BAYESIAN ANALYSIS FOR POLITICAL RESEARCH

Simon Jackman

*Department of Political Science, Stanford University, Stanford,
California 94305-6044; email: jackman@stanford.edu*

Key Words Bayesian statistics, sensitivity analysis, Markov chain Monte Carlo,
roll call analysis, legislative ideal points

■ **Abstract** Bayesian data analysis relies on Bayes' Theorem, using data to update
prior beliefs about parameters. In this review I introduce and contrast Bayesian analy-
sis with conventional frequentist inference and then distinguish two types of Bayesian
analysis in political science. First, Bayesian analysis is used to merge historical infor-
mation with current data in an analysis of likely election outcomes in Florida in 2000;
particular attention is paid to the sensitivity of the results to the choice of prior (i.e., how
much confidence one places in the historical information). Second, a more "modern"
style of Bayesian analysis is reviewed, relying on Markov chain Monte Carlo algo-
rithms to generate computationally intensive "random tours" of the high dimensional
posterior distributions that are the focus of many contemporary Bayesian analyses;
the example used is a central problem in political science, the analysis of legislators'
preferences using roll call data.

1. WHAT IS BAYESIAN ANALYSIS?

Generally, "Bayesian analysis" refers to statistical analysis of data that relies on
Bayes' Theorem, presented below. Bayes' Theorem tells us how to update "prior
beliefs" about parameters or hypotheses in light of data and arrive at "posterior
beliefs." Or, even more simply, Bayes' Theorem tells us how to learn about para-
meters from data, in a way consistent with the laws of probability. As we shall see,
Bayesian analysis is often more easily said than done, or at least this was the case
until recently. In the 1990s there was a veritable explosion of interest in Bayesian
analysis in the statistics profession, which has now crossed over into quantitative
social science. The mathematics and computation underlying Bayesian analysis
have been dramatically simplified via a suite of algorithms known collectively as
Markov chain Monte Carlo (MCMC), to be discussed below. The combination of
MCMC and vast increases in computing power available to most social scientists
means that Bayesian analysis is now truly part of the mainstream of quantitative
social science.

In particular, Bayesian analysis is now feasible and attractive for a set of re-
searchers who are not especially interested in the use of Bayes' Theorem and prior

beliefs in their data analyses but are attracted to the power and ease of MCMC algorithms as a tool for estimation and inference for complex models. In many applications of MCMC, researchers employ "diffuse priors," which (as explained below) effectively mean that the conclusions of the analysis are driven almost entirely by the data, just as they are in traditional likelihood-based analyses. In these cases, MCMC is often adopted as a set of techniques for finding the mode and exploring the shape of a likelihood function. Of course, it is always possible to bring prior information about model parameters to bear via "informative" priors.

Bayes' Theorem is at the heart of Bayesian analysis. Before proceeding to review Bayesian analysis in political science, I review the theorem, introducing some notation and definitions and showing Bayesian analysis at work in some simple applications.

2. PRELIMINARIES: LIKELIHOOD AND BAYES' THEOREM

2.1. Likelihood

The overwhelming bulk of data analysis in the social science uses probability models to relate observed data $\mathbf{y} = (y_1, \ldots, y_n)'$ to unknown parameters θ. A simple example involves modeling independent and identically distributed normal data in terms of its mean μ and its variance σ^2: $y_i \sim N(\mu, \sigma^2)$, where i indexes the observations $1, \ldots, n$. The familiar linear regression model follows by replacing the single parameter μ with $\mathbf{x}_i \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of unknown parameters. Generically, we can write these probability models as $\mathbf{y} \sim f(\mathbf{y} \mid \theta)$.

The "likelihood function" summarizes the information about θ in $\mathbf{y}$, defined as any function of θ proportional to $f(\mathbf{y} \mid \theta)$ (e.g., Tanner 1996, p. 14):

$$\mathcal{L}(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta). \qquad 1.$$

Both the frequentist and Bayesian approaches to statistical inference exploit the likelihood function. Frequentist inference treats θ as fixed but unknown, and sample-based estimates of θ, $\hat{\theta}$, as random (since repeated sampling, if undertaken, could yield different values of $\hat{\theta}$). The most widely used estimate of θ is the "maximum likelihood estimate," that value of θ that maximizes Equation 1. Frequentists use the likelihood function to evaluate the plausibility of the other $\hat{\theta}$ that might hypothetically result from repeated sampling, relative to the observed sample estimate $\hat{\theta}$. Neyman-Pearson–type inferential procedures, such as likelihood ratio tests, follow fairly straightforwardly from this perspective (e.g., Edwards 1992; Bernardo & Smith 1994, pp. 450–55). This approach to statistical inference has been championed within political science by King (1989).

Bayesian inference takes $\hat{\theta}$ as fixed (a feature of the observed data $\mathbf{y}$) and θ as random (in the sense that the researcher is uncertain of its true value). Bayesians use probability—the formal quantification of uncertainty—to characterize this uncertainty. Bayesian analysis produces posterior probability statements about

θ ("posterior" literally in the sense of "after" observing the data). The likelihood function summarizes the sample information about θ and thus provides an essential ingredient in Bayesian statistics, as we shall now see.

2.2. Bayes' Theorem

Bayes' Theorem is also frequently referred to as Bayes' Rule and is one of the most important results in all of statistics. In its simplest form, for discrete events, if H is a hypothesis and E is evidence (i.e., data), then Bayes' Theorem is

$$\Pr(H \mid E) = \frac{\Pr(E \cap H)}{\Pr(E)} = \frac{\Pr(E \mid H)\Pr(H)}{\Pr(E)},$$

provided $\Pr(E) > 0$, so that $\Pr(H \mid E)$ is the probability of belief in H after obtaining E, and $\Pr(H)$ is the prior probability of H before considering E. The left-hand side of the theorem, $\Pr(H \mid E)$, is usually referred to as the posterior probability of H. The theorem thus supplies a solution to the general problem of inference or induction (e.g., Hacking 2001), providing a mechanism for learning about a hypothesis H from data E.

Bayes' Theorem itself is uncontroversial. It is merely an accounting identity that follows from the axiomatic foundations of probability that link joint, conditional, and marginal probabilities, i.e., $\Pr(A \mid B)\Pr(B) = \Pr(A \cap B)$, where $\Pr(B) \neq 0$. Thus, Bayes' Theorem is sometimes referred to as the rule of inverse probability, since it shows how a conditional probability B given A can be "inverted" to yield the conditional probability A given B (Leamer 1978, p. 39).

In most analyses in the social sciences, we want to learn about a continuous parameter rather than the discrete parameters considered in the discussion thus far. Examples include the mean of a continuous variable in some population, or a proportion (a continuous parameter on the unit interval), or a regression coefficient. As above, we refer to a generic parameter as θ and denote the data available for analysis as $\mathbf{y} = (y_1, \ldots, y_n)'$. In this case, beliefs over the parameter are represented as probability density functions. Generically, we denote the prior as $\pi(\theta)$ and the posterior as $\pi(\theta \mid \mathbf{y})$. Bayes' Theorem for a continuous parameter is

$$\pi(\theta \mid \mathbf{y}) = \frac{\pi(\mathbf{y}; \theta)\pi(\theta)}{\int \pi(\mathbf{y}; \theta)\pi(\theta)\, d\theta}$$

but is more commonly expressed as

$$\pi(\theta \mid \mathbf{y}) \propto \pi(\mathbf{y}; \theta)\pi(\theta), \qquad\qquad 2.$$

where $\pi(\mathbf{y}; \theta)$ is the likelihood function (i.e., the likelihood function is simply the probability of the data given the parameter). In words, we can state this version of Bayes' Theorem as "the posterior is proportional to the prior times the likelihood." This well-known "Bayesian mantra" highlights a particularly elegant feature of the Bayesian approach: The likelihood function can be turned into a probability statement about θ, given data $\mathbf{y}$.

Several interesting features of Bayesian analysis are immediately apparent via inspection of Bayes' Theorem:

1. The Bayesian approach treats the parameter θ as a random variable and makes inferences conditional on the data, whereas frequentist approaches consider θ a fixed (but unknown) property of a population from which we randomly sample data **y**.

2. If the prior for θ is uniform [i.e., $\pi(\theta) = c > 0$, or any value of θ is as likely as any other, a priori], then the posterior density is simply proportional to the likelihood, since the uniform prior is absorbed into the constant of proportionality in Equation 2. This means that the posterior density has the same shape as the likelihood function. In turn, this means that when prior beliefs about parameters are uniform, reflecting "prior ignorance" about θ, then the results of a Bayesian analysis and a likelihood-based analysis will coincide. The maximum likelihood estimate (that value of θ where the likelihood function is maximized) corresponds to the location of the mode of the posterior density, and if the posterior is symmetric, the maximum likelihood estimate will also correspond to the location of the posterior mean. Put differently, from a Bayesian perspective, likelihood-based analyses of data assume prior ignorance, although seldom is this assumption made explicit, even if it is plausible.

3. If the prior density assigns zero weight to specific ranges of θ, then those ranges have zero posterior probability. Priors that have this property have the effect of truncating the range of feasible estimates for θ.

2.3. Historical Note

Bayes' Theorem is named for the Reverend Thomas Bayes, who died in 1761. Bayes' Theorem first appeared in an essay attributed to Bayes (1763) and communicated to the Royal Society after Bayes' death by Richard Price in 1763, which has been republished many times (e.g., Bayes 1958). Bayes himself only stated the result for a uniform prior. According to Stigler (1986b), in 1774 Laplace, apparently unaware of Bayes' work, stated the theorem in its more general form (for discrete events). Additional historical detail can be found elsewhere (Bernardo & Smith 1994, ch. 1; Lindley 2001; Stigler 1986a, ch. 3).

3. EXAMPLE 1: POLLING RESULTS

In early March of 2000, Mason-Dixon Polling and Research conducted a poll of voting intentions in Florida for the November U.S. presidential election. The poll considered Bush and Gore the presumptive nominees of their respective political parties. The poll had a sample size of 621 and resulted in the following breakdown of reported vote intentions: Bush 45%, Gore 37%, Buchanan 3%, and undecided

15%. For simplicity, we ignore the undecided and Buchanan vote share, leaving Bush with 55% of the two-party vote intentions and Gore with 45%, and $n = 509$ respondents expressing a preference for one of the major-party candidates. Although the data are binomial, with this relatively large n, a normal distribution provides an excellent approximation to the (frequentist) sampling distribution of the Bush and Gore vote shares, with standard error $\sqrt{(0.55 \times 0.45)/509} = 0.022$; in turn, this implies that a 95% confidence interval on Bush vote share extends from 50.7% to 59.3%. Put differently, this early poll unambiguously points to Bush leading Gore in Florida in the 2000 presidential race.

But how realistic is this early poll result? Is there other information available that bears on the election result? Previous presidential elections are an obvious source of information, which I exploit by using a simple regression model. I model all state-level presidential election outcomes from 1932 to 1996 with fixed effects (dummy variables) for state and year; as in the analysis of the survey above, the dependent variable is the Republican proportion of the two-party vote. I then use this regression model to make a prediction for Florida in 2000, picking up the fixed effect for Florida and selecting the median fixed effect for year (i.e., assuming that 2000 would be a typical election in terms of the national swing toward either major party). This prediction is 49.1%, with a standard error of 2.2 percentage points.

We can combine the information yielded by this analysis of previous elections with the survey via Bayes' Theorem. We can consider the prediction from the regression analysis as supplying a prior and consider the survey as data—but mathematically, it does not matter what label we assign each piece of information. As we shall now see, applying Bayes' Theorem is identical to pooling information about an underlying parameter.

To apply Bayes' Theorem in this case, I assume that the information supplied by the regression analysis can be characterized as a normal distribution, with mean equal to the point prediction (49.1%) and standard deviation equal to the standard error of the point prediction (2.2 percentage points). This is the prior density for this problem. Likewise, the information in the survey is summarized with a normal distribution, with mean 55% and standard deviation of 2.2 percentage points. Thus, both the prior and the likelihood for this problem are normal distributions, meaning that the posterior density will also be normal; in general, when the posterior density is of the same functional form as the prior density, then the prior is said to be "conjugate" to the likelihood (and indeed, priors are often chosen so as to have this property).

Generally, given a normal prior $\pi(\theta) \equiv N(\mu_0, \sigma_0^2)$ and a normal likelihood $\pi(y; \theta) = N(\mu_1, \sigma_1^2)$, with σ_1^2 known the posterior density is $\pi(\theta \mid y) = N(\mu_2, \sigma_2^2)$, where $\mu_2 = (\mu_0/\sigma_0^2 + \mu_1/\sigma_1^2)/(1/\sigma_0^2 + 1/\sigma_1^2)$; and $\sigma_2^2 = (1/\sigma_0^2 + 1/\sigma_1^2)^{-1}$. [For a proof, see any introductory text on Bayesian statistics, such as Lee (1989) or Gelman et al. (1995).] The mathematics of this conjugate normal/normal problem has a simple interpretation: The posterior mean is a weighted average of the prior mean and the maximum likelihood estimate, with weights equal to the precision of each, where the precisions of each distribution are the inverses of the respective

variances. Thus, the posterior is a compromise between prior and data, where the precisions (inverse variances) tell us how to weight each.

Applying this result to the problem at hand, we have $\mu_0 = 0.491$, $\mu_1 = 0.55$, and $\sigma_0 = \sigma_1 = 0.022^2$, and so the posterior over Bush's two-party vote share in Florida is a normal distribution with mean 52.1% and standard deviation 1.6 percentage points. This means that the 95% (posterior) confidence interval ranges from 48.9% to 55.2%, with 90% of the posterior probability density lying above 50% (i.e., the probability that Bush defeats Gore is 90%). Combining the survey result with the prior still leaves us reasonably confident that Bush would beat Gore, but the evidence is now more mixed and perhaps more plausible. Figure 1 provides a graphical display of the Bayesian analysis. The posterior mean lies between the prior and the survey result; it is also obvious that the posterior density has less dispersion (more precision) than the prior or the survey result alone.

3.1. Sensitivity Analysis

Of course, reasonable people could hold differing prior beliefs as to Bush's vote support. For instance, someone might dispute the relevance of the historical election data. Even if one considered the historical election data relevant, one might dispute the idea that, a priori, 2000 would be on a par with the "median year" for the Democratic presidential candidate. Thus, in many Bayesian analyses, it is useful to perform a sensitivity analysis, examining how the posterior density changes
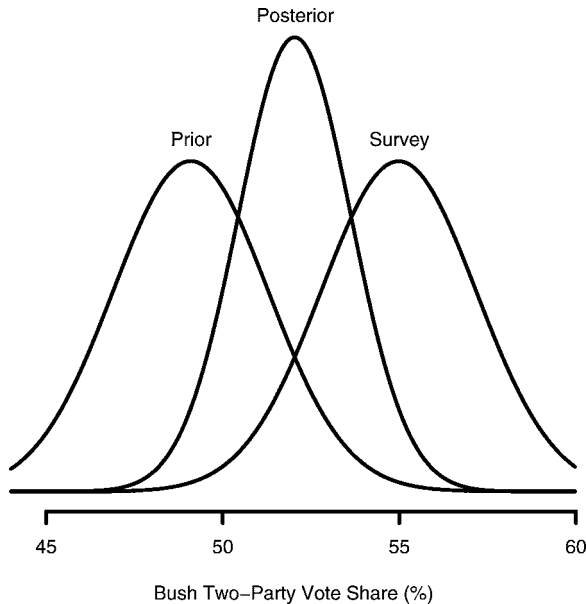

Figure 1 Example 1, Bayesian analysis of Bush support in Florida, 2000 election.

as one adopts different priors, so as to reassure oneself that one's own conclusions would also be reached by someone else with similar (albeit not identical) priors.

There are two primary ways of performing sensitivity analyses in practice. First, we can arbitrarily weaken or strengthen the prior adopted above by multiplying the prior variance by a scale parameter $\lambda \geq 0$. Setting $\lambda = 0$ generates a degenerate "spike prior" with zero variance and hence infinite precision, a prior from which no amount of data can move us. For $\lambda > 1$ we obtain a less precise version of the prior, with successively larger values of λ generating successively weaker or "uninformative" priors; the resulting posterior densities are successively dominated by the data. As $\lambda \rightarrow \infty$, the data completely dominate the prior, the posterior has the same shape as the likelihood (i.e., the prior is tending toward a "locally uniform prior," in the sense of being flat over the region of the parameter space supporting the likelihood), and the Bayesian analysis yields the same conclusions as analysis based on the likelihood.

The other approach to sensitivity analysis does not simply weaken or strengthen the analyst's prior but instead repeats the Bayesian analysis with an array of more or less plausible different prior densities. We can then identify the set of priors that lead to qualitatively similar conclusions and note the size of this set. One might use this type of sensitivity analysis to ask whether there is any substantively plausible prior that leads to other conclusions than those obtained with one's own prior.

For the current example, the first type of sensitivity analysis yields the posterior densities summarized in Figure 2. In the left panel, the thick, S-shaped solid line represents the location of the posterior mean (vertical axis) as a function of the sensitivity parameter λ, and the thinner solid lines show the 2.5% and 97.5% quantiles of the posterior. As $\lambda \rightarrow 0$, the posterior tends toward a degenerate
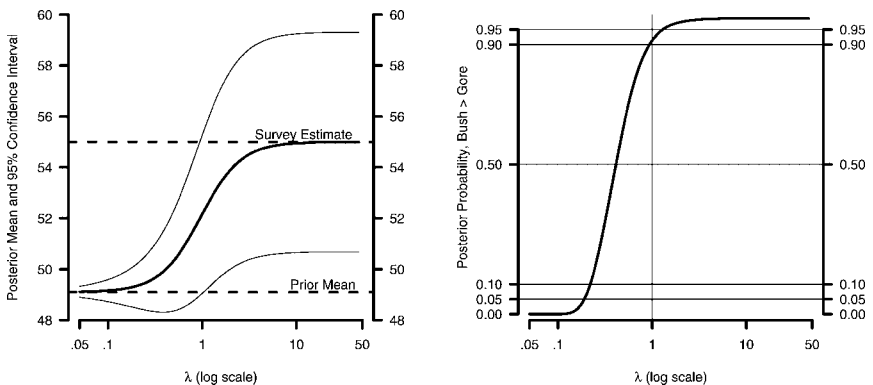

Figure 2 Example 1, sensitivity analysis. Left: The thick solid line shows the posterior mean as a function of the sensitivity parameter, λ, and the thinner solid lines show the 2.5% and 97.5% quantiles of the posterior. Right: The posterior probability that Bush's vote share exceeds Gore's, as a function of λ.

distribution with point mass on the prior mean of 49.1%. On the other hand, as $\lambda \to \infty$, the posterior becomes the normal distribution implied by the data. The right panel of Figure 2 shows the posterior probability that Bush's vote share is greater than Gore's as a function of λ. With $\lambda > 1$, this probability is greater than 0.90, and only with a substantially stronger version of the prior (say, $\lambda < 0.5$) does the probability that Bush's vote share exceeds Gore's drop below 0.5. On balance, we see that the finding that Bush's support exceeds Gore's is robust over a wide range of priors. We would need a prior twice as precise as the one we specified to find the balance of evidence pointing the other way.

Figure 3 shows another type of sensitivity analysis. The mapping from a set of conjugate normal priors to posteriors is presented as a contour plot, with the contours indicating the posterior probability that Bush's vote share exceeds Gore's. Each prior is defined by a mean (horizontal axis) and standard deviation (vertical axis), with the priors becoming less precise as we move up the graph. The graph
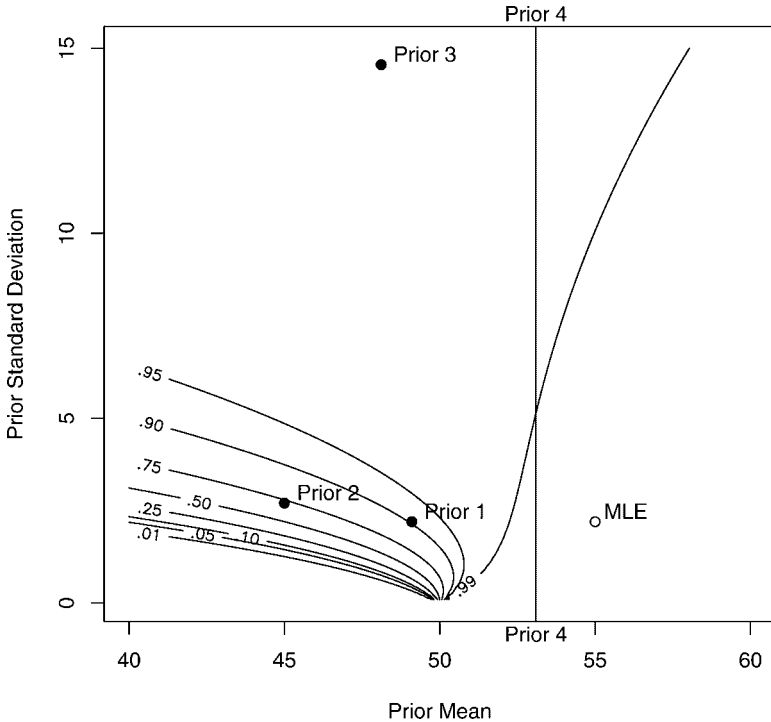

Figure 3 Example 1, sensitivity analysis via a "Bayesian map." This contour graph shows the mapping from (conjugate) normal priors to the posterior probability that Bush's vote share exceeds Gore's. Each point on the graph defines a normal prior with a mean (horizontal axis) and standard deviation (vertical axis). The contour lines indicate points in the prior space with the indicated posterior probability of Bush's two-party vote share exceeding Gore's. MLE, maximum likelihood estimate.

shows three types of prior. Prior 1 is the prior used above, suggested by the analysis of state-level election returns (mean $=$ 49.1% and standard deviation $=2.2$ percentage points); this prior does not strongly contradict the survey data, and the resulting posterior probability that Bush's vote share exceeds Gore's is 0.90. The sensitivity analysis in the previous section corresponds to a vertical line through Prior 1. Prior 2 is generated similarly to Prior 1 (regression analysis of state-level presidential election results with fixed effects for state), but instead of fixed effects for year, it includes a dummy variable for whether Democrats held the White House. This yields a prior more favorable to Gore (45% Republican vote share), but the posterior probability that Bush's vote share exceeds Gore's remains above 0.5. Prior 3 results from taking the average (and standard deviation) of Florida presidential elections, 1932–1996; there is considerable variation in these election results, giving rise to the large standard deviation, and the posterior is largely shaped by the survey data. Prior 4 is actually not one single prior but a set of priors generated by taking the average of the Republican share of the two-party vote in the three previous presidential elections in Florida (1988, 1992, and 1996), which is 53.1%, and considering a range of prior standard deviations (corresponding to different degrees of confidence in the prior mean). Three of the four priors largely agree with the survey data, leading to the conclusion that Bush's vote share exceeds Gore's; Prior 2 is a stronger "Democratic" prior and results in an inconclusive posterior, although the balance of the posterior uncertainty lies in a pro-Bush direction.

Finally, the map in Figure 3 shows just how strong one's priors would have to be so as to receive the "message" of the survey data but still conclude that on balance, Gore's support exceeded Bush's. The corresponding region of the "conjugate prior space" lies below the 0.50 probability contour. For example, a prior mean of 44% and a prior standard deviation of one percentage point would yield a posterior favorable to Gore, although the historical data do not readily admit such a prior.

4. PROPERTIES OF CONJUGATE ANALYSIS

Some other properties of conjugate Bayesian analysis should also be noted. For conjugate problems with normal distributions with known variances, the posterior precision is (weakly) greater than either the prior precision or the precision of the data, and indeed equals the sum of these two precisions. Put differently, we never lose precision by combining the information about θ in data with the information about θ in a conjugate prior. Further, with enough data, say from repeated applications of the data generation process, the posterior precision will continue to increase and will eventually overwhelm any nondegenerate prior. The upshot is that analysts with different (nondegenerate) prior beliefs over θ will eventually agree about θ provided they (*a*) see enough data and (*b*) update their beliefs using Bayes' Theorem. Because of these properties, Bayesian analysis has been heralded as a model for scientific practice: Reasonable people may differ (at least prior to

seeing data) but our views will tend to converge as scientific knowledge accumulates, provided we update our views consistent with the laws of probability (i.e., via Bayes' Theorem).

5. BAYESIAN COMPUTATION VIA MARKOV CHAIN MONTE CARLO ALGORITHMS

I now turn to consider what is sometimes called "modern" Bayesian analysis: the exploration of posterior densities via computationally intensive Markov chain Monte Carlo (MCMC) algorithms. The use of these algorithms has made Bayesian analysis feasible for a wide class of models and priors.

5.1. The Goal of Bayesian Computation

As shown above, in practice, Bayesian analysis amounts to computing the posterior density for the parameters in the model being analyzed. For simple models, it is possible to use analytic results to characterize the posterior density; recall the example in Section 3 where the posterior density for Bush's vote share turned out to be that of a normal distribution, with its mean and variance easily computed. Other features of a normal posterior density are also immediately accessible. For instance, 95% confidence intervals are trivially computed as the posterior mean plus or minus 1.96 times the standard deviation (assuming the variance to be known). Other confidence intervals or a critical quantile can be computed just as easily.

But for more interesting statistical models, the resulting posterior densities are not so straightforward to compute. Indeed, it is somewhat surprising how difficult Bayesian analysis can become once we leave behind "toy" problems. Even for models as familiar as probit or logit for binary responses, Bayesian analysis can become mathematically demanding and require considerable skill to work through the integrals inherent in Bayes' Theorem or to obtain summaries of the posterior density such as the posterior mean. These problems often become more challenging as the number of parameters in the model increases, since the posterior density becomes a multivariate density. Moreover, there is no guarantee that the posterior density can be neatly summarized by reporting its mean and standard deviation; the posterior density might be asymmetric or multimodal, in which case we might require a graphical summary (a histogram or density plot) to understand what the data are telling us about θ.

5.2. The Monte Carlo Principle

Fortunately, there exists an easily implemented solution to the general problem of computing posterior densities. The solution comes in two parts. The first part relies on the Monte Carlo principle:

> Anything we want to know about a random variable θ, we can learn by repeated sampling from the probability density function of θ.

Moreover, the precision with which we learn about features of θ is limited only by the number of random samples we are willing to wait for our computer to draw for us. The Monte Carlo method has long been known to statisticians (e.g., Metropolis & Ulam 1949), but only recently have increases in computing power made this principle useful to quantitative social scientists.

This simulation-based approach greatly simplifies Bayesian analysis. To learn about θ we simply tell a computer to sample many times from the posterior density for θ. To communicate what we have learned about θ from the data, we can summarize those samples in a histogram or via some numerical summary. Traditional hypothesis testing amounts to simply noting how many of the sampled values of θ lie above or below zero, or any other threshold of interest. The second "MC" in "MCMC" stands for the use of the Monte Carlo principle. The first "MC" stands for "Markov chain," which we consider next.

5.3. Using Markov Chains to Explore Multidimensional Parameter Spaces

Consider the generic problem of using data $\mathbf{y}$ to learn about d parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)'$. The posterior density for this problem, $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, is d-dimensional. Via the Monte Carlo principle, we know that we can learn about the posterior density by sampling from it many times. But the multivariate posterior density may not have a convenient, analytical form corresponding to one of the probability distributions from which our computer knows how to sample.

Happily, an algorithm known as the Gibbs sampler makes it easy to sample $\boldsymbol{\theta}$ from its (multivariate) posterior density, $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, by generating a random tour of the parameter space that supports the posterior. If the Gibbs sampler is at some arbitary point $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_d^{(t)})$ at iteration t, then the (random) transition to $\boldsymbol{\theta}^{(t+1)}$ occurs via the following scheme:

1. Sample $\theta_1^{(t+1)}$ from $p(\theta_1 | \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_d^{(t)}, \mathbf{Y})$.
2. Sample $\theta_2^{(t+1)}$ from $p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_d^{(t)}, \mathbf{Y})$.

$\vdots \qquad \vdots$

d. Sample $\theta_d^{(t+1)}$ from $p(\theta_d | \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{d-1}^{(t+1)}, \mathbf{Y})$.

A useful way to think about what the Gibbs sampler does is to see that the full joint posterior density for $\boldsymbol{\theta}$ has been broken down into a series of lower-dimensional conditional densities, circumventing the "curse of [high] dimensionality" (Gelfand 1997, p. 284). In turn this is driven by the fact (well known to Bayesians) that conditional densities determine marginal densities (Casella & George 1992, pp. 170–71).

The sequence of sampled vectors produced by this scheme, $\langle \boldsymbol{\theta}^{(t)} \rangle = \{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t)}, \ldots\}$, form a Markov chain. More specifically, under a large set of

conditions, the sampled vector $\boldsymbol{\theta}^{(t)}$ is the state vector of a convergent Markov chain that has the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$ as its "invariant," "stationary," or "limiting" distribution.

"Very minimal conditions turn out to be sufficient and essentially necessary to ensure convergence of the distribution of the [MCMC] sampler's state to the invariant distribution and to provide a law of large numbers for sample path averages" (Tierney 1996, p. 59). It is not possible to summarize these conditions in the space available here. A key condition for the existence of an invariant distribution for a Markov chain over a continuous state-space (a parameter space, in the context of MCMC) is irreducibility, which (informally) means that "the chain must be able to reach all interesting parts of the state-space" (Tierney 1996, p. 62). That is, if regions of the parameter space with positive posterior probability are noncontiguous, the Markov chain must be able to "jump" the zero-probability regions in a finite number of transitions; otherwise, the Markov chain is exploring only a subset of the feasible parameter space, yielding a misleading characterization of the posterior density. In most statistical applications this condition holds, but interesting counterexamples can be easily constructed (e.g., Gamerman 1997, p. 124).

If these conditions are met, then the output of the Gibbs sampler, $\boldsymbol{\theta}^{(t)}$, converges in distribution to the target posterior density as $t \rightarrow \infty$. More simply, when the Markov chain has been run for a sufficient "burn-in" period, each subsequent realization of the state vector is a sample from this posterior distribution. These samples from the posterior distribution are stored and summarized for inference. Any other relevant quantities that are functions of $\boldsymbol{\theta}$ can also be calculated with each Gibbs sample, once the Markov chain reaches its invariant distribution. Examples include the proportion of sampled θ that lie above or below zero, the observed data log-likelihood, residuals in a regression setting, or the percentage of cases correctly classified in a qualitative dependent-variable context.

5.4. Generalizations: Metropolis-Hastings

The Gibbs sampler is actually a special case of a more general random-tour algorithm known as the Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970), which I briefly describe here; Chib & Greenberg (1995) provide a useful explanation of the Metropolis-Hastings algorithm and practical tips for its implementation. The Metropolis-Hastings algorithm defines a set of "jumping rules" that govern how the algorithm randomly traverses the parameter space. At the start of iteration t, we have $\boldsymbol{\theta}^{(t-1)}$ and we make the transition to $\boldsymbol{\theta}^{(t)}$ as follows (Gelman et al. 1995, pp. 324–26):

1. Sample $\boldsymbol{\theta}^*$ from a "candidate," "proposal," or "jumping" distribution $J_t(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$.

2. Calculate the ratio

$$ r = \frac{p(\boldsymbol{\theta}^* \mid y)/J_t(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)} \mid y)/J_t(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)}, $$

which reflects the plausibility of the candidate point θ^* relative to the current value $\theta^{(t-1)}$.

3. Set

$$\theta^{(t)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise.} \end{cases}$$

This scheme means that if the candidate point increases the posterior density, it is accepted with probability 1; if the candidate point does not increase the posterior density, it is accepted with probability r. It can be shown that this scheme generates a Markov chain that has the posterior density $p(\theta|y)$ as its invariant distribution. The power of the Metropolis-Hastings method stems from the fact that no matter what the form of the proposal distribution, the invariant distribution of the resulting Markov chain will still be the desired posterior distribution, $p(\theta|y)$; for proofs, see Gilks et al. (1996) and references cited therein. Gibbs sampling is a special case of the Metropolis-Hastings algorithm in the sense that each component of θ is updated sequentially and the implicit jumping distributions are simply the conditional densities $p(\theta_j|\theta_{-j}^{(t-1)}, y)$; this means that $r = 1$ and each candidate point is always accepted.

6. EXAMPLE 2: THE ANALYSIS OF ROLL CALL DATA

I illustrate MCMC algorithms with an empirical problem of wide interest in political science: the analysis of roll call data. The Bayesian analysis of roll call data is described in detail elsewhere (Jackman 2000, 2001; Clinton et al. 2004); here I provide an abbreviated discussion of the model and the statistical issues.

Roll call data are generated by n legislators voting on m different roll calls. Each roll call $j = 1, \ldots, m$ presents legislators $i = 1, \ldots, n$ with a choice between a "Yea" position ζ_j and a "Nay" position Ψ_j, locations in $\mathbb{R}^d$, where d denotes the dimension of the policy space. Let $y_{ij} = 1$ if legislator i votes "Yea" on the jth roll call and $y_{ij} = 0$ otherwise. Political scientists typically analyze roll call data using a Euclidean spatial voting model (Enelow & Hinich 1984). Legislators are assumed to have quadratic utility functions over the policy space, $U_i(\zeta_j) = -\|\mathbf{x}_i - \zeta_j\|^2 + \eta_{ij}$ and $U_i(\Psi_j) = -\|\mathbf{x}_i - \Psi_j\|^2 + v_{ij}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the "ideal point" of legislator i, and η_{ij} and v_{ij} are the errors or stochastic elements of utility, and $\|\cdot\|$ is the Euclidean norm. Utility maximization implies

$$y_{ij} = \begin{cases} 1 & \text{if } U_i(\zeta_j) > U_i(\Psi_j), \\ 0 & \text{otherwise.} \end{cases} \qquad 3.$$

The specification is completed by assigning a distribution to the errors. We assume that the errors η_{ij} and v_{ij} have a joint normal distribution with $E(\eta_{ij}) = E(v_{ij})$, $\text{var}(\eta_{ij} - v_{ij}) = \sigma_j^2$ and that the errors are independent across both legislators and

roll calls. It follows that

$$
\begin{aligned}
P(y_{ij} = 1) &= P\left[U_i(\boldsymbol{\zeta}_j) > U_i(\boldsymbol{\Psi}_j)\right] \\
&= P\left(\nu_{ij} - \eta_{ij} < \|\mathbf{x}_i - \boldsymbol{\Psi}_j\|^2 - \|\mathbf{x}_i - \boldsymbol{\zeta}_j\|^2\right) \\
&= P(\nu_{ij} - \eta_{ij} < 2(\boldsymbol{\zeta}_j - \boldsymbol{\Psi}_j)'\mathbf{x}_i + \boldsymbol{\Psi}_j'\boldsymbol{\Psi}_j - \boldsymbol{\zeta}_j'\boldsymbol{\zeta}_j) \\
&= \Phi(\boldsymbol{\beta}_j'\mathbf{x}_i - \alpha_j), \qquad\qquad\qquad 4.
\end{aligned}
$$

where $\boldsymbol{\beta}_j = 2(\boldsymbol{\zeta}_j - \boldsymbol{\Psi}_j)/\sigma_j$, $\alpha_j = (\boldsymbol{\zeta}_j'\boldsymbol{\zeta}_j - \boldsymbol{\Psi}_j'\boldsymbol{\Psi}_j)/\sigma_j$, and $\Phi(\cdot)$ denotes the standard normal distribution function. This corresponds to a probit model (a logit model results if the errors have extreme value distributions), with an unobserved regressor $\mathbf{x}_i$ corresponding to the legislator's ideal point. The coefficient vector $\boldsymbol{\beta}_j$ is the direction of the jth proposal in the policy space relative to the "Nay" position.

Given the assumptions of independence across legislators and roll calls, the likelihood is

$$
L(\mathbf{B}, \boldsymbol{\alpha}, \mathbf{X} \mid \mathbf{Y}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \Phi(\mathbf{x}_i'\boldsymbol{\beta}_j - \alpha_j)^{y_{ij}} \left[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}_j - \alpha_j)\right]^{1-y_{ij}}, \qquad 5.
$$

where $\mathbf{B}$ is an $m \times d$ matrix with jth row $\boldsymbol{\beta}_j'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$, $\mathbf{X}$ is an $n \times d$ matrix with ith row $\mathbf{x}_i'$, and $\mathbf{Y}$ is the $n \times m$ matrix of observed votes with (i, j)th element y_{ij}. Without restrictions on the parameters, the model parameters are not identified, since any rescaling and/or rotation of $\mathbf{X}$ is indistinguishable from an offsetting transformation of the $\mathbf{B}$ parameters. A simple solution when fitting unidimensional models is to impose the constraint that the ideal points have mean zero and unit variance; I adopt this identifying constraint in the analysis below. Note that, in a Bayesian analysis, we can always ignore the lack of identification. The lack of identification is a feature of the likelihood and not necessarily a feature of the posterior density (because when the likelihood is flat with respect to the parameters, the posterior will coincide with the prior). Nonetheless, here I impose the identifying restriction.

The unidimensional spatial voting model is equivalent to the two-parameter item-response model used in educational testing, where β_j is the item-discrimination parameter and α_j is the item-difficulty parameter. But in the roll call context, the latent trait or "ability" parameter x_i is the ideal point of the ith legislator. Albert (1992), Patz & Junker (1999), and Johnson & Albert (1999) show how the Gibbs sampler can be used for a Bayesian analysis of this model, which I briefly summarize.

In a Bayesian setting, data analysis involves computing the joint posterior density of all model parameters; we denote this joint posterior density as $\pi(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\alpha}|\mathbf{Y})$. In most applications, this density is extremely high dimension; with n legislators, m roll calls, and a d-dimensional spatial voting model, $\mathbf{X}$ is an n-by-d matrix of unknown parameters, $\boldsymbol{\beta}$ is an m-by-d matrix, and $\boldsymbol{\alpha}$ is a vector of length m, for a

TABLE 1 Number of parameters in roll call analyses

	Legislators	Roll calls	Dimensions (*d*)		
	n	*m*	1	2	3
U.S. Supreme Court, 1994–1997	9	213	435	657	879
105th U.S. Senate	100	534	1168	1802	2436
93rd U.S. House	442	917	2276	3635	4994
U.S. Senate, 1789–1985	1714	37,281	76,276	115,271	154,266
U.S. House, 1789–1985	9759	32,953	75,485	118,017	160,549

total of $nd + m(d+1)$ parameters. Table 1 presents values of p for five different data sets. A moderately sized roll call data set (say the 105th U.S. Senate) with $n = 100$, $m = 534$ non-unanimous roll calls, and $d = 1$ yields $p = 1168$ unknown parameters, whereas a two-dimensional model yields $p = 1802$ parameters. A typical House of Representatives (e.g., the 93rd House) set has $n = 442$ and $m = 917$, and so a one-dimensional model has $p = 2276$ parameters whereas a two-dimensional model has $p = 3635$ parameters. Pooling across years dramatically increases the number of parameters. For instance, Poole & Rosenthal (1997) report that fitting a two-dimensional model to roughly 200 years of U.S. House of Representatives roll call data gives rise to an optimization problem with $p > 150,000$ parameters.

This proliferation of parameters causes several problems. The usual optimality properties of conventional estimators, such as maximum likelihood, may not hold when, as in this case, the number of parameters is a function of the sample size (see Lancaster 2000 for a recent survey). In particular, the customary asymptotic standard error calculations, using the inverse of the information matrix, are not valid. As a practical matter, the information matrix is too large for direct inversion. Poole & Rosenthal (1997, p. 246) take the obvious shortcut of fixing the bill parameters at their estimated values before calculating standard errors for the ideal point estimates. They acknowledge that this is invalid, but it reduces the computational burden by an order of magnitude.

The Bayesian methods of estimation and inference proposed here are valid for finite samples and do not employ any large-sample approximations. The number of parameters is fixed for any particular estimation problem by the actual number of legislators and roll calls, and Bayes' Theorem gives the exact posterior distribution of the parameters conditional on the observed data. The only approximation that is required involves the simulation of the posterior distribution, and this approximation can be made to any desired degree of accuracy by increasing the number of simulations (not the sample size). Thus, the Bayesian approach offers both practical and theoretical improvements over traditional approaches based on maximum likelihood.

6.1. The Gibbs Sampler for the Analysis of Roll Call Data

The idea that drives the Gibbs sampler for this problem is to recognize that any probit model can be rewritten as latent linear regression (e.g., Albert & Chib 1993). In the context of roll call analysis, this latent linear regression takes the form $y_{ij}^* = \mathbf{x}_i \boldsymbol{\beta}_j - \alpha_j + \varepsilon_{ij}$, where $\varepsilon_{ij} \overset{iid}{\sim} N(0,1)$ gives us a probit model and we observe a "Yea" vote ($y_{ij} = 1$) if $y_{ij}^* \geq 0$ and a "Nay" vote ($y_{ij} = 0$) if $y_{ij}^* < 0$. If we also specify normal priors on the unknown $\mathbf{x}_i$, $\boldsymbol{\beta}_j$, and $\boldsymbol{\alpha}_j$, we have a conjugate regression Bayesian model for those parameters, where the latent y_{ij}^* is the dependent variable(s). The only issue is how to generate y_{ij}^*. This is rather straightforward; we simply add the y_{ij}^* to the parameter vector carried around by the Gibbs sampler. Thus, the Gibbs sampler for this problem consists of the sampling from the following conditional distributions, where t indexes iterations:

1. $g(y_{ij}^* \mid y_{ij}, \mathbf{x}_i^*, \boldsymbol{\beta}_j \alpha_j)$. At the start of iteration t, we have $\boldsymbol{\beta}_j^{(t-1)}$, $\alpha_j^{(t-1)}$, and $\mathbf{x}_i^{(t-1)}$. We sample $y_{ij}^{*(t)}$ from one of the two following densities, depending on whether we observed a "Yea" ($y_{ij} = 1$) or a "Nay" ($y_{ij} = 0$):

$$y_{ij}^* \mid \left(y_{ij} = 0, x_i^{(t-1)}, \beta_j^{(t-1)}, \alpha_j^{(t-1)}\right) \sim N\left(\mu_{ij}^{(t-1)}, 1\right) I(y_{ij}^* < 0)$$

   (i.e., truncated normal)   or

$$y_{ij}^* \mid \left(y_{ij} = 1, x_i^{(t-1)}, \beta_j^{(t-1)}, \alpha_j^{(t-1)}\right) \sim N\left(\mu_{ij}^{(t-1)}, 1\right) I(y_{ij}^* \geq 0)$$

   (i.e., truncated normal),

   where $\mu_{ij}^{(t-1)} = \mathbf{x}_i^{(t-1)} \beta_j^{(t-1)} - \alpha_j^{(t-1)}$ and $I(\cdot)$ is an indicator function. For abstentions and other missing roll calls, we sample $y_{ij}^{*(t)}$ from the untruncated normal density $N(\mu_{ij}^{(t-1)}, 1)$, effectively generating multiple imputations for these missing data over iterations of the MCMC algorithm.

2. $g(\boldsymbol{\beta}_j, \alpha_j \mid \mathbf{X}, y_{ij}^*)$. For $j = 1, \ldots, m$, sample $\beta_j^{(t)}$ and $\alpha_j^{(t)}$ from the multivariate normal density with mean vector $[\mathbf{X}^{*\prime}\mathbf{X}^* + \mathbf{T}_0^{-1}]^{-1}[\mathbf{X}^{*\prime}\mathbf{y}_{\cdot j}^{*(t)} + \mathbf{T}_0^{-1}\boldsymbol{\tau}_0]$ and variance-covariance matrix $[\mathbf{X}^{*\prime}\mathbf{X}^* + \mathbf{T}_0^{-1}]^{-1}$, where $\mathbf{X}^*$ is an n-by-$(d+1)$ matrix with typical row $\mathbf{x}_i^* = (\mathbf{x}_i^{(t-1)}, -1)$, $\mathbf{y}_{\cdot j}^{*(t)}$ is an n-by-1 vector of sampled latent utility differentials for the jth roll call, and $N(\boldsymbol{\tau}_0, \mathbf{T}_0)$ is the prior for $\boldsymbol{\beta}_j$ and α_j. I set $\boldsymbol{\tau}_0 = 0$ and $\mathbf{T}_0 = \kappa \cdot \mathbf{I}_d$, where κ is an arbitrarily large constant (e.g., 25 or 100), to give a vague prior centered at the origin. This amounts to running Bayesian regressions of $\mathbf{y}_{\cdot j}^{*(t)}$ on $\mathbf{x}_i^{(t-1)}$ and a negative intercept and then sampling from the posterior density for the coefficients $\boldsymbol{\beta}_j$ and α_j, for $j = 1, \ldots, m$.

3. $g(\mathbf{x}_i \mid y_{ij}^*, \boldsymbol{\beta}_j, \alpha_j)$. Rearranging the latent linear regression yields $w_{ij} = y_{ij}^* + \alpha_j = \mathbf{x}_i' \boldsymbol{\beta}_j + \varepsilon_{ij}$. Collapse these equations over the j subscript to yield the n regressions $\mathbf{w}_i = \mathbf{B}\mathbf{x}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{B}$ is the m-by-d matrix with the jth row given by $\boldsymbol{\beta}_j'$. That is, we have n regressions, with the ideal points $\mathbf{x}_i$ as parameters to be updated. Again exploiting conjugacy, the update is performed by sampling each $\mathbf{x}_i^{(t)}$ from the d-dimensional normal density with

mean vector $(\mathbf{B'B} + \mathbf{V}_i^{-1})^{-1}(\mathbf{B'w}_i + \mathbf{V}_i^{-1}\mathbf{v}_i)$ and variance-covariance matrix $(\mathbf{B'B} + \mathbf{V}_i^{-1})^{-1}$, where $\mathbf{v}_i$ and $\mathbf{V}_i$ are the prior means and variance-covariance matrices for the ideal point of the ith legislator. Here I set $\mathbf{v}_i = 0$ and $\mathbf{V}_i = \mathbf{I}_d$ for all legislators.

Implementing this scheme requires only modest programming skill, and I have free software (Jackman 2003) that implements this algorithm via the R statistical package (also free). See also the implementation of Martin & Quinn (2003).

As an example, I use the Gibbs sampler to estimate a unidimensional spatial voting model fit to a small set of roll call data: all 213 non-unanimous decisions of the seventh "natural" Rehnquist court ($n = 9$) between 1994 and 1997 (Justices Rehnquist, Stevens, O'Connor, Scalia, Kennedy, Souter, Thomas, Ginsberg, and Breyer), which appear in the Spaeth (2001) data set. The decisions of the justices are coded as $y_{ij} = 1$ if justice i joins the majority on case j, and $y_{ij} = 0$ if he or she dissents; there are ten abstentions in the data. The Gibbs sampler is initialized at arbitrary values and allowed to run for half a million iterations; we retain only every thousandth iteration so as to produce an approximately independent sequence of sampled values from the joint posterior density. I (somewhat conservatively) discard the first half of the run as "burn-in," leaving 250 sampled values for analysis.

Figure 4 shows the iterative history or "trace plots" of the Gibbs sampler for the nine ideal points (one for each justice). The plots strongly suggest that the Gibbs sampler has converged on the posterior density for the ideal points; the traces appear to be random walks around the respective posterior means. Figure 5 shows the posterior means and confidence intervals for the justices' ideal points. Stevens is far away from the other justices, and Thomas and Scalia anchor the conservative end of the Court.

Finally, Figure 6 shows all possible pairwise slices through the joint posterior density of the ideal points. It is interesting to note that the ideal points are not independent a posteriori, a fact that is easily overlooked when we simply consider ideal point estimates and their standard errors via classical approaches. That is, in determining whether justice a was to the right or left of justice b, we would not simply rely on the pointwise confidence intervals in Figure 5, which are summaries of the uncertainty in each ideal point, and ignore the substantial covariation in ideal points.

Note that it is also possible to perform inference with respect to any function of the parameters; for instance, if we are interested in an estimand $\boldsymbol{\eta} = g(\mathbf{X})$, then we simply compute $\boldsymbol{\eta}^{(t)} = g(\mathbf{X}^{(t)})$ at iteration t of the Gibbs sampler and store the output. We can then perform inference for $\boldsymbol{\eta}$. Examples include pairwise comparisons of ideal points, rank ordering the ideal points, the location of the median justice, cutting planes for each bill (or case), residuals, and goodness-of-fit summaries. The basic model introduced here has been extended in numerous directions to handle change in ideal points over time (Martin & Quinn 2002) and evolution of the legislative agenda (Clinton & Mierowitz 2001); as Clinton
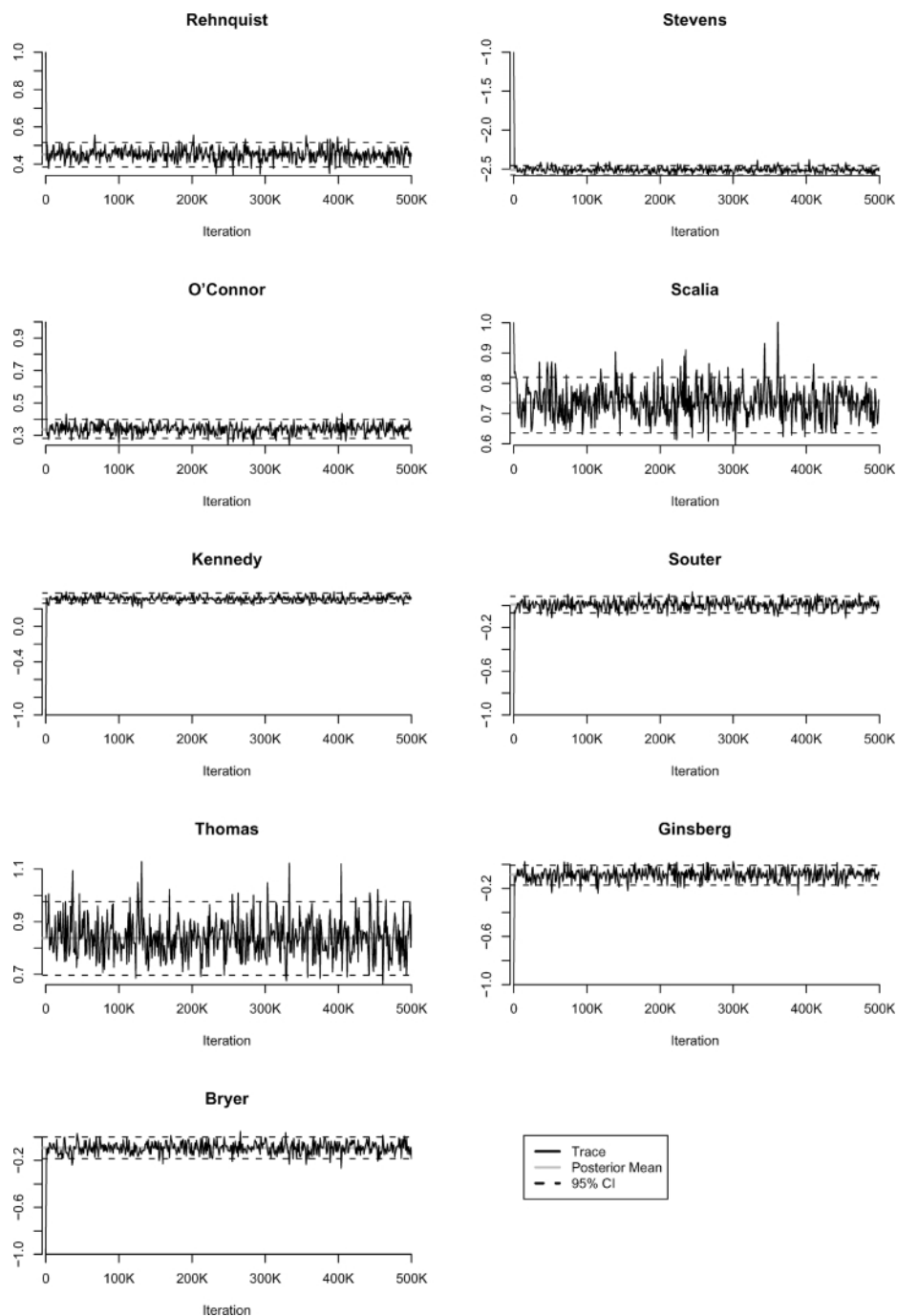
Figure 4   Trace plots.

Figure 5 Posterior means and pointwise 95% confidence intervals. Means and confidence intervals are computed using the last 250,000 iterations (thinned by 1000).
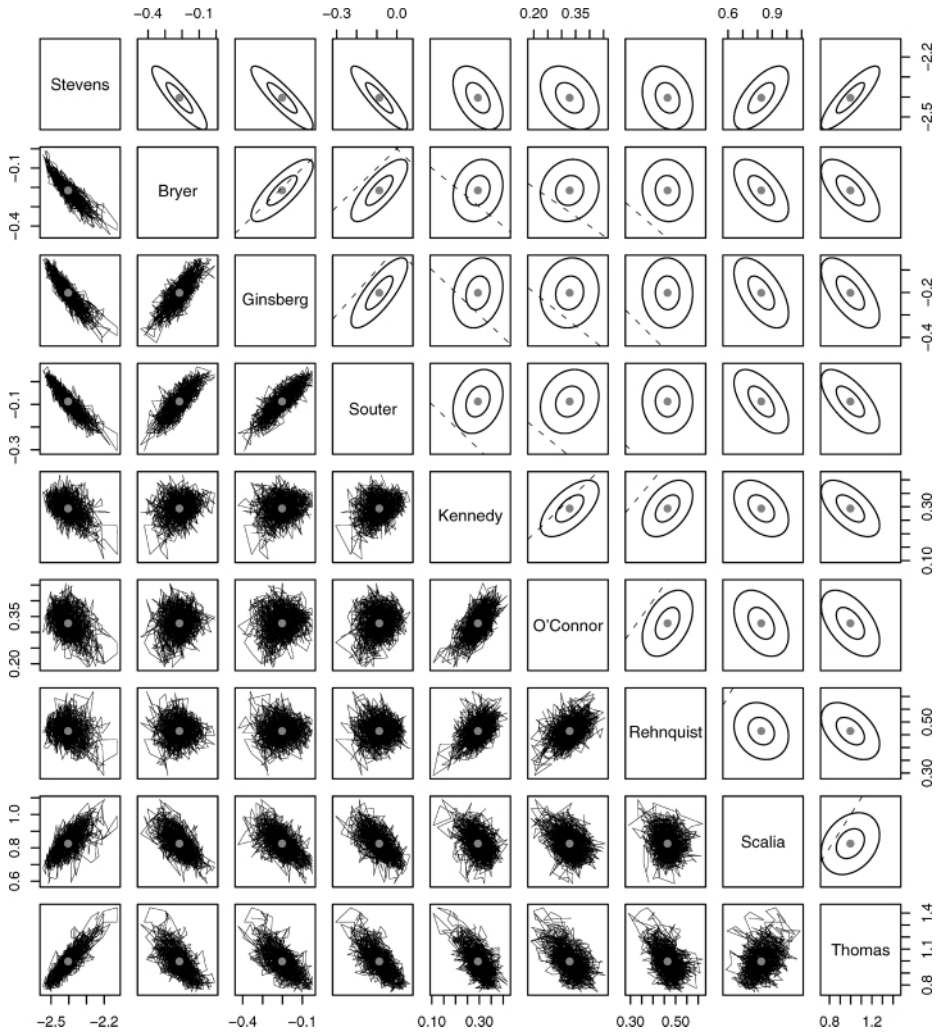
Figure 6 Pairwise trace plots and posterior densities. Traces and summaries are for the last 250,000 iterations (thinned by 1000). The ellipses mark 50% and 95% confidence intervals around the mean, using a bivariate normal approximation.

et al. (2004) highlight, these extensions are easily implemented using the Bayesian machinery presented here.

7. CONCLUSION

This review has contrasted two styles of Bayesian analysis used in political science. To illustrate the basics of Bayesian analysis (and the central role of the prior), I showed how a simple conjugate Bayesian analysis lets us integrate historical data

with current information about an estimand of substantive political interest (likely election outcomes in Florida). In a Bayesian analysis of this sort, where the prior is not innocuous, sensitivity analysis is worth doing and worth communicating to one's readers. In the second part of this review, the emphasis is on Bayesian computation. In the example of roll call data analysis, I attacked a high-dimensional statistical problem with the Gibbs sampler. Although I placed vague priors on all model parameters, the analysis is nonetheless Bayesian. I exploited Bayesian tools for estimation (MCMC algorithms) and the Bayesian approach to inference; that is, the posterior density summarizes uncertainty in the parameters after looking at the roll call data, and unlike in classical approaches, we need not rely on speculative notions of asymptotics and/or repeated sampling to support our inferences.

Constraints of space mean that other political science applications of Bayesian statistics cannot be considered here. These include hierarchical modeling (e.g., Western 1998) and model averaging (e.g., Bartels & Zaller 2001). These applications also exploit the essential ingredients of Bayesian analysis discussed in this review: priors and the new tools that dramatically simplify Bayesian computation. Detailed treatments of these topics and others appear in the book-length treatments of Carlin & Louis (2000) and Gill (2002), the latter specially tailored for a social science readership. Bayesian analysis via MCMC has been extremely simplified by the WinBUGS computer program (Spiegelhalter et al. 2000); books by Congdon (2001, 2003) contain many examples of Bayesian analysis with Win-BUGS, largely drawn from the biostatistical literature but easily applicable to the social sciences.

The past 15 years have seen something of a revolution in statistics. The popularization and widespread adoption of MCMC algorithms mean that models and data sets long relegated to the "too-hard" basket are now being analyzed. The effects of this revolution are beginning to be felt in political science, with the result that Bayesian analysis is no longer just an exotic tool for methodological specialists and is finding a place in the toolkit of workaday political scientists.

ACKNOWLEDGMENTS

**The *Annual Review of Political Science* is online at
http://polisci.annualreviews.org**

LITERATURE CITED

Albert J. 1992. Bayesian estimation of normal ogive item response curves using Gibbs sampling. *J. Educ. Stat.* 17:251–69

Albert JH, Chib S. 1993. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* 88:669–79

Bartels LM, Zaller J. 2001. Presidential vote models: a recount. *PS: Polit. Sci. Polit.* 34:9–20

Bayes T. 1763. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.* 53:370–418

Bayes T. 1958. An essay towards solving a problem in the doctrine of chances. *Biometrika* 45:293–315

Bernardo J, Smith FM. 1994. *Bayesian Theory.* Chichester, UK: Wiley

Carlin BP, Louis TA. 2000. *Bayes and Empirical Bayes Methods for Data Analysis.* London: CRC Press. 2nd ed.

Casella G, George EI. 1992. Explaining the Gibbs Sampler. *Am. Stat.* 46:167–74

Chib S, Greenberg E. 1995. Understanding the Metropolis-Hastings algorithm. *Am. Stat.* 49:327–35

Clinton JD, Mierowitz A. 2001. Agenda constrained legislator ideal points and the spatial voting model. *Polit. Anal.* 9:242–59

Clinton JD, Jackman S, Rivers RD. 2004. The statistical analysis of roll call data. *Am. Polit. Sci. Rev.* 98(2). In press

Congdon P. 2001. *Bayesian Statistical Modelling.* Chichester, UK: Wiley

Congdon P. 2003. *Applied Bayesian Modelling.* Chichester, UK: Wiley

Edwards AWF. 1992. *Likelihood.* Baltimore, MD: Johns Hopkins Univ. Press. Expanded Ed.

Enelow J, Hinich M. 1984. *The Spatial Theory of Voting: An Introduction.* New York: Cambridge Univ. Press

Gamerman D. 1997. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* London: Chapman & Hall

Gelfand AE. 1997. Gibbs sampling. In *Encyclopedia of the Statistical Sciences,* ed. S Kotz, CB Read, DL Banks, 1:283–92. New York: Wiley

Gelman A, Carlin JB, Stern HS, Rubin DB. 1995. *Bayesian Data Analysis.* London: Chapman & Hall

Gilks WR, Richardson S, Spiegelhalter DJ. 1996. Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice,* ed. WR Gilks, S Richardson, DJ Spiegelhalter, pp. 1–16. London: Chapman & Hall

Gill J. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach.* New York: Chapman & Hall

Hacking I. 2001. *An Introduction to Probability and Inductive Logic.* Cambridge, UK: Cambridge Univ. Press

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains, and their applications. *Biometrika* 57:97–109

Jackman S. 2000. Estimation and inference are missing data problems: unifying social science statistics via Bayesian simulation. *Polit. Anal.* 8:307–32

Jackman S. 2001. Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference and model checking. *Polit. Anal.* 9:227–41

Jackman S. 2003. *Rollcall: A R Library for the Analysis of Roll Call Data.* Stanford, CA: Stanford Univ., Polit. Sci. Comput. Lab.

Johnson VE, Albert JH. 1999. *Ordinal Data Modeling.* New York: Springer-Verlag

King G. 1989. *Unifying Political Methodology.* New York: Cambridge Univ. Press

Lancaster T. 2000. The incidental parameter problem since 1948. *J. Econometrics* 95:391–413

Leamer E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley

Lee PM. 1989. *Bayesian Statistics: An Introduction.* Oxford, UK: Oxford University Press

Lindley DV. 2001. Thomas Bayes. In *Statisticians of the Centuries,* ed. CC Heyde, Seneta E, pp. 68–71. New York: Springer-Verlag

Martin AD, Quinn KM. 2002. Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Polit. Anal.* 10:134–53

Martin AD, Quinn KM. 2003. *MCMCpack.* St. Louis, MO: Washington Univ. Dep. Polit. Sci.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equations

of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087–91

Metropolis N, Ulam S. 1949. The Monte Carlo method. *J. Am. Stat. Ass.* 44:335–41

Patz RJ, Junker BW. 1999. A straigtforward approach to Markov chain Monte Carlo methods for item response models. *J. Educ. Behav. Stat.* 24:146–78

Poole KT, Rosenthal H. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford Univ. Press

Spaeth HJ. 2001. United States Supreme Court judicial database: 1953–1997 terms. ICPSR Study 9422

Spiegelhalter DJ, Thomas A, Best N. 2000. *WinBUGS Version 1.3*. Cambridge, UK: MRC Biostat. Unit

Stigler S. 1986a. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Belknap Press of Harvard Univ. Press

Stigler S. 1986b. Laplace's 1774 memoir on inverse probability. *Stat. Sci.* 1:359–78

Tanner MA. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer-Verlag. 3rd ed.

Tierney L. 1996. Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice*, ed. WR Gilks, Richardson S, Spiegelhalter DJ, pp. 59–74. London: Chapman & Hall

Western B. 1998. Causal heterogeneity in comparative research: a Bayesian hierarchical modelling approach. *Am. J. Polit. Sci.* 42:1233–59

Ⓡ*Annual Review of Political Science*
Volume 7, 2004

CONTENTS

ERRATA
An online log of corrections to *Annual Review of Political Science*
chapters may be found at http://polisci.annualreviews.org/