Political Analysis (2004) 12:323–337
doi:10.1093/pan/mph021

Introduction to the Special Issue

Jeff Gill
Department of Political Science, University of California, Davis,
One Shields Avenue, Davis, CA 95616
e-mail: jgill@ucdavis.edu

1 Introduction

Welcome to the special issue of *Political Analysis* dedicated to Bayesian methods. We hope that you enjoy the varied and interesting contributions herein featuring Bayesian statistical methods. For many people in empirical political science, Bayesian statistics seems like a weird offshoot of probability that surfaces occasionally in journals and books but does not occupy a particularly central role. This perception appears to be changing. In fact, it appears to be changing quite rapidly. The purpose of this issue is to support and accelerate this momentum by further demonstrating the full flexibility and power of Bayesian methodology.

So why this change? Why are people suddenly more interested in developing Bayesian models in their own research? The first apparent reason for this change is that the Bayesian model specifications have distinct advantages over traditional alternatives, such as probabilistic descriptions of model results and the systematic incorporation of prior information. The second reason for this change has to do with computing. Prior to 1990, when the watershed review article by Gelfand and Smith appeared, statisticians working outside of statistical physics and image restoration (i.e., the vast majority) were unaware of a flexible set of estimation tools based on Markov chains. Two primary tools were described in that article: the Metropolis-Hastings algorithm from the 1953 article by Metropolis et al. (although curiously only the Hastings 1970 paper is directly cited, p. 400), and the Gibbs sampler from the 1984 article by Geman and Geman. These were relatively inaccessible pieces published in the *Journal of Chemical Physics* and *IEEE Transactions on Pattern Analysis and Machine Intelligence,* respectively. What Gelfand and Smith demonstrated was that these are actually very powerful *general* tools for describing posterior distributions of interest and subsequently producing inferential statements when standard analytical methods are difficult or impossible. As a result, we see an explosion of Bayesian work by previously frustrated researchers in the leading statistics journals of the 1990s who apply Markov chain Monte Carlo (MCMC) to otherwise intractable problems.

Now we come to the social sciences, and political science in particular. Gradually over the last five years or so Bayesian applications (generally using MCMC) have appeared. Through the work of Gelman and King (1994), Bartels (1996, 1997), Smith (1998, 1999), Western (1998), Quinn et al. (1999), Jackman (2000a, 2000b, 2001), Clarke (2001), Hill and Kriesi (2001), Martin and Quinn (2002), and others, we have seen a steady increase in awareness among general readers. An underlying theme in this work so far is that

important model characteristics cannot be developed without a fully Bayesian specification. A particularly appealing feature of Bayesian model specifications is the direct inclusion of prior information that allows political scientists to recognize divergent theoretical perspectives, expert opinion, and previous work in the studied area.

It is important to note at this point that Bayesian methods are not just another "fad" sweeping through social science methodology like Lisrel or HLM. As opposed to picking up another tool or technique, Bayesian methods require a philosophical commitment on behalf of the researcher. This commitment essentially boils down to accepting two basic premises: (1) phenomena of interest are uncertain and changing, and (2) available prior information should be used in model specifications. Both of these underlying principles are well suited to research in the social sciences. First, it is rare to find authors asserting that estimated quantities are fixed and unchanging in the real political world when there are variables such as political ideology, probability of going to war, stability of governments, and legislative productivity. Second, we all commence model specification with extensive substantive knowledge about the problem (consider, for instance, the ubiquitous footnote about coefficients that are "signed in the expected direction"). So unlike researchers in certain natural science fields, we view a shifting, uncertain world in which change is the norm but previous observations and previous scholarly findings provide a substantial guide to theory and conjecture.

Bayesian inference actually predates the classical approaches of Neyman-Pearson frequentism and Fisher likelihoodism. These powerful giants of early-twentieth-century statistics were openly hostile to "inverse probability," and researchers were therefore discouraged from applying Bayesian methods for quite some time. Interestingly, Fisher (1935) created *fiducial inference,* which was an attempt to apply inverse probability without uniform priors. This approach failed to do what Fisher wanted, and Lindley (1958) eventually proved that fiducial inference is consistent only when it is made equivalent to Bayesian inference with a uniform prior. Fortunately, Bayesians such as Good (1950), Savage (1954, 1962), Jeffreys (1961), Lindley (1961, 1965) and de Finetti (1972, 1974, 1975) preserved interest throughout the middle of the century. One of the hallmarks of this dark era was that many of the Bayesian specifications, while arguably superior in theoretical foundation, led to mathematical forms that were intractable. The problem of marginalizing difficult multidimensional integrals was eventually solved by MCMC techniques, leading to the current Bayesian renaissance. So Bayesian statistics is increasingly popular at the start of the twenty-first century because it has finally outlived and outlasted active hostility by influential figures and because the computational tools for general estimation have only recently became available and easy to use.

2 Bayesian Mechanics

The core philosophical foundation of Bayesian inference in statistics is the consideration of both observables and parameters as random quantities for description. In practice, all observed quantities are treated as fixed to be conditioned on, and all unobserved quantities are assumed to possess distributional qualities to be treated as random variables. Unobserved quantities can be both parameters to be estimated as well as missing data. Thus underlying parameter values are now no longer treated as fixed and unmoving in the total population, and all statements are made in probabilistic terms.

The Bayesian inference process begins with explicitly assigning prior information to the unknown quantities from sources that can be empirical, qualitative, narrative, statistical, or intuitive. These *prior distributions* range from very informative descriptions

based on previous research in the field to deliberately vague and uncertain forms that reflect high levels of uncertainty or previous ignorance. Furthermore, this prior distribution is not seen as an inconvenience imposed by the treatment of unknown quantities; it is the means by which existing knowledge is systematically included in the model.

Next, a likelihood function is specified in the conventional manner by assigning a parametric form and plugging in the observed data. The third step produces a *posterior distribution* by multiplying the prior distribution by the likelihood function. In this manner, the likelihood function uses the data to update the specified prior knowledge conditionally on the data (through the likelihood function):

$$\text{Posterior Probability} \propto \text{Prior Probability} \times \text{Likelihood Function}.$$

This is just Bayes's law, in which the denominator on the right-hand side has been ignored by using proportionality. Stated more formally, and including the proportionality constant in the denominator, we have

$$\pi(\beta \mid \mathbf{X}) = \frac{p(\beta)L(\beta \mid \mathbf{X})}{\int_{\mathfrak{B}} p(\beta)L(\beta \mid \mathbf{X})d\beta}\ , \tag{1}$$

where β is the parameter vector of interest with defined support $\mathfrak{B}$, and $\mathbf{X}$ represents the data. Standard notation gives $p(\beta)$ as the prior distribution on β, and $L(\beta \mid \mathbf{X})d\beta$ as the likelihood function. Thus by conditioning on the data through the likelihood function, we are updating the information contained in the prior distribution.

The idea of Bayesian updating is more general than it first appears. Since prior information comes from virtually any source, we can treat the current posterior distribution as a prior should new data be observed that we would like to condition on. Obviously this process of updating continues indefinitely, or as long as we like, with old posteriors becoming new priors and our level of knowledge constantly improving. A very neat consequence of Bayesian updating is that the final posterior distribution after a series of such updating steps is identical to the posterior constructed in the standard way as if all the data had arrived at once.

The final step is to evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions, including the prior distribution. This is often done by varying the prior in a systematic or ad hoc manner and observing the magnitude of changes in the posterior. Global sensitivity analysis evaluates a wide range of alternative prior specifications, forms of the link function, missing data implications, error sensitivity, and perturbations of the likelihood and prior specifications. Local sensitivity analysis is the more modest and realizable process of making minor changes in the prior parameterization (but generally keeping the same parametric form) while looking at the resulting posterior effects.

When the researcher is satisfied with the fit of the model and the range of assumptions, the results are then described to readers. Unlike the null hypothesis significance test (NHST) method of deciding strength of conclusions based on the magnitude of p values, evidence is presented in the Bayesian inference process by simply summarizing the posterior distribution, and therefore there is no artificial decision based on the assumption of a true null hypothesis. The posterior summary is typically given by quantiles and probability statements such as the probability that the parameter occupies some region of its support: $p(\beta \in [\beta_L : \beta_H])$.

The most useful interval measure for model results is the *highest posterior density* (HPD) region, which is the Bayesian version of a confidence interval. The HPD region is the (possibly multidimensional) region of the posterior distribution with the highest probability at some threshold, regardless of whether or not it is contiguous. More formally, the $100(1-\alpha)\%$ HPD region is the subset of the support of the posterior distribution for some parameter, β, that meets the criteria:

$$C = \{\beta : \pi(\beta \mid \mathbf{X}) \geq k\},$$

where k is the largest number such that

$$1 - \alpha = \int_{\beta:\pi(\beta \mid \mathbf{X})>k} \pi(\beta \mid \mathbf{X}) d\beta$$

(Casella and Berger 2001, p. 448). Unlike the analogous frequentist construct, we treat β as a random quantity and therefore do not have to be bothered with nebulous concepts like *confidence* here. This means that we can speak probabilistically: "The probability that the posterior mean is greater than zero is ...," etc.

We can also consider the *posterior predictive distribution* as a way to check model integrity and to make predictions if desired. Start with the prior predictive distribution of a new data value, x_{new}, before observing the full data set:

$$p(x_{new}) = \int_{\mathfrak{B}} p(x_{new}, \beta) d\beta = \int_{\mathfrak{B}} p(x_{new} \mid \beta) p(\beta) d\beta.$$

This is just the marginal distribution of an unobserved data value from the product of the prior for β and the PDF or PMF, integrating out this parameter. This makes intuitive sense because uncertainty in β is averaged out to reveal a distribution for the data point. More useful is the distribution of a new data point, x_{new}, *after* the data, $\mathbf{X}$, have been observed, which is the posterior predictive distribution, calculated by

$$\begin{aligned} p(x_{new} \mid \mathbf{X}) &= \int_{\mathfrak{B}} p(x_{new}, \beta \mid \mathbf{X}) d\beta = \int_{\mathfrak{B}} \frac{p(x_{new}, \beta \mid \mathbf{X})}{p(\beta \mid \mathbf{X})} p(\beta \mid \mathbf{X}) d\beta \\ &= \int_{\mathfrak{B}} p(x_{new} \mid \beta, \mathbf{X}) p(\beta \mid \mathbf{X}) d\beta. \end{aligned}$$

This can be simplified since x_{new} and $\mathbf{X}$ are assumed independent:

$$= \int_{\mathfrak{B}} p(x_{new} \mid \beta) p(\beta \mid \mathbf{X}) d\beta.$$

Because of the integral the posterior predictive distribution is the product of the single variable PDF or PMF times the full data likelihood in which we integrate over uncertainty in β to give a probability statement that is dependent on the observed data only. Now the degree to which this predicted distribution differs from observed data is a measure of model fit.

It is also possible to compare models with the *Bayes factor*. This is easily related to conventional hypothesis testing since one of them can be a "null" specification. Suppose

we wish to test two competing (not necessarily nested) models, M_1 and M_2, for explaining the same data $\mathbf{X}$, with corresponding estimated coefficient vectors β_1 and β_2. The posterior odds ratio in favor of Model 1 versus Model 2, incorporating both prior and posterior information, is produced by Bayes's law:

$$\underbrace{\frac{\pi(M_1 \mid \mathbf{X})}{\pi(M_2 \mid \mathbf{X})}}_{\text{posterior odds}} = \underbrace{\frac{p(M_1)/p(\mathbf{X})}{p(M_2)/p(\mathbf{X})}}_{\text{prior odds/data}} \times \underbrace{\frac{\int_{\beta_1} f_1(\mathbf{X} \mid \beta_1) p_1(\beta_1) d\beta_1}{\int_{\beta_2} f_2(\mathbf{X} \mid \beta_2) p_2(\beta_2) d\beta_2}}_{\text{Bayes factor}}.$$

By rearranging we get the standard form of the Bayes factor, which can be thought of as the magnitude of the evidence for Model 1 over Model 2, contained in the data:

$$B(\mathbf{X}) = \frac{\pi(M_1 \mid \mathbf{X})/p(M_1)}{\pi(M_2 \mid \mathbf{X})/p(M_2)}.$$

Interestingly, for nested models with the same priors, the Bayes factor reduces to a standard likelihood ratio.

3 Relation to Classical Methods

By classical methods here we mean both *frequentist* and *likelihoodist* approaches. A frequentist posits a very large number of repeated trials of the same experiment and estimates (assumed) fixed, but unknown, population parameters by conditioning on these parameters and integrating over the observed data. A likelihoodist, like the Bayesian, assumes that data are fixed once observed, but finds that all useful information for estimating the unknown parameters is contained in the likelihood function. In contrast, the Bayesian balances information between the prior and the likelihood function, conditioning on the data and integrating over the parameters. Thus the greatest distance actually lies between the Bayesian perspective and the canonical Neyman-Pearson frequentist view.

Since the Neyman-Pearson frequentist setup is not one that social scientists generally prescribe to directly (our standard testing method is an inconsistent blend of Fisher and Neyman-Pearson, something I have been quixotically complaining about for some time [Gill 1999, 2001, 2002]), more appropriate comparisons are drawn with Fisher's *test of significance.* A noticeable difference is the means by which results are described. Rather than give the mode of the likelihood function and the curvature around it, the Bayesian approach is intended to more fully describe the posterior in probabilistic terms. Therefore it is counter to the Bayesian mentality to talk simply about "the value" of some parameter, since parameters possess distributional rather than fixed properties.

A common question asked by skeptics and enthusiasts is to what degree Bayesian estimates resemble traditional likelihood estimates. Since the posterior is, by definition, a data-weighted compromise between the prior and the likelihood, the Bayesian posterior resembles a classical sampling distribution to the degree to which the likelihood swamps the prior. This happens in two important circumstances. First, if the information in the prior is weak relative to the likelihood then the latter will dominate. For this reason researchers sometimes deliberately establish weak priors in the form of uniforms, dispersed normals, or other forms in order to let the data dominate. Furthermore, the Bayesian posterior can be setup as equal to the Fisherian sampling distribution if the appropriate uniform prior is used. Often such diffuse priors are used with nuisance parameters where it is not worth the effort to specify highly informed prior distributions.

Second, as the data size increases the likelihood progressively dominates the prior, and in the limit the prior is immaterial. This means that the maximum likelihood estimate is asymptotically equal to a Bayesian posterior mean for any proper (noninfinite density) prior. There is a great amount of statistical theory behind these relations. For instance, Freedman (1963, 1965) and Diaconis and Freedman (1986) give mathematically rigorous conditions for the consistency of Bayesian estimates in standard terms.

The primary differences are seen in small sample problems in which the asymptotic equivalence is not applicable. A common frequentist criticism of the Bayesian approach in these settings is that subjective priors have a great impact on the posterior distribution. However, since the prior is an overt, integral part of the model development process, researchers must be direct and clear about this part of the specification. There is nothing inherently wrong with the prior dominating likelihood function as long as the prior is defensible. As Western and Jackman (1994) note, certain literatures are destined to have small samples and therefore important priors. There is also a developing literature on robust Bayesian analysis specifically focused on producing estimators that are insensitive to a wide range of possible prior distributions (Berger 1990).

4 Obtaining Marginal Posterior Distributions through Stochastic Simulation

As noted, Markov chain Monte Carlo techniques solve a lingering problem in Bayesian analysis. Often Bayesian model specifications produce joint posterior expressions that are analytically intractable. The core principle behind MCMC techniques is that if an iterative chain of consecutive values can be carefully set up and run long enough, then *empirical* summaries of quantities of interest can be obtained from chain values. So to marginalize multidimensional probability structures (such as desired posteriors), we start a Markov chain in the appropriate sample space and let it run until it settles into the target distribution. When it runs for a time confined to this particular distribution, we can collect summary statistics such as means, variances, and quantiles from the simulated values. So the process replaces usually difficult or impossible analytical work with empirical summaries from the simulated values. As long as the simulated values are from the distribution of interest, there is no qualitative difference in the answers.

The most common method of producing Markov chains for MCMC work is the Gibbs sampler (the default mechanism in the package *WinBUGS*), which produces an empirical estimate of the marginal posterior distributions by iteratively sampling from *full conditional distributions* for each parameter. The Gibbs method is popular because it is usually easy to stipulate these conditional specifications: the distribution for each parameter when candidate or real values are established for all others.

For convenience define β as a k-dimensional vector of unknown parameters. Call $\beta_{[i]}$ the β vector where the i^{th} parameter is omitted. The Gibbs sampler draws from the complete conditional distribution for the "left-out" value: $\pi(\beta_i \mid \beta_{[i]})$, repeating for each value in the vector each time conditioning on the most recent draw of the other parameters. When each of the parameters has been updated in this way, then the cycle recommences with the completely new vector β.

This procedure will eventually converge permanently to a limiting (stationary) distribution that is the target posterior, provided that the Markov chain is *ergodic*. Ergodicity results from aperiodicity plus positive recurrence of the Markov chain. Aperiodic chains have no defined pattern whereby they repeat the same series of values in any arbitrary period. A Markov chain is recurrent if it is defined on an irreducible state space such that every substate can be reached from every other substate, and a Markov

chain is *positive* recurrent if the mean time to transition back to the same state is finite. The ergodic theorem is foundational to MCMC work since it is essentially the strong law of large numbers in a Markov chain sense: the mean of chain values converge almost surely to strongly consistent estimates of the parameters of the limiting distribution, despite mild dependence (on some state space $S \in \Re$ for a given transition kernel and initial distribution). These properties for the Gibbs sampler are well known and are described in detail elsewhere (Robert and Casella 2004).

Unfortunately the word *eventually* as used above in claiming convergence is a big caveat, and a large part of the MCMC literature in statistics journals focuses on understanding and assessing Markov chain limiting behavior. Two primary philosophies compete for adherents among applied researchers. Gelman and Rubin (1992) suggest using the EM algorithm (or some variant) to find the mode or modes of the posterior, then using overdispersed points throughout the posterior as a starting point for multiple chains. Convergence is then assessed by comparing within-chain variance against between-chain variance in a standard ANOVA manner with the idea that at convergence, variability within each chain should be similar and will resemble the estimated target variance. Conversely, Geyer (1992) recommends implementing one long chain and using well-known time series statistics to assess convergence. In this vein, Geweke (1992) suggests a difference of means test using an asymptotic approximation of the standard error for the difference. Since the test statistic is asymptotically standard normal, then for long chains, large values imply nonconvergence. In practice, most experienced users perform some combination of diagnostic approaches.

5 A Brief Illustrative Application

As stated, the two primary advantages of Bayesian models are systematic integration of prior information and probabilistic treatment of all unknown quantities. The former is considerably more obvious, so this example highlights the value of the latter. The previously described method of comparing models with the Bayes factor is somewhat limiting in its restriction to two candidates. This could be extended with iterative comparisons but the process would get quite tiresome, and a more generalized process is desired.

Suppose we are in the process of determining a final model specification from among several, possibly many, alternatives. Classical methods say relatively little about this process except in cases in which nested models are tested with likelihood ratios. Furthermore, most authors in political science admit only one specification in their finished work, despite admonitions by Leamer (1978, 1983) and others that competing models can contain important information about parameter reliability and model fit. It turns out that the probabilistic foundation here, which is literally not allowed in classical analysis, provides an effective means of making such comparisons and evaluations.

We need to choose between alternative *not necessarily nested* models: $M_1, M_2, \ldots, M_K$. For each $k = 1 : K$, we determine a model prior, $p(M_k)$, which represents our a priori belief in this model. These priors can be determined from assertions in the relevant substantive literature or by our own beliefs, or can be left deliberately vague by specifying uniform probabilities, $\frac{1}{K}$. Note also that model priors are distinct from the prior distributions assigned to each coefficient within the individual models (part of what determines their identity) and are necessary only when we wish to develop a systematic comparison between models. Also, by convention these model priors sum to one for the set of models tested, hence the (unfortunately) standard situation in which specifying only a single model is equivalent to a prior probability of one for the model and zero for all possible alternatives.

For comparative purposes, we can use a simple averaging scheme from Raftery (1995). The standard posterior distribution in alternative models can be reexpressed to explicitly note the dependence on the model specification choice:

$$\pi(\beta \mid \mathbf{X}, M_k) = \frac{p(\beta \mid M_k)L(\beta \mid \mathbf{X}, M_k)}{p(\mathbf{X} \mid M_k)},$$

the k^{th} model in this case (recall that β_k is the varying length coefficient vector corresponding to the k^{th} model only). The *integrated likelihood* is the denominator of Bayes law calculated here by

$$p(\mathbf{X} \mid M_k) = \int \underbrace{\ell(\beta_k \mid M_k, \mathbf{X})p(\beta_k \mid M_k)}_{\text{likelihood} \times \text{prior}} d\beta_k$$

(also called the *marginal likelihood*, the *marginal probability of the data*, or the *predictive probability of the data*). For additional details on the setup for Bayesian model averaging see Bartels (1997), Raftery et al. (1997), or Hoeting et al. (1999) and the references contained therein.

Since we treat parameters and models in probabilistic terms, it is possible to continue to use standard probability calculus to produce quantities of interest. For instance, using Bayes law we can produce the posterior probability of any model of interest, given the set of models evaluated:

$$p(M_k \mid \mathbf{X}) = \frac{p(M_k)p(\mathbf{X} \mid M_k)}{\sum_{\ell=1}^{K} p(M_\ell)p(\mathbf{X} \mid M_\ell)},$$

which is just a generalization of the Bayes factor to accommodate more reference models in the denominator.

Often, though, it is parameter posterior information that is of greater interest, and we would certainly want a way to average these posterior distributions across model specifications. Consider now

$$P(\beta_j \neq 0 \mid \mathbf{X}) = \sum_{\beta_j \in M_k} p(M_k \mid \mathbf{X}),$$

where the notation $\beta_j \in M_k$ requires summation only over models where B_j is included in the specification. This form is essentially just the posterior probability that β_j is reliable in the "true" model conditional on the model set compared. Conditional the assumption that $\beta_j \neq 0$, we can analyze the posterior mean and variance for the j^{th} coefficient of the k^{th} model:

$$E[\beta_j \mid \mathbf{X}] \approx \sum_{i=1}^{K} \hat{\beta}_j(k)p(M_k \mid \mathbf{X})$$

and

$$Var[\beta_j \mid \mathbf{X}] \approx \sum_{i=1}^{K}[(Var_{\beta_j}(k) + \hat{\beta}_j(k)^2)p(M_k \mid \mathbf{X})) - E[\beta_j \mid \mathbf{X}]^2],$$

where the approximation notation is used to account for categorical granularity. So what we are left with is the probability that the explanatory variable of interest actually matters

to us averaging across models, and the expectation and statistical reliability also both averaged across models. Note again that we are now considering models and posterior coefficients in a direct probabilistic way.

As an empirical example of how this is useful, we will highlight here Raftery's (1995) reanalysis of Ehrlich's 1933–1969 state-level crime data. Ehrlich (1975, 1977) provides famous and highly criticized studies of the economic motivations for crime, which included the first published instance of a multivariate linear model on deterrence. The outcome variable is crime rate, and 15 possible explanatory variables are considered: % young male, south, education, police 1960, police 1969, labor participation, sex ratio, population, nonwhites, unemployment 14–24, unemployment 35–39, GDP, inequality, probability of prison, and prison time. Therefore with $K = 15$ explanatory variables, there are 32,768 possible model specifications calculated from $Num = \sum_{r=0}^{K} \binom{K}{r}$ (assuming no interactions, polynomial, or time series components).

Raftery (1995) considers 14 possible model specifications (all with diffuse prior distributions on the coefficients) to produce the results given in Table 1. This table actually contains quite a bit of information. The black boxes indicate variable inclusion in the numbered models. A single model specification is therefore understood by looking down a particular column. It is easy to see that two variables (education and inequality) are incorporated in every model and four variables (south, labor participation, sex ratio, GDP) are incorporated in none. Since this is a linear model R^2 values are provided, and more informatively below them is the model probability ($p(M_k \mid \mathbf{X})$). These latter values necessarily sum to one (subject to rounding) for all of the models *specifically incorporated*, since nonincluded models are explicitly given zero prior probability. Certainly the first specification stands out according to this criterion as well as by *BIC* [*BIC* is the Bayesian analog of the Akaike Information Criterion, calculated by $-2\ell(\hat{\beta} \mid \mathbf{X}) + p\log(n)$, for p explanatory variables and n observations].

On the right-hand side $P(\beta_j \neq 0 \mid \mathbf{X})$ gives the probability that the j^{th} coefficient is statistically "reliable" conditional on the set of models. That is, given the specifications analyzed, what is the probability that this variable is nonzero in the traditional Fisherian sense? This quantity is somewhat less useful than the others in the table since it is substantially (but not completely) affected by the number of models that the researcher selected to include this variable. More interestingly we can see in the following columns the model-averaged coefficient value and its standard error. Choosing to view these in the traditional manner (rather than through Bayesian posterior description), we can see that they all meet the 95% level of statistical reliability. Actually, two coefficients (population and prison time) have 95% HPD intervals that edge up to (but do not cross) zero. So the worst thing we could say is that 95% of the posterior density for these two parameters is bounded away from zero.

The key point from this brief example is that there is great advantage in being able to treat model quantities probabilistically. Such an approach is intuitive and it facilitates comparison in a way that classical models do not permit. While this is a somewhat stylized example provided to show high-level points, it is also a starting point for more nuanced and flexible approaches. For instance, some authors treat model space as a continuous metric rather than a discrete one, with the objective of integrating over all of the alternatives rather than just a chosen set. So if we were interested in making predictions about some variable of interest ($\mathbf{y}$) across models, then the weighted average predictive distribution is given by

$$p(\mathbf{y}|\mathbf{X}) = \int_{\mathfrak{B}} \int_{\mathfrak{M}} p(\mathbf{y} \mid \mathbf{X}, \beta, M) p(\beta \mid \mathbf{X}, M) p(M \mid \mathbf{X}) d\beta dM,$$

Table 1 Posterior summary of models

Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	$P(\beta_j \neq 0 \mid \mathbf{X})$	$E[\beta_j \mid \mathbf{X}]$	$\sqrt{Var[\beta_j \mid \mathbf{X}]}$
% young male	■	■	■	■	■	■	■		■	■		■	■	■	0.94	1.40	0.50
South																	
Education	■	■	■	■	■	■	■	■	■	■	■	■	■	■	1.00	2.12	0.50
Police 1960	■	■		■	■		■	■	■			■		■	0.76	0.95	0.20
Police 1969			■			■				■	■		■		0.24	0.97	0.19
Labor part.																	
Sex ratio																	
Population						■		■			■				0.12	−0.08	0.04
Nonwhites	■	■	■	■		■		■	■	■	■	■			0.83	0.10	0.04
Unemp. 14–24																	
Unemp. 35–39	■	■	■		■		■						■		0.68	0.32	0.13
GDP																	
Inequality	■	■	■	■	■	■	■	■	■	■	■	■	■	■	1.00	1.33	0.32
Prob. prison	■	■	■	■	■	■	■	■	■	■	■	■		■	0.98	−0.24	0.10
Prison time	■			■						■					0.35	−0.30	0.15
R^2	0.84	0.83	0.82	0.82	0.80	0.82	0.80	0.80	0.80	0.81	0.79	0.79	0.78	0.78			
$p(M_k \mid \mathbf{X})$	0.24	0.18	0.11	0.08	0.08	0.06	0.05	0.04	0.04	0.03	0.02	0.02	0.02	0.02			
$BIC + 60$	4.1	4.6	5.5	6.2	6.4	6.9	7.3	7.6	7.6	7.5	8.7	8.8	9.1	9.1			

which is the posterior probability of **y** times the probability of the model and the coefficient posterior, integrating across coefficient and model space (cf. Draper 1995). Another group of authors (Spiegelhalter et al. 2002) take a slightly different (but probabilistic) approach to obtain a measure for the effective number of parameters in a model based on comparing the difference between posterior means of deviances and the deviances at the posterior means of parameters. Accordingly, specification and parameter decisions are simultaneously considered.

6 Concluding Points

Bayesian methods are not a panacea, and Bayesian approaches are not going to solve every problem that one encounters. What is offered is a grounded way of thinking about probability models and inference that is typically more flexible than known alternatives. Consider some important features of Bayesian inference:

- *A tradition of specifying overt and clear model assumptions.* By convention and necessity, prior information and posterior uncertainty are given in Bayesian research as direct statements. That is, it is necessary to fully describe to readers how one's prior beliefs influence the model results and how sensitive these model results are to changes in those prior beliefs. Classical researchers are "let off the hook" here even though their results are based on prior beliefs as well, a characteristic called *incoherence* in the Bayesian literature (Cornfield 1969).
- *A rigorous way to make probability statements about the real quantities of interest.* Consider the murky meaning of "confidence" in classical models. Anyone who has lectured to undergraduates about how to interpret a standard confidence interval should appreciate the probabilistic meaning of the Bayesian HPD interval (which is actually how many would *like* to interpret a confidence interval). Conversely, adopting the Bayesian perspective is to say that all claims and summaries will be made on a probabilistic basis, which is much more flexible and convenient.
- *An ability to update these statements as new information is received.* The Bayesian updating process by which today's posteriors become tomorrow's priors is a simple, and fully consistent, way to change conclusions (posteriors) as new information is observed.
- *Systematic incorporation of previous knowledge on the subject.* Prior distributions are not encumbrances; they are opportunities. We all have prior beliefs and suspicions before commencing a data analysis project. One should actually be suspicious about claims of total ignorance, since the substantive interpretation is therefore itself suspect. So the ability to systematically insert information from previous work, competing theories, and qualitative data is an opportunity to ground the model in its substantive context. Furthermore, we are not actually required to provide strongly informed prior information, and weak forms of prior distributions are easy to stipulate.
- *Missing values handled seamlessly as part of the estimation process.* In a Bayesian model everything is designated as either "known" or "unknown" whether this is data, parameters, latent variables, prior parameters, or anything else. So missing data are estimated as a parameter conditional on observed or known quantities just like standard parameters (usually as a nuisance parameter in MCMC estimation). Thus missing data are seamlessly accommodated and do not have to be accommodated as a special problem. In fact, *multiple imputation,* the state-of-the-art standalone

procedure for handling missing data, is a Bayesian estimation process (whether the users of convenient software know it or not).

- *Recognition that population quantities are changing over time rather than fixed immemorial.* One of the defining characteristics of the social sciences is that we rarely have fixed constants to estimate, such as the speed of light, the orbits of planets, or various molecular values (real problems in early natural science statistics). Instead we regularly contend with ideas that are mobile and shifting with changing social systems and uncertain or ambivalent underlying attributes. So if we are to recognize that the probability of war between India and Pakistan changes for both systematic and stochastic reasons, then we should systematize this outcome to the greatest extent possible and summarize the remaining uncertainty probabilistically. In other words, we should create a Bayesian posterior distribution.
- *Direct assessment of both model quality and sensitivity to assumptions.* It is the norm rather than the exception in Bayesian inference to test the sensitivity of the model results to assumptions, particularly the prior. Conversely, it is the exception rather than the norm to do so in classical inference. This partly resulted from defensiveness by twentieth-century Bayesian practitioners, but also because it highlights the degree to which the posterior is a weighted compromise between the prior and the data, something of vital interest.

7 This Issue of *Political Analysis*

This special issues provides a wide variety of well-constructed Bayesian research. The applications include issues of measurement, specification, dimensionality, and estimation. In all cases Bayesian solutions are given to problems where alternatives are either impossible, difficult, or theoretically undesirable.

In "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses," **Kevin Quinn** addresses an underappreciated problem with latent variable analysis: there can be both continuous and ordinal measurements in the same model. In these cases neither factor analysis nor item response theory are fully appropriate, so biases can occur from strict use of either. The Bayesian measurement model presented here provides a major step forward in this nettlesome problem by generalizing measurement to a higher level that specifically incorporates mixed latent traits. The MCMC estimation process developed for this model is made freely available to interested readers through the R package *MCMCpack*.

Bruce Western and **Meredith Kleykamp** provide a Bayesian look at change point analysis in "A Bayesian Change Point Model for Historical Time Series Analysis." Change point estimation is a perfect application for the Bayesian approach because these events in the social sciences do not generally occur instantaneously as they might in some physical experiment. More accurately, their commencement has a probabilistic interpretation due to various delaying and anticipatory effects. So why would we presuppose that a classical binary test of change at some exact point in time would be more appropriate than a posterior distribution for the change? The answer, of course, is that we would not. The authors also show that a simple linear specification can be very flexible when put in the Bayesian context, and the results turn out to be quite interesting.

The "Columbia Group" of **David Park, Andrew Gelman,** and **Joseph Bafumi** present us with a new way to specify multilevel Bayesian models in "Bayesian Multilevel Estimation with Poststratification." The article gives a new means of estimating state-level opinions from national data using a Bayesian model specification. Furthermore, the article takes a novel approach to the problem by constructing a design matrix (in the classical

sense) and using the resulting coefficient estimates to produce predicted probabilities across these characteristics. Finally, the authors show how to construct verifying model checks (a practice that is unfortunately infrequent in standard empirical political science).

Jack Buckley writes a very different type of article from the others in this special issue. In "Simple Bayesian Inference for Qualitative Political Research," he argues for Bayesian inference as a way to "build rapprochement" between qualitative and quantitative political science. Since priors can come from many sources, such information can be completely qualitative in nature as long as the end product for the model is a probability distribution of some kind (very generally defined). The vehicle for illustrating Bayesian flexibility here is the underappreciated Behrens-Fisher problem, which has bedeviled classical statisticians for decades.

Simon Jackman departs from conventional data sources in "What Do We Learn from Graduate Admissions Committees? A Multiple Rater, Latent Variable Model, with Incomplete Discrete and Continuous Indicators" to analyze applications data for a political science Ph.D. program with a measurement model of latent applicant quality based on the ratings from admissions committee members and (the few) observable characteristics of the applicants. Anyone who has served on these committees will certainly appreciate the assessment problem addressed with his technique for combining ordinal rankings with other covariates. This is a surprisingly complex task and one in which the Bayesian approach using data augmentation is quite helpful. It also turns out that this measurement model can be applied to more conventional data-analytic settings in political science where latent traits are important. The findings also imply that such ratings actually do reflect applicant quality, but there exist noticeable biases among individual raters, and the uncertainty of point estimates can be large.

Finally, **Jeff Gill** and **George Casella** address a pervasive issue in the applied MCMC world: mixing problems from high-dimension, multimodal target posterior distributions. Such distributions provide many local maxima to attract and trap Markov chains for extended periods of time that can exceed reasonable chain lengths and thus prevent a full exploration of the posterior distribution of interest. The article, "Dynamic Tempered Transitions for Exploring Multimodal Posterior Distributions," gives a new MCMC algorithm based on simulated annealing. This is a means of melting down the posterior space to allow more free traversal of the Markov chain before cooling back to the original state. The authors then demonstrate the properties and performance of the new technique and apply it to a well-known problem in voting theory.

What is apparent from these brief descriptions is that there is a broad set of applications suited to Bayesian analysis. This underscores my assertion that Bayesian methods are not simply a temporary trend or fashionable application but represent a fundamentally different way of thinking about statistical modeling and inference.

References

Bartels, Larry M. 1996. "Pooling Disparate Observations." *American Journal of Political Science* 40:905–942.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41:641–674.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis,* 2nd ed. New York: Springer-Verlag.

Casella, G., and R. L. Berger. 2001. *Statistical Inference,* 2nd ed. Belmont, CA: Duxbury.

Clarke, Kevin A. 2001. "Testing Nonnested Models of International Relations: Reevaluating Realism." *American Journal of Political Science* 45:724–744.

Cornfield, Jerome. 1969. "The Bayesian Outlook and Its Application." *Biometrics* 25:617–657.

de Finetti, B. 1972. *Probability, Induction, and Statistics.* New York: John Wiley & Sons.

de Finetti, B. 1974. *Theory of Probability, Volume 1.* New York: John Wiley & Sons.

de Finetti, B. 1975. *Theory of Probability, Volume 2.* New York: John Wiley & Sons.
Draper, David. 1995. "Asessment and Propagation of Model Uncertainty." *Journal of the Royal Statistical Society, Series B* 57:45–97.
Ehrlich, I. 1975. "The Deterrent Effect of Capital Punishment: A Question of Life or Death." *American Economic Review* 65:397–417.
Ehrlich, I. 1977. "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence." *Journal of Political Economy* 85:741–788.
Gelfand, A. E., and A. F. M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85:389–409.
Gelman, Andrew, and Gary King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38:514–554.
Gelman, A., and D. B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7:457–511.
Geman, S., and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–741.
Geweke, J. 1992. "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." In *Bayesian Statistics 4,* eds. J. M. Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger. Oxford: Oxford University Press, pp. 169–193.
Geyer, C. J. 1992. "Practical Markov Chain Monte Carlo." *Statistical Science* 7:473–511.
Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52:647–674.
Gill, Jeff. 2001. "Whose Variance Is It Anyway? Interpreting Empirical Models with State-Level Data." *State Politics and Policy Quarterly* 1:313–338.
Gill, Jeff. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach.* New York: Chapman and Hall/CRC.
Good, I. J. 1950. *Probability and the Weighting of Evidence.* London: Griffin.
Hastings, W. K. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika* 57:97–109.
Hill, Jennifer L., and Hanspeter Kriesi. 2001. "Classification by Opinion-Changing Behavior: A Mixture Model Approach." *Political Analysis* 9:301–324.
Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14:382–401.
Jackman, Simon. 2000a. "Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation." *Political Analysis* 8:307–332.
Jackman, Simon. 2000b. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44:375–404.
Jeffreys, H. 1961. *Theory of Probability.* Oxford, UK: Oxford University Press.
Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: John Wiley & Sons.
Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73:31–43.
Lindley, D. V. 1958. "Fiducial Distributions and Bayes' Theory." *Journal of the Royal Statistical Society, Series B* 20:102–107.
Lindley, D. V. 1961. "The Use of Prior Probability Distributions in Statistical Inference and Decision." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, pp. 453–468.
Lindley, D. V. 1965. *Introduction to Probability and Statistics from a Bayesian Viewpoint, Parts 1 and 2.* Cambridge, UK: Cambridge University Press.
Martin, Andrew, and Kevin Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10:134–153.
Quinn, Kevin M., Andrew D. Martin, and Andrew B. Whitford. 1999. "Voter Choice in Multi-party Democracies: A Test of Competing Theories and Models." *American Journal of Political Science* 43: 1231–1247.
Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111–163.
Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. 1997. "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association* 92:179–191.
Robert, C. P., and G. Casella. 2004. *Monte Carlo Statistical Methods,* 2nd ed. New York: Springer-Verlag.
Savage, L. J. 1954. *The Foundations of Statistics.* New York: Wiley.
Savage, L. J. 1962. *The Foundations of Statistical Inference.* London: Methuen.

Smith, Alastair. 1998. "A Summary of Political Selection: The Effect of Strategic Choice on the Escalation of International Crises." *American Journal of Political Science* 42:698–701.

Smith, Alastair. 1999. "Testing Theories of Strategic Choice: The Example of Crisis Escalation." *American Journal of Political Science* 43:1254–1283.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society, Series B* 64:583–639.

Western, B. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42:1233–1259.

Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88:412–423.