{v})q(\boldsymbol{\tau})q(\boldsymbol{\theta})q(\alpha)$. The modeling assumptions imply that the approximating distribution can be written as follows:

$$q(\boldsymbol{v},\boldsymbol{\tau},\boldsymbol{\theta},\alpha) = \prod_{k=1}^{K-1} q(v_k) \prod_{i=1}^{N} q(\boldsymbol{\tau}_i) \prod_{k=1}^{K} q(\boldsymbol{\theta}_k) q(\alpha). \tag{13}$$

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

To state the algorithm, we first provide the functional forms for each component of the approximating distribution and then describe the specific update steps for the parameters of these distributions. The distributional forms for the approximating distribution are given by

[4]Blei and Jordan (2006) show that the approximation improves very quickly as the number of components included in the approximating distribution increase.

$$\begin{aligned} q(\boldsymbol{\tau}_i) &= \text{Multinomial}(1, \boldsymbol{r}_i), \\ q(\boldsymbol{v}_k) &= \text{Beta}(\gamma_{k,1}, \gamma_{k,2}), \\ q(\boldsymbol{\theta}_k) &= \text{Dirichlet}(\boldsymbol{\eta}_k), \\ q(\alpha) &= \text{Gamma}(w_1, w_2). \end{aligned} \tag{14}$$

Given these parametric forms for the approximating distribution components, the variational approximation proceeds by sequentially updating their parameters, $\boldsymbol{r}_i$, $\gamma_{k,1}$, $\gamma_{k,2}$, $\boldsymbol{\eta}_k$, w_1, w_2. Suppose that the current values of the parameters are given by $\gamma_{k,1}^{\text{old}}, \gamma_{k,2}^{\text{old}}, \boldsymbol{\eta}_k^{\text{old}}, w_1^{\text{old}}, w_2^{\text{old}}$. An iteration of the variational approximation algorithm will update the parameters using the following steps:

1. $$r_{i,k}^{\text{new}} \propto \exp\Big\{\text{E}[\log v_k] + \text{E}[\log(1-v_k)] + y_{i,k}\text{E}[\log \boldsymbol{\theta}_k]\Big\}, \quad \text{where}$$
$$\begin{aligned} \text{E}[\log v_k] &= \Psi(\gamma_{1,k}^{\text{old}}) - \Psi(\gamma_{1,k}^{\text{old}} + \gamma_{2,k}^{\text{old}}), \\ \text{E}[\log(1-v_k)] &= \Psi(\gamma_{2,k}^{\text{old}}) - \Psi(\gamma_{1,k}^{\text{old}} + \gamma_{2,k}^{\text{old}}), \\ \text{E}[\log \boldsymbol{\theta}_k] &= \sum_{j=1}^{w} \Psi(\eta_{k,j}^{\text{old}}) - \Psi\left(\sum_{m=1}^{w} \eta_{k,m}^{\text{old}}\right). \end{aligned} \tag{15}$$

2. $$\boldsymbol{\eta}_k^{\text{new}} = \boldsymbol{\lambda} + \sum_{i=1}^{N} r_{i,k}^{\text{new}} \boldsymbol{y}_i.$$

3. $$\gamma_{k,1}^{\text{new}} = 1 + \sum_{i=1}^{N} r_{i,k}^{\text{new}}.$$

4. $$\gamma_{k,2} = \frac{w_1^{\text{old}}}{w_2^{\text{old}}} + \sum_{i=1}^{N} \sum_{j=k+1}^{N} r_{i,j}^{\text{new}}.$$

5. $$w_1^{\text{new}} = s + K - 1.$$

6. $$w_2^{\text{new}} = s_2 - \sum_{k=1}^{K-1} \left[\Psi(\gamma_{k,2}^{\text{new}}) - \Psi(\gamma_{k,1}^{\text{new}} + \gamma_{k,1}^{\text{new}})\right].$$

6.3 *Political Attention in Senate Press Releases*

We applied the algorithm to the full collection of 64,033 press releases. The variational approximation took approximately 45 min 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 Fig. 3 shows the approximated posterior distribution on the number of topics. To obtain this posterior distribution, we used the variational approximation to generate a posterior distribution on the stick-breaking proportions, $\boldsymbol{\pi}(\boldsymbol{v})$ and then drew topic labels, conditional $\boldsymbol{\pi}(\boldsymbol{v})$. Using this simulation approach, we find that the 95% credible interval on the number of topic stretches from 66 topics to 79.

Although 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 Fig. 3. Here, we present the expected number of documents for each topic k or $\sum_{i=1}^{N} r_{ik}$. Only 30 topics have an expected number of documents greater than 100 and 45 topics are expected to have more than 10 documents. This is 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 focus 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).

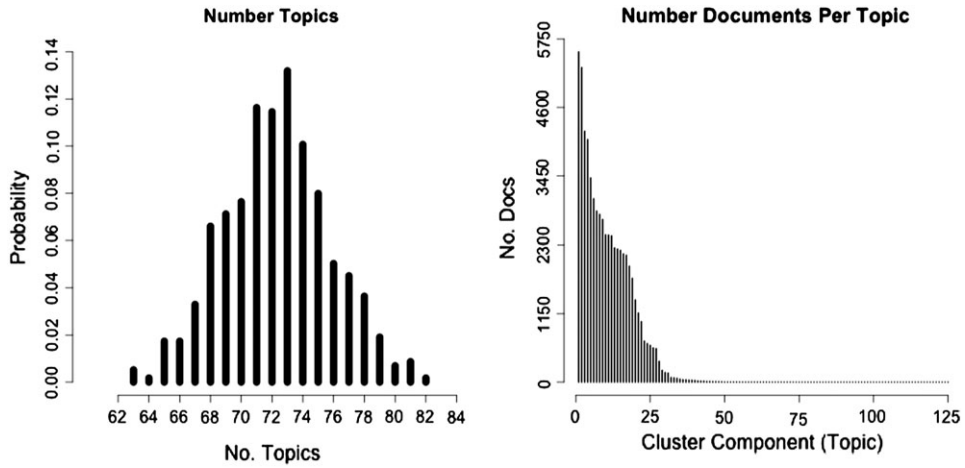

Fig. 3 Number of topics employed and the distribution of documents over topics. This figure presents the distribution on the number of topics and documents per topic, as estimated by the Dirichlet process prior. The left-hand plot shows that the model identifies about 78 topics in the collection of press 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 depends on the modeling characteristics.

The most important quantities 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. These provide one measure of how senators divide their attention when communicating with constituents (Grimmer 2010). 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 10 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, whereas also showing that the model is able to identify substantively interesting topics of press releases.

This section demonstrates how variational approximations make possible the application of nonparametric Bayesian methods to large data sets. This is difficult using MCMC, which tends to converge slowly when estimating topic models applied to large collections of texts (Blei and Lafferty 2006). But variational

Table 2 Ten most discussed topics

Label	*Identifying stems*	*% Press releases*
Appropriations/grants	fund,project,000,million,water,transport,develop,improv,airport,citi	8.6
Honorary	honor,servic,school,serv,american,veteran,academi,famili,student,world	8.2
Iraq war	iraq,troop,war,iraqi,american,militari,polit,secur,support,countri	6.6
Health grants	health,program,educ,children,school,fund,student,care,servic,000	6.3
Homeland security	secur,homeland,port,border,depart,fund,guard,air,servic,transport	5.3
Judicial nominations	court,vote,justic,american,judg,case,hous,congress,constitut,protect	4.8
Hurricanes/disasters	disast,assist,hurrican,fema,flood,damag,fund,katrina,storm,declar	4.5
Taxes	tax,american,budget,social,secur,wage,famili,worker,increas,benefit	4.4
Defense projects	million,defens,fund,air,militari,base,facil,guard,armi,project	4.2
Health policy	health,care,drug,medicar,senior,prescript,plan,medic,program,cost	3.8

approximations provide a fast approximation to the true posterior, providing useful insights into what members of Congress communicate with their constituents.

7 Conclusions

In this paper, we 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 intraparty cleavages in the U.S. 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 make 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. As a result, political scientists using variational approximations will be to make inference about politics that would otherwise be impossible.

References

Ansolabehere, Stephen, Jonathan Rodden, and James Snyder, Jr. 2008. The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review* 102:215–32.

Antoniak, C. E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* 2:1152–74.

Bishop, Christopher. 2006. *Pattern recognition and machine learning*. New York: Springer.

Blei, David, and Michael Jordan. 2006. Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis* 1:121–44.

Blei, David, and John Lafferty. 2006. Dynamic Topic Models. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA.

Blei, David, and John Lafferty. 2009. *Text mining: Theory and applications, chapter topic models*. Oxford, UK: Taylor and Francis.

Celeux, G., M. Hurn, and C. P. Robert. 2000. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* 95:957.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review* 98:355–70.

Cowles, Mary Kathryn, and Bradley Carlin. 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* 91:883–904.

Dempster, Arthur, Nathan Laird, and Donald Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39:1–38.

Escobar, Michael, and Mike West. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90:577–88.

Fenno, Richard. 1978. *Home style: House members in their districts*. Boston: Addison Wesley.

Ferguson, Thomas. 1973. Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1:209–30.

Gelfand, Alan, and A. F. M. Smith. 1990. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85:398–409.

Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin. 1995. *Bayesian data analysis*. New York: Chapman & Hall.

Gelman, Andrew, and Donald Rubin. 1992. Inference from iterative simulation: Simulation using multiple sequences. *Statistical Science* 7:457–72.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–41.

Ghahramani, Zoubin, and Matthew Beal. 2001. Propagation algorithms for variational Bayesian learning. *Advances in Neural Information Processing Systems* 13:852–59.

Gill, Jeff. 2004. Is partial-dimension convergence a problem for inferences from MCMC algorithms? *Political Analysis* 12:153–78.

Gill, Jeff, and George Casella. 2004. Dynamic tempered transitions for exploring multimodal posterior distributions. *Political Analysis* 12:425.

———. 2009. Nonparametric priors for ordinal Bayesian Social Science models: Specification and estimation. *Journal of the American Statistical Association* 104:453–54.

Gill, Jeff, and Lee Walker. 2005. Elicited priors for Bayesian model specifications in political science research. *Journal of Politics* 67:841–72.
Grimmer, Justin. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis* 18:1–35.
Hoff, Peter, and Michael Ward. 2004. Modeling dependencies in international relations networks. *Political Analysis* 12:160–75.
Jackman, Simon. 2000. Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science* 44:375–404.
Jasra, A., C. C. Holmes, and D. A. Stephens. 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture models. *Statistical Science* 20:50.
Jordan, Michael, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence Saul. 1999. An introduction to variational methods for graphical models. *Machine Learning* 37:183–233.
Kass, Robert, and Adrian Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90:773–95.
King, Gary. 1998. *Unifying political methodology: The likelihood theory of statistical inference*. Ann Arbor: University of Michigan Press.
Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220:671.
Kottas, A., M. D. Branco, and A. E. Gelfand. 2002. A nonparametric Bayesian modeling approach for cytogenetic dosimetry. *Biometrics* 58:593–600.
Lax, Jeffrey, and Justin Phillips. 2009. Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review* 103:367–86.
Londregan, John. 2000. Estimating legislators' preferred points. *Political Analysis* 8:35–56.
MacKay, David. 2003. *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.
Manning, Christopher, Pabhakar Raghavan, and Hinrich Shutze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.
Martin, Andrew, Kevin Quinn, and Jong Hee Park. Markov chain Monte Carlo package (MCMCpack). *Journal of Statistical Software*. Forthcoming.
McLachlan, Geoffrey, and David Peel. 2000. *Finite mixture models*. San Francisco: John Wiley & Sons.
Medvedovic, M., and S. Sivaganesan. 2002. Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics* 18:1194–206.
Neal, Radford. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 9:249–65.
Petrone, S., and A. Raftery. 1997. A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics & Probability Letters* 36:69–83.
Quinn, Kevin, and Arthur Spirling. 2010. Identifying intra-party voting blocs in the UK house of commons. *Journal of the American Statistical Association.* Forthcoming.
Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54:209–28.
Sethuraman, Jayaram. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica* 4:639–50.
Teh, Yee Weh. 2010. Dirichlet processes. In *Encyclopedia of machine learning*, eds. Claude Sammut and Geoffrey Webb. New York: Springer.
Teh, Yee Weh, Michael Jordan, Matthew Beal, and David Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101:1566–81.
Teh, Y. W., D. Newman, and M. Welling. 2007. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Advances in Neural Information Processing Systems* 19:1353.
Trier, Shawn, and Simon Jackman. 2008. Democracy as a latent variable. *American Journal of Political Science* 52:201–17.
Wang, Bo, and D. M. Titterington. 2004. Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* 20:577–84.
Western, Bruce, and Simon Jackman. 1994. Bayesian inference for comparative research. *American Political Science Review* 88:412–23.