Bayesian Modeling and Inference: A Postmodern Perspective

Jeff Gill and Simon Heuberger[1]

INTRODUCTION

Bayesian methods and Bayesian inference are now ubiquitous in the social sciences, including political science and international relations. The reasons for this are numerous and include: a superior way to describe uncertainty, freedom from the deeply flawed Null Hypothesis Significance Testing paradigm, the ability to include previous information, more direct description of model features, and, most, recently statistical computing tools that make model computations easy. Yet this is not a static area in social science methodology, and new methodological developments are published at a rapid pace. The objective of this handbook chapter is to describe basic Bayesian methods, chronicle the recent history of their application in political science and international relations, and then highlight important recent developments.

The key tenets of Bayesian inference are that all unknown quantities are assigned probability statements, these probability statements are updated when new information (data) are observed, and the results are described distributionally. This does not sound like a radical departure from traditional statistical thinking, but it constitutes a different way of thinking about uncertainty that is strictly probability based, eschewing assumptions like an unending stream of independent and identically distributed (IID) data. Furthermore, this paradigm comes along with some historical and practical conventions:

- Overt and clear model assumptions.
- A rigorous way to make *probability* statements about the real quantities of theoretical interest.
- An ability to update these statements (i.e. learn) as new information is received.
- Systematic incorporation of *qualitative* knowledge on the subject of interest.
- Recognition that population quantities are changing over time rather than fixed immemorial.
- Straightforward assessment of both model quality and sensitivity to assumptions.
- Freedom from the flawed null hypothesis significance testing (NHST) paradigm.

These come mostly from a different intellectual path than more conventionally accepted procedures, including relentless criticism during a roughly 100-year period by giants of the field. For the purpose of this exposition we divide the history of Bayesian statistics, as just described, into three eras: classical, modern, and postmodern. Each of these is discussed with attention to issues relevant to social science research. Our objective is to provide a chapter that is both introductory and a summary of recent research trends.

CLASSICAL BAYES

Bayes' Law was published in 1763, but the classical era of Bayesian statistical inference occupied the middle of the 20th century up until 1990. Much has been written about this time (Gelman and Shalizi, 2013, Strevens, 2006) but the defining hallmark was a sense of principled philosophical correctness combined with analytical frustration. That is, it was not particularly difficult to write logical and mathematical arguments that demonstrated the inferential, and general scientific, superiority of Bayesian inference over classical methods for real problems, but it was often hard to execute this process for real and pragmatic data problems. Practical work often took incremental and modest directions based on what was mathematically possible. So, one sees publications during this time that give 'recipes' for broad classes of models that one could adapt to their specific question and therefore get results. The key problem here was the common creation of Bayesian models where the final joint specification of the model could not be mathematically decomposed into constituent parts to create a standard regression table of results.

Our starting point is much earlier. Bayes' (1763) famous essay was published two years after his death. The paper's release was facilitated by his friend Richard Price rather than Bayes himself (both are now buried in Bunhill Cemetery in central London). A reverend and amateur mathematician, Bayes developed a probability theorem that relates the order of conditional probabilities. It turns out that Laplace (1774, 1781) was developing similar ideas on the continent but communication was not without problem at the time. Let A and B be two non-independent events. Basic probability laws tell us that the conditional probability of A given B is given by $p(A \mid B) = \frac{p(A,B)}{p(B)}$, with $p(A, B)$ being the probability of A and B both occurring and $p(B)$ the unconditional probability that the event B occurs. We thus obtain the probability of A conditioned on B happening from $p(A \mid B)$. Conditional probability can also be defined in the reversed order, with A happening before considering B, that is, $p(B \mid A) = \frac{p(B,A)}{p(A)}$. The joint probability remains the same in both directions, meaning that $p(A, B) = p(B, A)$. Since $p(A,B) = p(A \mid B)p(B)$, we can thus surmise that $p(A|B)p(B) = p(B|A)p(A)$ and by rearranging $p(A \mid B) = \frac{p(A)}{p(B)} p(B \mid A)$.

The statistical application of Bayes' Law begins with the definition of the likelihood function (Fisher, 1925a, Birnbaum, 1962), which is nothing more than the joint probability of the observed data. Define the collected data $\boldsymbol{X}$ that is observed once and therefore treated as a fixed quantity. We assume that we know the respective probability mass/density function for these data to accurately describe the data-generating process: $p(X|\beta)$, conditioned on some unknown parameter β to be estimated. This joint function is then defined by

$$L(\beta \mid \boldsymbol{X}) = \prod_{i=1}^{n} p((X_i \mid \beta) = p(X_1 \mid \beta) \times p(X_2 \mid \beta) \times \cdot \times p(X_n \mid \beta). \quad (1)$$

The switching of the order of conditionality in this statement is not an application of Bayes' Law, but is Fisher's (1925a) recognition that once the data are observed, they are fixed for inferential purposes, and it is the unknown parameter of interest that is the focus of inference at this point. This perspective differs dramatically from canonical Frequentist inference whereby there is an unending stream of independent and identically distributed data available to the researcher (Gill, 1999).

Fisher (1925b) and others provide the standard and convenient inferential treatment of the likelihood function, which is to find the multi-dimension mode of this – usually log-concave to the x-axis – function and to treat curvature around this mode as a measure of uncertainty. These procedures are incredibly well-established and well-accepted and need not be reviewed here. See Gill and Torres (2019) for a recent detailed exposition.

The next required quantity for Bayesian inference is the *prior distribution*: the presumed distribution of the parameters of interest before the current data for analysis are collected and analyzed: $p(\beta)$. Notice that this is not conditioned on anything explicitly in its formulation. In truth it is conditioned on previous knowledge about the effect of interest that is symbolized by the parameter β in the regression context. So if the effect of interest is size of nations' militaries and the outcome of interest is the proclivity to wage militarized conflict, then we would typically notate the magnitude (regression coefficient) of the former as β and the vector outcomes as $\boldsymbol{Y}$. The likelihood function, $L(\beta|\boldsymbol{X})$, is central to Bayesian methods because it enables the researcher to find a most probable value $\hat{\beta}$ by testing differing values of β. The value $\hat{\beta}$ is the most likely to have generated the data given H_0 expressed through a specific parametric form relative to other possible values in the sample space of β. In this regard, the likelihood function resembles the inverse probability, $p(H_0|\boldsymbol{X})$, but as we will see, this clearly is not the socially desired value that readers sometimes quest for but is not provided under this framework: the probability of the null hypothesis given some data. Crucially, however, the produced likelihood function from this setup now matters only in relation to other likelihood functions with differing values of β since it is no longer bounded by 0 and 1.

Bayesian inference combines data information in the likelihood function with researchers' prior information. This is done by conditioning the prior on the likelihood using Bayes' Law,

$$\pi(\beta \mid \boldsymbol{X}) = \frac{p(\beta)}{p(\boldsymbol{X})}\, p(\boldsymbol{X} \mid \beta), \tag{2}$$

which 'flips' the order of the conditional distribution by using the ratio of the prior to the unconditional distribution of the data. Now recall the Bayesian maxim that once a quantity is observed, it is treated as fixed. Here the data are assumed to be observed once and fixed in perpetuity for that given collection. Therefore, the probability of observing the observed data is simply one: $p(\boldsymbol{X}) = 1$. This means that the denominator in (2) can be treated as simply a normalizing constant to ensure that $\pi(\beta|\boldsymbol{X})$ sums or integrates to one and is therefore a normalized distribution. Since we can always recover this normalizing information later in the inferential process, it is customary and convenient to express (2) with proportionality:

$$\pi(\beta \mid \boldsymbol{X}) \propto p(\beta) L(\beta \mid \boldsymbol{X}). \tag{3}$$

Here we have also used the more intuitive form of the likelihood function, as described above. So it is common to say that the posterior is just proportional to the prior times the likelihood. So Bayesian inference is codified by a compromise between prior information and data information.

One source of discomfort for Bayesians during this era was the specification of prior distributions for unknown parameters. This is primarily because non-Bayesian scholars of the era criticized prior specification as

'subjective' since it came from sources outside of $\boldsymbol{X}$, as determined by the researcher. There were many aspects of this view that were misguided. First, all modeling decisions, in all statistical settings, are subjective. It is subjective what data to use. It is subjective what likelihood function to specify (e.g. link function, etc.). It is subjective which estimation procedure to use. It is subjective what software to use. It is subjective how to present results, and so on. In addition, prior specifications are opportunities to include qualitative information, known earlier research on a question of interest, values to be updated over time, and more. Also, prior distributions can be constructed to have many different mathematical properties. Despite all of these facts, the specification of priors remains as a pointedly discussed issue well into the 21st century. We will return to this topic at several points over the course of this chapter.

The difficult classical period of Bayesian statistics produced different strains of prior forms for different purposes. Sometimes these were in competition and sometimes they overlapped. Importantly, the bulk of the 20th century was a period when Bayesians needed to be very circumspect in selecting and describing prior selection since non-Bayesians routinely used this as leverage for criticism. Zellner (1971: 18) divides prior distributions during this time into two types: *data-based* (DB) types that are 'generated in a reasonably scientific manner and which are available for further analysis', and *nondata-based* (NDB) types that 'arise from introspection, casual observation, or theoretical considerations'. The critical problem faced by Bayesians of the classical era is that 'one person's NDB prior information can differ from that of another's' (Zellner, 1971: 19). However, this issue is not limited to NDB priors: 'It is possible that two investigators working with the same model and DB prior information can arrive at different posterior beliefs if they base prior information on different bodies of past data' (Zellner, 1971: 19). Thus, something as simple as a difference of opinion on relevant prior information divided some Bayesians and left the door open for the classical 'subjective' criticism in prior selection.

In truth there was a defined set of principled prior approaches that emerged from this challenging era. Proper Bayes is the group most accurately described by Zellner above: those that constructed prior distributions from compiled evidence, such as earlier studies or published work, researcher intuition, or substantive experts. This is a rich literature that seeks to be build on existing scientific knowledge but emboldens the subjective criticism. Empirical Bayes uses the data not only in the likelihood function but also to estimate these hyperpriors values (parameters of prior distributions). This can be done with or without a specific distributional form at this level (parametric versus nonparametric empirical Bayes, see Morris, 1983). This approach offends some Bayesians because the data are used in both right-hand-side elements of (Equation 3). Lindley (in Copas, 1969: 421) accordingly observed that '…there is no one less Bayesian than an empirical Bayesian'. Reference Bayes seeks priors that move the prior as little as possible from the likelihood function (Bernardo, 1979) – minimizing the distance between the chosen likelihood and the resulting posterior according to a criteria like the Kullback-Leibler distance, which comparatively measures distributions (this is also referred to as dominant likelihood prior). Note that the term 'reference prior' sometimes confusingly also refers to a prior that is used as a default (Box and Tiao, 1973: 23). Related to this approach, but different in purpose, are priors that seek to minimize any sense of subjective information: diffuse or uninformative priors such as parametric forms with large variance or uniform specification (bounded and unbounded). These were often specified because it reduced arguments with non-Bayesians and sometimes led to easily calculated results.

Continuous unbounded uniform priors were referred to as 'improper' since they did

not integrate to one over the real line, yet they generally yielded proper posteriors due to a cancellation in Bayes' Law. Consider an unbounded uniform prior for a regression parameter defined by $p(\beta) = hf(\beta)$ over $[-\infty{:}\infty]$. This is essentially a rectangle of height $hf(\beta)$ and width $k = \infty$. So, re-expressing Bayes' Law with this prior gives

$$\pi(\beta \mid \boldsymbol{X}) = \frac{p(\beta)p(\boldsymbol{X} \mid \beta)}{\int_{-\infty}^{\infty} p(\beta)p(\boldsymbol{X} \mid \beta)d\beta}, \quad (4)$$

which is (2) with the replacement

$$\begin{aligned} &\int_{-\infty}^{\infty} p(\beta)p(\boldsymbol{X} \mid \beta)d\beta \\ &= \int_{-\infty}^{\infty} p(\beta, \boldsymbol{X})d\beta = p(\boldsymbol{X}) \end{aligned} \quad (5)$$

by the definition of conditional probability and the process of integration for β, which is hypothetical to us in practical terms but exists in nature. Therefore, replacing the definition of the improper rectangular prior gives

$$\begin{aligned} \pi(\beta \mid \boldsymbol{X}) &= \frac{(hf(\beta) \times k)p(\boldsymbol{X} \mid \beta)}{\int_{-\infty}^{\infty} (hf(\beta) \times k)p(\boldsymbol{X} \mid \beta)d\beta} \\ &= \frac{k}{k} \frac{hf(\beta)p(\boldsymbol{X} \mid \beta)}{\int_{-\infty}^{\infty} hf(\beta)p(\boldsymbol{X} \mid \beta)d\beta} \\ &\propto f(\beta)p(\boldsymbol{X} \mid \beta), \end{aligned} \quad (6)$$

where the cancellation of k / k occurs because these are the exact same flavor of infinity, otherwise it is undefined. Thus, the posterior is proportional to the likelihood times some finite prior (the h values can also be canceled or left off due to proportionality). A similar prior for a variance component is $p(\sigma) = 1/sigma$ over $[0{:}\ \infty]$. Unfortunately, as mathematically tractable as improper priors are, they did not overwhelmingly convince Bayesian skeptics, some of whom considered it a form of arithmetic trickery.

This classification of classical priors presents the most common forms but is by no means exhaustive. Others include maximum entropy priors, mixtures priors, decision-theoretic priors, conjugate priors, and more (O'Hagan, 1994). Conjugate priors can be informed or diffuse and provide an especially nice choice since, for a given likelihood function, the parametric form of the prior flows down to the posterior giving simple calculations. In decision-theoretic Bayesian inference (Gill, 2014), results are presented in a full decision-theoretic framework where utility functions determine decision losses and risk, which are minimized according to different probabilistic criteria. These approaches were especially appealing in an age with limited computational tools.

MODERN BAYES

An important hallmark of the classical era of Bayesian statistics was the relative ease with which a joint posterior distribution that could not be integrated into the constituent marginal distributions could be produced with actual social science data and a realistic model based on theoretical principles. Consider the following example that motivated the work in Gill and Waterman (2004) using the data collected by Mackenzie and Light (1987) on every US federal political appointee to full-time positions requiring Senate confirmation from November, 1964 through to December, 1984 (the collectors of the data needed to preserve anonymity so they embargoed some variables and randomly sampled 1,500 cases down to 512). This survey queries various aspects of the Senate confirmation process, acclamation to running an agency or program, and relationships with other functions of government. A key question is why these executives last only about two years on average after assuming the position, given the difficulty of the process. This work hypothesizes that running a federal agency (or sub-agency) is

considerably more stressful than alternative positions for these executives. The outcome variable of interest is `stress` as a surrogate measure for self-perceived effectiveness and job satisfaction, measured as a five-point scale from 'not stressful at all' to 'very stressful'. Explanatory variables specified for the model are: `Government Experience`, `Ideology`, `Committee Relationship`, `Career.Exec-Compet`, `Career.Exec-Liaison/Bur`, `Career.Exec-Liaison/Cong`, `Career.Exec-Day2day`, `Career.Exec-Diff`, `Confirmation Preparation`, `Hours/Week`, and `President Orientation` (Gill and Casella, 2009, for detailed descriptions). A Bayesian random effects specification for ordered survey outcomes is specified, so latent variable thresholds for $\boldsymbol{Y}$ are assumed to be on the ordering:

$$\boldsymbol{U}_i : \theta_0 \underset{c=1}{\Longleftrightarrow} \theta_1 \underset{c=2}{\Longleftrightarrow} \theta_2 \underset{c=3}{\Longleftrightarrow} \theta_3 \ldots \theta_{C-1} \underset{c=C}{\Longleftrightarrow} \theta_C.$$

The vector of (unseen) utilities across individuals in the sample, $\boldsymbol{U}$, is determined by a linear additive specification of K explanatory variables: $\boldsymbol{U} = -\boldsymbol{X}'\gamma + \epsilon$, where $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_p]$ does not depend on the θ_j, and $\varepsilon \sim F_\varepsilon$. This means that the probability that individual i in the sample is observed to be in category r or lower is

$$\begin{aligned} P(\boldsymbol{Y}_i \leq r \mid \boldsymbol{X}_i) &= P(\boldsymbol{U}_i \leq \theta_r) \\ &= P(\epsilon \leq \theta_r + \boldsymbol{X}_i'\gamma) \\ &= F_{\epsilon_i}(\theta_r + \boldsymbol{X}_i'\gamma), \end{aligned} \tag{7}$$

which is differently signed than in `R`: '**logit P(Y ¡= k — x) = zeta_k – eta**' from the help page (there is no fixed convention on the sign). Specifying a logistic distributional assumption on the errors and adding the random effect term produces this logistic cumulative specification for the whole sample:

$$\begin{aligned} F_\epsilon(\theta_r + \boldsymbol{X}'\gamma) &= P(\boldsymbol{Y} \leq r \mid \boldsymbol{X}) \\ &= [1 + \exp(-\theta_r - \boldsymbol{X}'\gamma)]^{-1}. \end{aligned} \tag{8}$$

Prior distributions are given to be either semi-informed or skeptical:

$p(\gamma_k) \sim \mathcal{N}(\mu_{\gamma_k}, \sigma_\gamma^2), k = 1, \ldots, k$ for each of the K explanatory variables,

$p(\theta_j) \sim \mathcal{N}(0, \sigma_\theta^2), j = 1, \ldots, C-1$ for the four latent variable thresholds, with assigned hyperparameter values μ_{γ_k}, σ_γ^2, σ_θ^2

All this produces a joint posterior distribution according to

$$\begin{aligned} &\pi(\gamma, \theta \mid \boldsymbol{X}, \boldsymbol{Y}) \propto L(\gamma, \theta \mid \boldsymbol{X}, \boldsymbol{Y}) p(\theta) p(\gamma) \\ &\propto \prod_{i=1}^{n} \prod_{j=1}^{C-1} \prod_{k=1}^{p} \left[\Lambda(\theta_j - \boldsymbol{X}_i'\gamma + \boldsymbol{b}_i) - \Lambda(\theta_{j-1} - \boldsymbol{X}_i'\gamma) \right]^{\tau_{ij}} \\ &\quad \times \exp\left(-\frac{(\gamma_k - \mu_{\gamma_k})^2}{2\sigma_\gamma^2} - \frac{\theta_j^2}{2\sigma_\theta^2} \right), \end{aligned} \tag{9}$$

which is kind of ugly (and hard marginalize). While this form tells us everything we need to know about the *joint distribution* of the model parameters, we need to marginalize (integrate) it for every one of these parameters to create an informative regression table. That is, this is a joint probability statement like $p(A, B, C)$ for arbitrary example random variables A, B, and C, and we need to create the marginal (solitary) probability statements for each. Continuing the contrived example for one of these, say A, we get $P(A) = \int_B \int_B p(A, B, C) dB dC$ from elementary integral calculus. Using $P(A)$, we can get the mean and standard error to present as the first and second column of a regression table for A in the conventional way. The key point is

that it was easy in the middle of the 20th century to produce a model such as this whereby it was prohibitively difficult or impossible to integrate-out each parameter for a series of marginal posterior distributions to describe inferentially. This led Evans (1994) to retrospectively describe Bayesians of the time as 'unmarried marriage guidance councilors' because they could tell others how to do inference when they often could not do it themselves.

This world, and *the* world, changed in 1990 with a review essay by Gelfand and Smith in the *Journal of the American Statistical Association*. In what is without a doubt one of the ten most important articles published in statistics, they found that a tool, *Gibbs sampling*, hiding in engineering and image restoration (Geman and Geman, 1984), solved this very problem for the Bayesians. Gibbs sampling replaced analytical derivation of marginals from a joint distribution with work from the computer. In this algorithm, the full conditional distributions for each parameter to be estimated are expressed with the conditionality on the other parts of the model that are relevant for this parameter. Then the sampler draws consecutively at each step with the latest drawn versions of the conditioned parameters until the Markov chain reaches its stationary (converged) distribution and these conditional draws can then be treated as marginal samples. This means that the marginal distributions of interest from complicated model specifications can simply be described empirically from a large number of draws sampled by the computer. The date of 1990 is not just important for the publication of this article, but also because, by 1990, statisticians (and others) were free from centralized mainframe computing (with attendant punch cards, JCL, batch processing, and other encumbrances) via ubiquitous and reasonably powerful desktop machines. The popularity of bootstrapping in the 1980s presaged this development for the same computational reason.

The general name for the new tool is *Markov chain Monte Carlo* (MCMC), which includes Gibbs sampling as well as the *Metropolis-Hastings algorithm* that was dormant in statistical physics but has existed since the 1953 paper by Metropolis et al. in the *Journal of Chemical Physics* (as was the slightly sexist, somewhat quaint, custom of the time in some of the natural science fields, their wives typed the paper and were added as coauthors). The principle behind MCMC is that a Markov chain can be setup to describe a high dimension posterior distribution by wandering around the state space visiting subregions in proportion to the density that needs to be summarized. The linkage between 'MC' and 'MC' is the *ergodic theorem* that says that if the Markov chain is created according to some specific technical criteria and is run long enough such that it converges to its stationary (limiting) distribution, the draws from the path of the chain can be treated as if they are IID from the respective marginal distributions. Thus, in the stationary distribution, each step is recorded as a multidimensional draw as if from a regular Monte Carlo process, even though these draws are actually produced from a serial, conditional Markovian algorithm. Robert and Casella (2004), Gill (2014), and Gamerman and Lopes (2006) provide the necessary theoretical background, and Robert and Casella (2011) give an entertaining history of this development in statistics during the early 1990s.

A convenient way to visualize how the MCMC process works practically is given in Figure 50.1, which shows the general data structure as produced by running the Markov chain for a model with 5 parameters to be estimated $[\theta_1,\ldots,\theta_5]$ over $m = 1,\ldots, M$ MCMC simulations after convergence is asserted. The arrow on the right indicates the direction of the production of simulated vectors. The serial Markovian process produces one row at a time that is conditional on previous rows. Thus, row I_1 is followed by row I_2 on up to row I_{499999} and row I_{500000} (the second half of the full chain run of

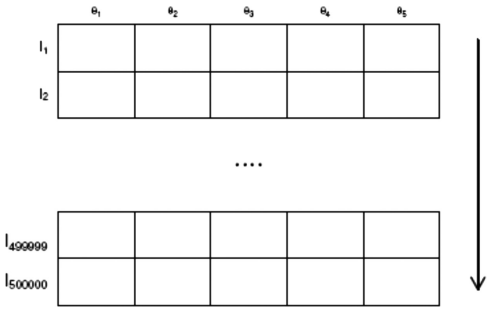

Figure 50.1 How Markov Chain Monte Carlo works

1 million iterations) as the chain explores the six-dimensional space (5 parameters plus posterior density). Each row is a sample, and the ergodic theorem states that under specific circumstances (met in practice with standard algorithms) in the converged state, each of these rows can be treated like an independent multidimensional draw from the posterior of interest produced by the Bayesian regression model. Inference is now produced by analyzing down columns for each parameter. From the ergodic theorem we can take the column vector, for say θ_1, and summaries for this parameter are performed by simply applying means, variances, quantiles, etc. on the draws down the first column in this case. Thus the 'MC' (Markov chain) that produces the rows is later ignored and the second 'MC' (Monte Carlo) is used to get inferences.

To say that MCMC revolutionized Bayesian inference and general Bayesian work is to say that commercial air flights slightly improved global travel. It was in fact the tool that freed the Bayesians; no longer were models too complex for the marginalization of joint posteriors to create regression tables and other intuitive summaries. The result was a flood of Bayesian applications that were pent-up in researchers' file drawers giving a Bayesian perspective to a wide swath of model types and general statistical approaches. What got relatively left behind in this revolution was introspection about the choice of priors. Actually, prior choice was sublimated down to diffuse normals, uniforms, and other vague forms since the focus had turned to issues of estimation: the 1990s saw an explosion of customized MCMC procedures, calculations of quantities of interest from posterior samples (Chib, 1995; Chib and Jeliazkov, 2001), and a deeper understanding of Markov chain theoretical principles (Gilks et al., 1996; Polson, 1996; Roberts and Tweedie, 1996; Brooks et al., 1997; Roberts and Rosenthal, 1998; Brooks and Roberts, 1999; Robert and Mengersen, 1999; Neal, 2003). In the social sciences, applied Bayesian researchers defaulted almost exclusively to these diffuse forms and thus became *Bayesians of convenience* in the theoretical sense since prior informative was routinely ignored.

Returning to the example of political executives in the US government, Gill and Waterman (2004) used a Gibbs sampler (implemented in the `BUGS` language) to marginalize the model described in (Equation 9), running the chain for 1,000,000 iterations and disposing of the first 500,000 to feel assured that the chain was exploring its stationary distribution during the second period (a standard set of empirical and graphical diagnostics were used as well). The results are summarized in Table 50.1.

Notice in Table 50.1 that the results are given in terms of posterior quantiles only. For a traditional view of these results, readers can look at the posterior median (0.5) and the 95% credible interval ([0.025:0.975]) for each parameter estimated, as shown in the lighter grey columns, but many Bayesians prefer the more detailed descriptive view of the more posterior distributions given here, which incorporates all columns along with associated general probability statements. So, for instance, we can say that with `President Orientation` there is a 97.5% probability that this effect is above 0.2462, even if a 98% credible interval ([0.01:0.99]) covers zero. Bayesians tend not to be fixated with arbitrary thresholds, however. It is not directly

Table 50.1 Posterior quantiles, ordered model for survey of political executives

	0.01	*0.025*	*0.25*	*0.5*	*0.75*	*0.975*	*0.99*
Explanatory Variables:							
Government Experience	–2.0644	–1.8422	–1.0640	–0.6558	–0.2481	0.5332	0.7552
Ideology	–1.2583	–1.1360	–0.7136	–0.4917	–0.2696	0.1544	0.2761
Committee Relationship	0.2345	0.3784	0.8809	1.1446	1.4089	1.9128	2.0546
Career.Exec-Compet	–0.2450	0.0668	1.1570	1.7304	2.3053	3.3996	3.7096
Career.Exec-Liaison/Bur	–4.1108	–3.9079	–3.1929	–2.8188	–2.4435	–1.7309	–1.5265
Career.Exec-Liaison/Cong	–0.1072	0.0708	0.7036	1.0362	1.3680	1.9991	2.1798
Career.Exec-Day2day	–1.6484	–1.5090	–1.0223	–0.7660	–0.5096	–0.0200	0.1182
Career.Exec-Diff	–0.4171	–0.3076	0.0770	0.2780	0.4791	0.8625	0.9725
Confirmation Preparation	–0.0389	0.1277	0.7154	1.0242	1.3333	1.9223	2.0903
Hours/Week	–1.7215	–1.5653	–1.0156	–0.7273	–0.4390	0.1095	0.2660
President Orientation	–0.0712	0.2462	1.3650	1.9504	2.5355	3.6539	3.9720
Threshold Intercepts:							
None—Little	–10.2633	–9.5795	–7.1895	–5.9355	–4.6826	–2.2947	–1.6125
Little—Some	–6.3966	–5.8141	–3.7648	–2.6912	–1.6194	0.4197	0.9985
Some—Significant	–2.3605	–1.8037	0.1625	1.1935	2.2229	4.1847	4.7451
Significant—Extreme	3.5151	4.2837	6.9931	8.4080	9.8269	12.5227	13.2905

in the table but from the sorted MCMC values for this parameter, we can see that there is only a 1.2% probability that this effect is negative. Increased levels of Committee Relationship are reliably associated with increased stress from these results since all of the observed posterior quantiles are positive for this coefficient. This is a somewhat paradoxical finding but may be attributed to a closer relationship with the relevant oversight committee that is not voluntary but mandated by congressional concern over agency policies. We see another reliable but subtle positive relationship between Career.Exec-Compet and stress since the 95% credible interval is bounded away from zero. This may be because underlings challenge political executives' decisions or because, as career public servants, they know the agency and its mission or history better. There are other statistically reliable findings here as well. These statements are very Bayesian in that we are describing regions of the marginal posterior space for this variable in strictly probabilistic terms. There is no notion of confidence or p-values required. In fact, when people misinterpret standard statistical inference, it is often the Bayesian interpretation, such as we have done here, that is desired.

The modern Bayes era was typified with the ease of production of standard models with standard assumptions that were beyond the abilities of Bayesians in the classical era. One after another, complex regression style models – including item response theory (IRT), multinomial probit, complex hierarchical forms, causal specifications, and more – were simply solved. Thus, the 1990s into the 2000s saw article after article in statistics as well as methodological social sciences that used MCMC tools to estimate increasingly intricate models from increasingly complex theories, all from a Bayesian perspective.

POSTMODERN BAYES

It gradually became apparent that the MCMC revolution was about more than just estimating models that had frustrated Bayesians of

the previous generation. It turns out that estimation with Bayesian stochastic simulations (MCMC) provides opportunities to extend modeling and to produce quantities of interest beyond regular posterior inference. So, if the modern Bayesian era is characterized by freedom to estimate pent-up models from 100 years of frustration, then the post-modern Bayesian era is defined by a realization that Bayesian stochastic simulation does not just *allow* estimation (marginalization) of previously inestimable models, it also (almost inadvertently) gives additional inferential information that can be exploited for enhanced purposes. The underlying point is that empirical description, rather than just analytical math-stat description, of posterior parameters of interest gives distributional information that can be used for other purposes such as model checking, model comparison, and enhanced specifications. This section provides a sense of these new tools through several described examples.

Poster Predictive Checks

A very useful way to judge model quality is the posterior predictive check, an approach that compares fabricated data implied by the model fit to the actual outcome data. The general idea is that, if the predictions closely resemble the truth, then the model is fitting well. In addition, deviations from an ideal fit are often informative about where the model could fit better in terms of direction or category. In the most basic form, we are simply comparing (usually plotting) the $y_i,\ldots, y_n$ outcome values from the data with the $\hat{y}_i,\ldots,\hat{y}_n$ values produced from the model. In the case of a linear model, this is incredibly simple in the Bayesian or non-Bayesian context as $\hat{y} = \boldsymbol{X}\hat{\beta}$, but with more complex specifications, it may require more involved calculations. In the non-Bayesian sense, or in the simplified Bayesian sense, we can often analytically calculate $\hat{y}_i$ values and, importantly, measures of uncertainty for these values that allow us to measure or plot accuracy.

As it turns out, it is an ancillary benefit of the MCMC process that it is not only easy to calculate these posterior predictive values for the outcome variable, but it is also almost 'free' to get measures of uncertainty since the empirical distributional descriptions of the estimated model parameters can be made to flow through to the predicted quantities. More specifically, for K explanatory variables (including a constant if appropriate) define $\beta^{(s)} = \beta_{k=1,\ldots,K,s=1,\ldots,n_{sims}}$ for n_{sims} (a large) number of MCMC iterations collected after the assertion of convergence. In a fairly general sense, this lets us produce $\hat{\boldsymbol{y}}^{(s)} = f(\boldsymbol{X}, \boldsymbol{y}, \beta^{(s)})$ values from the Bayesian model specification (including prior distributions), giving n_{sims} values from the full distribution of the poster predicted values.

Returning to the public executives data from Gill and Waterman (2004), subsample n_{sample} = 10,000 values from the total post-convergence MCMC runs to create $\beta^{(s)}$ for the regression parameters and set $\overline{\theta} = (\theta_{r=1},\ldots,\theta_{r=5}$ as the mean of the thresholds parameters across all of the 500,000 saved MCMC runs (although we can also make this component stochastic if desired). The first, and most elementary, summary uses the mean of the data vector $\overline{\boldsymbol{X}}$ to create an archetypal simplified data case (e.g. the mean taken down columns of the K explanatory variable matrix). Define first $\overline{\mu} = \overline{\boldsymbol{X}}\overline{\beta}^{(s)}$ temporarily averaging the simulated coefficient values, then for the ordered logit specification we create the cumulative and marginal outcome probabilities according to

$$
\begin{aligned}
&P_{cumulative}(y \le r) = \left[1 + \exp(-\overline{\theta}_r - \overline{\mu})\right]^{-1}, \quad r = 1,\ldots,5 \\
&P_{marginal}(y = r) = P_{cumulative}(y \le r-1) - P_{cumulative}(y \le r), \quad r = 2,\ldots,4 \\
&P_{marginal}(y = 1) = P_{cumulative}(y = 1) \\
&P_{marginal}(y = 5) = 1 - P_{cumulative}(y \le 4).
\end{aligned}
\tag{10}
$$

Figure 50.2 Categories illustration

This is illustrated in Figure 50.2 showing the outcome values and the threshold values. Thus, we have five ordered categorical probabilities averaged across cases and averaged across simulated estimations given by [0.0001, 0.0019, 0.0865, 0.9040, 0.0075]. This compares somewhat unfavorably to the distribution of actual outcomes in the $n = 512$ size data: [51, 54, 96, 200, 131].

Now suppose we keep the mean vector of the data values but replace the mean of the MCMC parameter draws with the n_{sample} = 10,000. This is done exactly as just described, where (10) is done 10,000 separate times with 10,000 MCMC sub-draws producing a (10,000 × 5) matrix of probability vectors by row. From this we gain the measure of uncertainty on the parameter estimates that is provided by draws from the posterior distributions through the MCMC process. The result of this process is 10,000 draws describing the unconditional probabilities for the complete sample of 512 individuals (averaged) for each of $P_{marginal}(y)\ r = 1,\ldots,5$. These are summarized in Figure 50.3. Notice that the marginal probabilities differ substantially across each category with interesting bunching at extremes in some categories reflecting strong covariate information that flows through these predictions.

This analysis provides useful information, but it is still an incomplete picture because variance across the 512 cases is still suppressed (averaged over). To unlock this added level of uncertainty, we make further use of the MCMC draws using the same procedure as done with the average person case but now expand the data structures to let the

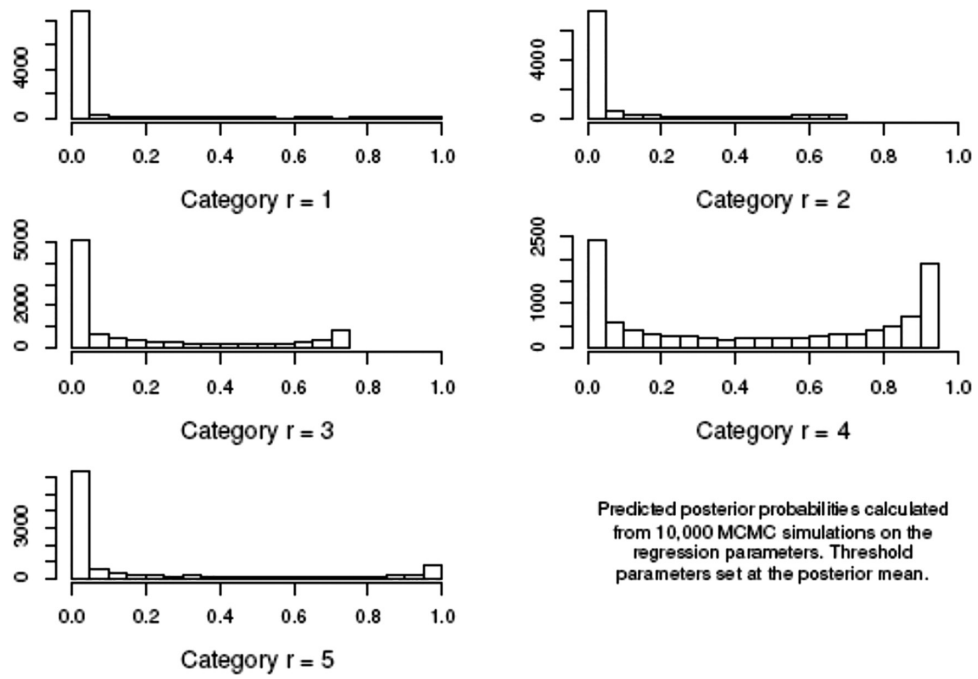

Figure 50.3 Posterior predictive probabilities by ordered outcomes

posterior variability in the estimation process flow down to individual level predictions that include individual level covariate differences. This is literally done in a loop in the code whereby each of the 512 cases is predicted instead of a mean case done at first above. Figure 50.4 shows a random selection of 10,000 of the 500,000 MCMC draws again (more can be done but it crowds the figure), where the resulting $\hat{y}$ posterior predicted value is plotted against the y value for each of the 512 data cases, where each are jittered (the addition of random noise) to slightly separate cases visually for such categorical data. What we see here is a pretty good fit to the data whereby many of the cross-plotted points are on the main diagonal of the plot where y and $\hat{y}$ take on the same values. What we also see is a slight underestimation for cases where the true observed values are in categories 4 and 5. This could not be shown without letting the full uncertainty flow down to the individual predictions, showing the usefulness of the distributional information from the MCMC simulations.

The point of this subsection is that the assessment of model quality through prediction is straightforward (easy actually) with MCMC output because the parameter estimation comes from a large number of draws from the posterior distribution of these parameters. Thus, each of these draws from the underlying Bayesian distributions of the model can be 'flowed' through to quantities of interest like outcome predictions with the uncertainty preserved through the empirical draws. Therefore, the MCMC process actually makes this process easier since complex mathematical-statistics analytical calculations are completely unneeded now, including lengthy derivations, transformations, use of the delta method, and more.

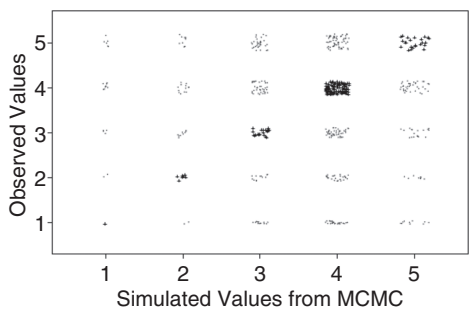

Figure 50.4 Posterior predictive probabilities by outcome categories

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (or analogously Hybrid Monte Carlo; henceforth HMC) is a modification of the standard Metropolis-Hastings algorithm that uses 'physical system dynamics' as a means of generating candidate values for a Metropolis decision to move to a new point in the state space. The Metropolis-Hastings algorithm (Metropolis et al., 1953) is a Markovian accept-reject procedure that moves through the multidimensional space of interest. The basic Metropolis-Hastings algorithm for a single selected parameter vector starts with a J-length parameter vector, $\theta \in \Theta^J$, to empirically describe target distribution of interest, $\pi(\theta)$, by 'visiting' substates proportionally to the density of interest. At the t^{th} step of the Markov chain ('t' stands for time), the chain is at the position indicated by the vector $\theta^{[t]}$. We then draw θ' from a distribution over the same support, from a *proposal distribution* denoted $q_t(\theta'|\theta)$. This function must be 'reversible', meaning that $\pi(\theta)p(\theta,\theta') = \pi(\theta')p(\theta',\theta) \forall \theta,\theta'$ in order to pick from a wide range of non-symmetric forms, where $p(\cdot)$ is the *actual transaction function* – the probability of generating a candidate *and* accepting it – and $\pi(\cdot)$ is the target density. We then decide to move with probability

$$\min\left[\frac{\pi(\theta')}{\pi(\theta)}\frac{q(\theta|\theta')}{q(\theta'|\theta)},1\right]. \qquad (11)$$

Importantly, rejecting θ' means moving to θ (the current position) as the new position

in the time series of samples. An important feature of the algorithm is the flexibility of the choice of the proposal distribution. Many variations are based on strategic choices for this distribution. For example, the Hit and Run algorithm (Chen and Schmeiser, 1993) separates the direction and distance decision in the proposal so that it can be optimized for highly constrained posterior structures and the algorithm does not reject a large number of candidate destinations.

The HMC algorithm exploits flexibility in the choice of candidate distribution by incorporating information about posterior topography to traverse the sample space more efficiently. Topography in this sense means the curvature of the target distribution, which is easy for humans to visualize in three dimensions and impossible for humans to see in the 8–12 dimensions that political science models often specify. The basic idea predates the modern advent of MCMC from Duane et al. (1987) and was developed in detail by Neal (1993, 2011).

Like the original Metropolis (1953) algorithm, Hamiltonian Monte Carlo comes from physics (Meyer, Hall and Offin 1992). Here we are concerned with an object's trajectory within a specified multidimensional system as a way to describe joint posterior distributions. Now define ϑ_t as a k-dimensional location vector and $\boldsymbol{p}_t$ as a k-dimensional momentum (e.g. mass times velocity) vector, both recorded at time t. The Hamiltonian system at time t with $2k$ dimensions is given by the joint Hamiltonian function

$$H(\vartheta_t, \boldsymbol{p}_t) = U(\vartheta_t) + K(\boldsymbol{p}_t), \qquad (12)$$

where $U(\vartheta_t)$ is the function describing the *potential energy* at the point ϑ_t, and $K(\boldsymbol{p}_t)$ is the function describing the *kinetic energy* for momentum $\boldsymbol{p}_t$. Neal (2011) gives the simple 1-dimensional example

$$U(\vartheta_t) = \frac{\vartheta_t^2}{2} \qquad K(p_t) = \frac{p_t^2}{2}, \qquad (13)$$

which is equivalent to a standard normal distribution for ϑ. Commonly, the kinetic energy function is defined as

$$K(\boldsymbol{p}_t) = \boldsymbol{p}_t' \Sigma^{-1} \boldsymbol{p}_t, \qquad (14)$$

where Σ is a symmetric and positive-definite matrix that can be as simple as an identity matrix times some scalar that can serve the role of a variance: $\Sigma = \sigma^2 \boldsymbol{I}$. This simple form is equivalent to the log PDF of the multivariate normal with mean vector zero and variance-covariance matrix Σ.

Hamiltonian dynamics describe the gradient-based way that potential energy changes to kinetic energy and kinetic energy changes to potential energy as the object moves over time throughout the system (multiple objects require equations for gravity, but that is fortunately not our concern here). The mechanics of this process are given by Hamilton's equations, which are the set of simple differential equations

$$\begin{aligned} \frac{\partial \vartheta_{it}}{\partial t} &= \frac{\partial H}{\partial \boldsymbol{p}_{it}} = \frac{K(\partial \boldsymbol{p}_{it})}{\partial \boldsymbol{p}_{it})} \\ \frac{\partial \boldsymbol{p}_{it}}{\partial t} &= -\frac{\partial H}{\partial \vartheta_{it}} = -\frac{U(\partial \vartheta_{it})}{\partial \vartheta_{it})} \end{aligned} \qquad (15)$$

for dimension i at time t. For continuously measured times these equations give a mapping from time t to time $t + \tau$, meaning that from some position ϑ_t and momentum $\boldsymbol{p}_t$ at time t we can predict ϑ_t and $\boldsymbol{p}_\tau$. Returning to the one-dimensional standard normal case, these equations are simply $d\vartheta_t/dt = p$ and $dp/dt = -\vartheta$.

There are three important properties of Hamiltonian dynamics that are actually *required* if we are going to use them to construct an MCMC algorithm (Neal, 2011). First, Hamiltonian dynamics is *reversible*, meaning that the mapping from $(\vartheta_t, \boldsymbol{p}_t)$ to $(\vartheta_{t+\tau}, \boldsymbol{p}_{t+\tau})$ is one-to-one and therefore also defines the reverse mapping from $(\vartheta_{t+\tau}, \boldsymbol{p}_{t+\tau})$ to $(\vartheta_t, \boldsymbol{p}_t)$. Second, *total* energy is conserved over time t and dimension k, and the Hamiltonian is invariant, as shown by

$$\begin{aligned}\frac{\partial H}{\partial t} &= \sum_{i=1}^{k}\left[\frac{\partial \vartheta_i}{\partial t}\frac{\partial H}{\partial \vartheta_i}+\frac{\partial \boldsymbol{p}_i}{\partial t}\frac{\partial H}{\partial \boldsymbol{p}_i}\right] \\ &= \sum_{i=1}^{k}\left[\frac{\partial H}{\partial \boldsymbol{p}_i}\frac{\partial H}{\partial \vartheta_i}-\frac{\partial H}{\partial \vartheta_i}\frac{\partial H}{\partial \boldsymbol{p}_i}\right]=0.\end{aligned} \tag{16}$$

This provides detailed balance (reversibility) for the MCMC algorithm. Second, Hamiltonian dynamics preserve volume in the $2k$ dimensional space. In other words, elongating some region in a direction requires withdrawing another region as the process continues over time. This ensures that there is no change in the scale of Metropolis-Hastings acceptance probability. Finally, Hamiltonian dynamics provides a *symplectic mapping* in $\mathcal{R}^{2k}$ space. Define first the smooth mapping $\psi : \mathcal{R}^{2k} \rightarrow \mathcal{R}^{2k}$ with respect to some constant and invertible matrix $\boldsymbol{J}$ with $\boldsymbol{J}' = -\boldsymbol{J}$ and $\det(\boldsymbol{J}) \neq 0$, along with having Jacobian $\psi(z)$ for some $z \in \mathcal{R}^{2k}$. The mapping ψ is symplectic if

$$\psi(z)'\boldsymbol{J}^{-1}\psi(z) = \boldsymbol{J}^{-1}. \tag{17}$$

Leimkuhler and Reich (2005, p. 53) give the following mapping in 2-dimensional space $z = (\vartheta, p)$:

$$\psi(\vartheta, p) = \begin{bmatrix} p \\ 1 + b\vartheta + ap^2 \end{bmatrix}, \tag{18}$$

with constants $a, b \neq 0$. The Jacobian of $\psi(\vartheta, p)$ is calculated by

$$\frac{\partial}{\partial \vartheta}\frac{\partial}{\partial p}\psi(\vartheta, p) = \begin{bmatrix} 0 & 1 \\ b & 2ap \end{bmatrix}. \tag{19}$$

We check symplecticness by

$$\begin{aligned}&\left[\frac{\partial}{\partial \vartheta}\frac{\partial}{\partial p}\psi(\vartheta, p)\right]'\boldsymbol{J}^{-1}\left[\frac{\partial}{\partial \vartheta}\frac{\partial}{\partial p}\psi(\vartheta, p)\right] \\ &= \begin{bmatrix} 0 & 1 \\ b & 2ap \end{bmatrix}'\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ b & 2ap \end{bmatrix} \\ &= -b\boldsymbol{J}^{-1}.\end{aligned} \tag{20}$$

Thus we say that $\psi(\vartheta, p)$ is symplectic for $b = -1$ and any $a \neq 0$.

Everything discussed so far assumed continuous time, but obviously for a computer implementation in a Markov chain Monte Carlo context, we need to discretize time. Thus, we will grid $t + \tau$ time into intervals of size υ: υ, 2υ, 3υ, ..., $m\upsilon$. We need a way to obtain this discretization while preserving volume, and so we use a tool called the *leapfrog methods*. The notation is more clear if we now move t from the subscript to functional notation: $\vartheta(t)$ and $\boldsymbol{p}(t)$, which is also a reminder that time is now discrete rather than continuous. To complete a single step starting at time t, first update each of the momentum dimensions by $\upsilon/2$ with the following:

$$\boldsymbol{p}_i\left(t + \frac{\upsilon}{2}\right) = p_i(t) - \frac{\upsilon}{2}\frac{\partial U(\vartheta_t)}{\partial \vartheta_i(t)}. \tag{21}$$

Now take a full υ-length step to update each of the position dimensions to leapfrog over the momentum:

$$\vartheta_i(t + \upsilon) = \vartheta_i(t) + \upsilon\frac{\partial K(\boldsymbol{p}_t)}{\partial \boldsymbol{p}_i(t + \frac{\upsilon}{2})}. \tag{22}$$

Then finish with the momentum catching up in time with the step:

$$\boldsymbol{p}(t + \upsilon) = \boldsymbol{p}_i\left(t + \frac{\upsilon}{2}\right) - \frac{\upsilon}{2}\frac{U(\vartheta_t)}{\partial \vartheta_i(t + \upsilon)}. \tag{23}$$

Notice that the leapfrog method is reversible since it is a one-to-one mapping from t to $t + \upsilon$. Obviously, running these steps M times completes the Hamiltonian dynamics for $M \times \upsilon$ periods of total time. The determination of υ is a key tuning parameter in the algorithm since smaller values give a closer estimation to continuous time but also add more steps to the algorithm.

A Metropolis-Hastings algorithm is configured such that the Hamiltonian function serves as the candidate-generating distribution. This requires connecting the

regular posterior density function, $\pi(\theta)$, to a potential energy function, $U(\vartheta_t)$, where a kinetic energy function, $K(\boldsymbol{p}_t)$, serves as a (multidimensional and necessary) auxiliary variable. This connection is done via the *canonical distribution* commonly used in physics,

$$p(x) = \frac{1}{Z}\exp\left[-\frac{E(x)}{T}\right], \tag{24}$$

where $E(x)$ is the energy function of some system at state x, T is the temperature of the system (which can simply be set at 1), and Z is just a normalizing constant so that $p(x)$ is a regular density function. In the Hamiltonian context (Equation 24) is

$$\begin{aligned} p(\vartheta, \boldsymbol{p}) &= \frac{1}{Z}\exp\left[-\frac{H(\vartheta, \boldsymbol{p})}{T}\right] \\ &= \frac{1}{Z}\exp\left[-\frac{U(\vartheta_t) + K(\boldsymbol{p}_t)}{T}\right] \\ &= \frac{1}{Z}\exp\left[-\frac{U(\vartheta_t)}{T}\right]\exp\left[-\frac{K(\boldsymbol{p}_t)}{T}\right], \end{aligned} \tag{25}$$

demonstrating that ϑ and $\boldsymbol{p}$ are independent. Finally, we connect the energy function metric with the regular posterior density metric with the function

$$E(\vartheta) = -\log(\pi(\theta)), \tag{26}$$

thus completing the connection. Notice that the θ variables must all be continuous in the model, although Hamiltonian Monte Carlo can be combined with other MCMC strategies in a hybrid algorithm.

The Hamiltonian Monte Carlo algorithm uses two general steps at time t:

- Generate, independent of the current ϑ_t, the momentum $\boldsymbol{p}_t$ from the multivariate normal distribution implied by $K(\boldsymbol{p}_t) = \boldsymbol{p}_t' \Sigma^{-1} \boldsymbol{p}_t$ with mean vector zero and variance-covariance matrix $\sigma^2\boldsymbol{I}$ (or some other desired symmetric and positive-definite form).
- Run the leapfrog method M times with υ steps to produce the candidate $(\tilde{\vartheta}, \tilde{\boldsymbol{p}})$.
- Accept this new location or accept the current location as the $t + 1$ step with a standard Metropolis decision using the H function

$$\min\left[1, \exp(-H((\tilde{\vartheta}, \tilde{\boldsymbol{p}})) + H(\vartheta, \boldsymbol{p})\right]. \tag{27}$$

While this process looks simple, there are several complications to consider. We must be able to take the partial derivatives of the log-posterior distribution, which might be hard. The chosen values of the leapfrog parameters, M and υ, are also critical. If υ is too small then exploration of the posterior density will be very gradual with small steps, and if υ is too big then many candidates will be rejected. Choosing M is important because this parameter allows the Hamiltonian process to explore strategically with respect to gradients. Excessively large values of M increase compute time, but excessively small values of M lead to many rejected candidates. In both cases where the parameters are too small, we lose the advantages of the gradient calculations and produce an inefficient random walk. Finally, σ^2 affects efficiency of the algorithm in the conventional sense of appropriating tuning the variance of the multivariate normal for the momentum. These can be difficult parameter decisions and Neal (2011) gives specific guidance on trial runs and analysis of the results. The Hamiltonian Monte Carlo dynamics are difficult to implement in complex multilevel generalized linear models that aim to apply full Bayesian inference. While these models can be carried out with BUGS or JAGS, this takes an enormous amount of time and computational resources. To circumvent this, a group of academics, among them Andrew Gelman and Bob Carpenter, developed STAN. `STAN` is written in C++ and, unlike `BUGS` and `JAGS`, employs reverse-mode algorithmic differentiation to implement HMC in a much faster way. It supports a range of functions (e.g. probability functions, log gamma, inverse logit etc.) and integrates matrix operations on linear algebra. See `https://mc-stan.org/` for details and downloads.

Bayes Factor Calculations

Another example where MCMC output makes mathematical calculations much easier is the calculation of the Bayes Factor for nonlinear regression models. This also follows the principle that MCMC simulation provides a natural distributional summary that can be used for multiple purposes besides the original purpose of simply producing marginal distributions from a complicated joint distribution from the Bayesian model specification. In the classical era, it was recognized that Bayes Factors were an extremely useful model assessment and comparison tool going back to Jeffreys (1983), but they were often very difficult to calculate for realistic regression-style models.

Bayes Factors start with observed data $\boldsymbol{x}$ for testing two models, with associated parameter vectors θ_1 and θ_2: M_1: $f_1(\boldsymbol{x} \mid \theta_1)$ M_2: $f_2(\boldsymbol{x} \mid \theta_2)$. Here these parameter vectors can define nested or non-nested alternatives, unlike the more simple likelihood ratio test. With prior distributions, $p_1(\theta_1)$ and $p_2(\theta_2)$, and prior probabilities on the two models, $p(M_1)$ $p(M_2)$, we can produce the odds ratio for Model 1 versus Model 2 by Bayes' Law:

$$\underbrace{\frac{\pi(M_1 \mid \boldsymbol{x})}{\pi(M_2 \mid \boldsymbol{x})}}_{\text{posterior odds}} = \underbrace{\frac{p(M_1) \,/\, p(\boldsymbol{x})}{p(M_2) \,/\, p(\boldsymbol{x})}}_{\text{prior odds/data}} \times \underbrace{\frac{\int_{\theta_1} f_1(\boldsymbol{x} \mid \theta_1) p_1(\theta_1) d\theta_1}{\int_{\theta_2} f_2(\boldsymbol{x} \mid \theta_2) p_2(\theta_2) d\theta_2}}_{\text{Bayes Factor}}. \tag{28}$$

So, we are actually interested in the ratio of marginal likelihoods – equation (5) from the two models. By canceling and algebraically rearranging, we get the common form of the Bayes Factor:

$$BF_{(1,2)} = \frac{\pi(M_1 \mid \boldsymbol{x}) \,/\, p(M_1)}{\pi(M_2 \mid \boldsymbol{x}) \,/\, p(M_2)} \tag{29}$$

(Gill, 2014). As suggested by these forms, analytical calculation for reasonably realistic social science regression models can be challenging. Fortunately, this is direct and easy for most Bayesian generalized linear models estimated with MCMC. Chib (1995) and Chib and Jeliazkov (2001), for instance, give a handy and generalizable recipe in the context of probit regression models. To begin, rearrange equation (4) and take logs for a single model (for the moment) using the log-likelihood:

$$\log p(\boldsymbol{x}) = \ell(\theta' \mid \boldsymbol{x}) + \log p(\theta') - \log \pi(\theta' \mid \boldsymbol{x}). \tag{30}$$

Here θ' is a completely arbitrary point in the appropriate sample space, such as a point in the high density region, for instance the posterior mean. To start, we use $\pi(\theta' \mid \boldsymbol{x})$ from simulation for a generic MCMC estimation approach (details in Chapter 14 of Gill, 2014). Define the probability of the Metropolis-Hastings Markov chain as transitioning to an arbitrary point θ' from a starting point θ with the candidate-generating distribution that produces θ' times the probability that it is accepted from above:

$$\alpha(\theta', \theta) = \min\left[\frac{\pi(\theta')}{\pi(\theta)} \frac{q_t(\theta \mid \theta')}{q_t(\theta' \mid \theta)}, 1\right]. \tag{31}$$

This candidate-generating distribution produces θ' times the probability that it is accepted from above: $p(\theta, \theta') = q(\theta' \mid \theta)\alpha(\theta', \theta)$, such that for any arbitrary point detailed balance is preserved, $\pi(\theta) q(\theta' \mid \theta) \alpha(\theta', \theta) = \pi(\theta') q(\theta \mid \theta') \alpha(\theta, \theta')$. Now take integrals of both sides with respect to θ, realizing that $\pi(\theta')$ is a function evaluation at an arbitrary point and can therefore be moved outside of

the integration process, and, with some algebra, reach

$$\pi(\theta') = \frac{\int \Theta \pi(\theta) q(\theta' \mid \theta) \alpha(\theta', \theta) d\theta}{\int_{\Theta} q(\theta \mid \theta') \alpha(\theta, \theta') d\theta} = \frac{E_{\pi(\theta)}[q(\theta' \mid \theta)\alpha(\theta', \theta)]}{E_{q(\theta \mid \theta')}[\alpha(\theta, \theta')]}, \quad (32)$$

which is simply the ratio of two expected value calculations: Chib and Jeliazkov (2001) observed that in the course of running an MCMC estimation process for marginal posterior distributions, we can get the marginal likelihood without extra trouble replacing (Equation 32) with its simulation version,

$$\pi_{sim}(\theta') = \frac{\frac{1}{M}\sum_{m=1}^{M} \alpha(\theta', \theta_m) q(\theta' \mid \theta_m)}{\frac{1}{N}\sum_{n=1}^{N} \alpha(\theta_N, \theta')}, \quad (33)$$

which uses known quantities readily at hand for some number of simulations M. Here θ' is chosen arbitrarily but within a high-density region of the posterior distribution. Therefore, this process substitutes a challenging integration process with simulation of the posterior density at a single point by completing (Equation 30) with the simulated result

$$\log p_{sim}(\boldsymbol{x}) = \ell(\theta' \mid \boldsymbol{x}) + \log p(\theta') - \log \pi_{sim}(\theta' \mid \boldsymbol{x}), \quad (34)$$

where all of these quantities are easily available.

Bayesian Nonparametrics

We are concerned in this section with how *nonparametric priors* can enhance the increasing use of Bayesian models for the presence of unobserved heterogeneity, which is a common problem across the social sciences. Researchers commonly deal with the problem by specifying non-unique random effect terms ψ_j, $j \in J < n$ to capture grouping or clustering information where the mapping from $i = 1, \ldots, n$ to $j = 1, \ldots, J$ is known.

Generally, the distribution of the ψ_i is unknown but safely assumed through custom or testing. The normal distribution is a useful default for both practical and asymptotic reasons. This convenience cannot be directly tested through residuals analysis but affects overall model fit, which can be tested. In the Bayesian setting, a better and more flexible alternative exists as a so-called nonparametric Bayesian prior in which ψ_j is drawn from a vastly more flexible distributional setup, starting with

$$(Y_1, \ldots, Y_n) \sim f(y_1, \ldots, y_n \mid \beta, \psi_1, \ldots, \psi_n) = \prod_i f(y_i \mid \beta, \psi_i), \quad \psi_i \sim G, \quad i = 1, \ldots, n, \quad (35)$$

where f can be taken as normal and where a popular choice for G is the Dirichlet Process ($\mathcal{DP}$),

$$\psi_i \sim G \sim \mathcal{DP}(\lambda, \phi_0), \quad i = 1, \ldots, n, \quad (36)$$

with base measure ϕ_0 and precision parameter λ. In particular, the observations are modeled as

$$Y_i = \boldsymbol{X}_i \beta + \psi_i + \epsilon_i, \quad (37)$$

where the ϵ_i are treated as independent normal random variables and ψ_i indicates the random effects assignment for the ith case. Alternatively, specification of a link function turns this into a generalized linear model (GLM) in the classic sense, $\hat{Y}_i = g^{-1}(\boldsymbol{X}_i \beta + \psi_i)$. Since the ψ_i are drawn from a $\mathcal{DP}$ distribution, they are not necessarily unique and thus can be represented by a K-vector, η, where $K < n$. Furthermore, the model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{A}\eta + \epsilon, \quad (38)$$

where $\psi = A\eta$ and A is an $n \times K$ matrix of zeros with a single one in each row which denotes the specific η_k assigned to ψ_i (Kyung et al., 2010).

Dirichlet process mixture models were originally formulated by Ferguson (1973), who defined the underlying process and derived the key properties. Blackwell and MacQueen (1973) then showed that the marginal distribution for the Dirichlet process can be treated as that of the n^{th} step of a Polya urn process. Other key theoretical work includes Korwar and Hollander (1973) and Sethuraman (1994). The contributions that have particular importance for the likelihood function development are that of Lo (1984), who derives the analytic form of a Bayesian density estimator, and Liu (1996), who derives an identity for the profile likelihood estimator of m. This is an interesting sociology of science story in that these works mostly predated the computational tools that made the models possible for real datasets. Follow-on work that changed this state are typified by Escobar and West (1995), MacEachern and Müller (1998), and Jain and Neal (2004).

The model specified in (35) is actually a classical semiparametric random effects model and, with further Bayesian modeling of the parameters, can be implemented with MCMC. Unfortunately, the presence of the Dirichlet term makes the use of the standard Gibbs sampler somewhat complicated in non-conjugate situations such as with is the model that was developed in Gill and Casella (2009). These authors find that this approach can model difficult data and produce results that existing alternative methods fail to discover. They then account for unobserved, important clustering structures with the nonparametric process that do not necessarily reflect intervening or confounding variables but still provide information about agency environment that was not explicitly available.

Gill and Casella (2009) introduced a GLMDM with an ordered probit link to model political science data, specifically modeling the stress, from the Gill and Waterman (2004) data already described here. Their Dirichlet precision parameter was not an influential model parameter and was therefore fixed at a value that made the MCMC sampler more efficient. Kyung et al. (2010, 2012), looked at the maximum likelihood estimation of the precision parameter and found that the standard approach to finding the maximum likelihood estimate, given in Liu (1996), could yield a maximum, a minimum, or even a ridge. Figure 50.5, from Kyung et al., (2010), shows some observed shapes of this likelihood function for simulated circumstances. Since likelihood estimation is not reliable for this parameter, Kyung, et al., (2010) proved that introducing a prior distribution on the precision parameter guarantees an interior mode and so stabilizes the estimation procedure.

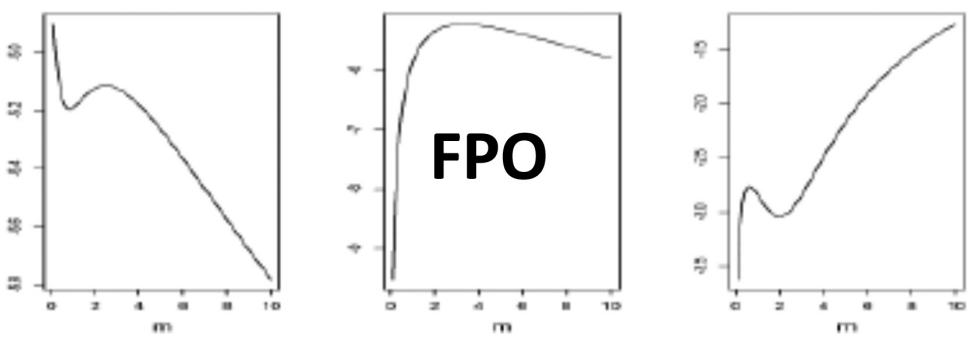

Figure 50.5 Log-likelihood functions for configurations of component likelihoods

Models with Dirichlet process priors are treated as hierarchical models (Gill and Womack, 2013) in a Bayesian framework, and the implementation of these models through Bayesian computation and efficient algorithms has had substantial progress. Escobar and West (1995) produced a Gibbs sampling algorithm for the estimation of posterior distribution for all model parameters and the direct evaluation of predictive distributions. MacEachern and Müller (1998) developed a Gibbs sampler with non-conjugate priors by using auxiliary parameters, and Neal (2000) provided an extended and more efficient Gibbs sampler to handle general Dirichlet process mixture models with non-conjugate priors by using a set of auxiliary parameters. Teh et al. (2006) extended the method of Escobar and West for posterior sampling of the precision parameter with a gamma distributed prior. Kyung, Gill and Casella (2011) extended these results to a generalized Dirichlet process mixed model with a probit link function. They derived a Gibbs sampler for the model parameters and the important subclusters of the Dirichlet process using new parameterization of the hierarchical model to derive a Gibbs sampler that more fully uses the structure of the model.

Again, $\mathbf{X}_i$ are the covariates associated with the i^{th} observation, β be the coefficient vector, and ψ_i be the random effect accounting for subject-specific deviation from the underlying model. Assume that $Y_i \mid \psi$ are conditionally independent, each with a density from the exponential family, where $\psi = (\psi_1, \ldots, \psi_J)$, $J < n$. Based on the notation on McCulloch et al. (2008), the Generalized Linear Mixed Dirichlet Model is expressed as

$$\begin{aligned} Y_i \mid \psi &\overset{ind}{\sim} f_{Y_i \mid \psi}(y_i \mid \psi), \quad i = 1, \ldots, n \\ f_{Y_i \mid \psi}(y_i \mid \psi) &= \exp\left[\left\{y_i \gamma_i - b(\gamma_i)\right\} / \xi^2 - c(y_i, \xi)\right], \end{aligned} \tag{39}$$

where y_i is assumed discrete valued. It is assumed that $[Y_i \mid \psi] = \mu_i = \partial b(\gamma_i) / \partial \gamma_i$. Using the arbitrary link function $g(\cdot)$, we can express the transformed mean of Y_i, $\mathrm{E}[Y_i \mid \psi]$, as a linear function, $g(\mu_i) = \boldsymbol{X}\beta + \psi_i$. For the Dirichlet process mixture models, we assume that

$$\begin{aligned} \psi_i &\sim G \\ G &\sim \mathcal{DP}(mG_0), \end{aligned} \tag{40}$$

where $\mathcal{DP}$ is the Dirichlet Process with base measure G_0 and precision parameter m. Blackwell and MacQueen (1973) proved that for $\psi_1, \ldots, \psi_n$ iid from $G \sim \mathcal{DP}$, the joint distribution of ψ is a product of successive conditional distributions of the mixture form

$$\begin{aligned} \psi_i \mid \psi_1, \ldots, \psi_{i-1}, m \sim{}& \frac{m}{i - 1 + m} g_0(\psi_i) \\ &+ \frac{1}{i - 1 + m} \sum_{l=1}^{i-1} \delta(\psi_l = \psi_i) \end{aligned} \tag{41}$$

where $\delta(\cdot)$ denotes the Dirac delta function and $g_0(\cdot)$ is the density function of base measure. The likelihood function from Liu (1996) and Lo (1984) is produced by integrating over the random effects,

$$L(\theta \mid \boldsymbol{y}) = \frac{\Gamma(m)}{\Gamma(m+n)} \sum_{k=1}^{n} m^k \sum_{C:|C|=k} \prod_{j=1}^{k} \Gamma(n_j) \int f(\boldsymbol{y}_{(j)} \mid \theta, \psi_j) dG_0(\psi_j), \tag{42}$$

where C defines the subclusters (not actual clusters in the social science since there is no penalty here for over-fitting in the algorithm), $\boldsymbol{y}_{(j)}$ is the vector of y_is that are in subcluster j, and ψ_j is the common parameter for that subcluster. There are $\mathcal{S}_{n,k}$ different subclusters C, the Stirling Number of the Second Kind (Abramowitz and Stegun, 1972: 824–825). Now we consider again the $n \times k$ matrix $\boldsymbol{A}$ defined by

$$\boldsymbol{A} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix},$$

where a_i is a $1 \times k$ vector of all zeros except for a 1 in one position that indicates which group the observation is from. So, each column of matrix $\boldsymbol{A}$ represents a partition of the sample of size n into k groups. If the subcluster C is partitioned into groups $\{S_1, \ldots, S_k\}$, then if $i \in S_j$, $\psi_i = \eta_j$ and the random effect can be rewritten as $\psi = \boldsymbol{A}\eta$, where $\eta = (\eta_1, \ldots, \eta_k)$ and $\eta_j \overset{iid}{\sim} G_0$ for $j = 1, \ldots, k$.

The results of implementing the GLMDM model for the data in Gill and Waterman (2004) are given in Table 50.2 as posterior quantiles. Notice first that these results differ markedly from the previous analysis of these data with a conventional Bayesian ordered choice model as summarized in Table 50.1. The nonparametric specific is a fundamentally different approach that includes and leverages underlying heterogeneity by accounting for subclusters in the estimation process. For instance, the effect of the variable `Committee Relationship` and stress is reliably in the opposite direction: closer ties to the oversight committee are associated with lower stress levels, when accounting for group level latent heterogeneity. This actually makes sense when considering the wide range of relationship types, policy spaces, and administrative histories that exist between congressional committees and administrative agencies. So, the Dirichlet process that accounts for such underlying grouping reveals a different type of relationship effect. Moreover, the previously seen, positive relationship between

Table 50.2 Posterior quantiles, GLMDM for survey of political executives

	0.01	*0.025*	*0.25*	*0.5*	*0.75*	*0.975*	*0.99*
Explanatory Variables:							
`Government Experience`	−0.1071	−0.0861	−0.0117	0.0275	0.0665	0.1409	0.1623
`Ideology`	−0.0421	−0.0309	0.0077	0.0280	0.0483	0.0870	0.0980
`Committee Relationship`	−0.3146	−0.3021	−0.2581	−0.2350	−0.2119	−0.1679	−0.1554
`Career.Exec-Compet`	−0.3600	−0.3431	−0.2823	−0.2505	−0.2186	−0.1579	−0.1404
`Career.Exec-Liaison/Bur`	−0.0371	−0.0240	0.0226	0.0470	0.0714	0.1181	0.1313
`Career.Exec-Liaison/Cong`	−0.1438	−0.1299	−0.0811	−0.0556	−0.0299	0.0191	0.0330
`Career.Exec-Day2day`	−0.3195	−0.3041	−0.2499	−0.2215	−0.1931	−0.1391	−0.1236
`Career.Exec-Diff`	−0.0383	−0.0241	0.0262	0.0525	0.0787	0.1288	0.1431
`Confirmation Preparation`	−0.6267	−0.5978	−0.4955	−0.4419	−0.3883	−0.2859	−0.2568
`Hours/Week`	0.3411	0.3509	0.3858	0.4040	0.4222	0.4571	0.4669
`President Orientation`	−0.6502	−0.6210	−0.5188	−0.4653	−0.4116	−0.3090	−0.2798
Threshold Intercepts:							
`None—Little`	−1.9915	−1.9580	−1.8402	−1.7782	−1.7160	−1.5979	−1.5644
`Little—Some`	−1.4407	−1.4096	−1.3010	−1.2439	−1.1866	−1.0778	−1.0466
`Some—Significant`	−0.9007	−0.7847	−0.3788	−0.1660	0.0473	0.4541	0.5718
`Significant—Extreme`	0.3811	0.4108	0.5157	0.5705	0.6254	0.7303	0.7598

`Career.Exec-Compet` and stress is now overturned: there is a reliably negative finding here likely recognizing agency heterogeneity in senior staffing. The hours per week worked have a positive relationship with stress. This statistically reliable finding is shown in the tightly bounded and positive quantiles for `Hours/Week`. Interestingly, political executives who required preparation for the hearings on their Senate confirmation later provided lower stress scores. This reliable finding is likely related to the committee relationship variable.

CONCLUSION

The purpose of this chapter is to introduce the Bayesian inferential process with a focus on its implementation in political science and international relations. It is designed to highlight the practical history of how results are obtained in Bayesian analysis over the course of time. The hope is that readers will see both the principled theoretical advantages of thinking 'Bayesianly' and the practical ease with which results can today be produced and extended.

Bayesian inference is characterized by the explicit use of probability for describing uncertainty, which means probability models (likelihood functions) for data given parameters and probability distributions (PDFs and PMFs) for parameters. From this basis, inference proceeds from inference for unknown values conditioned on observed data with the use of inverse probability with Bayes' Law to describe the full distribution of the unknown quantities with this update. Probability statements lie at the heart of Bayesian analysis. Everything a Bayesian does not know for a fact is modeled with probability distributions. At the core setup of Bayesian analysis, prior knowledge informs a specified probability model, which is then updated by conditioning on observed data and whose fit to the data is evaluated distributionally. The Bayesian paradigm fits closely with the core tenets of scientific discovery: current theories form the basis of stated prior information and informative evidence from data collection have the ability to update our theories. Contrary to traditional statistical thinking, however, it constitutes a different way of thinking about uncertainty that is strictly probability based, eschewing assumptions like an unending stream of independent and identically distributed data.

In the sections above, we identified three clear historical eras in the development of Bayesian methods: classical, modern, and postmodern. The classical era lasted until 1990 and was characterized by a determination that philosophical correctness should be recognized but was tempered with the challenges of estimating models with realistically large and complicated social science data. Unfortunately at the time, it was not hard to create logical and mathematical arguments that showed the superiority of Bayesian inference over more traditional methods, but it was very hard, if not impossible, to apply these arguments empirically.

Gelfand and Smith changed this state of the Bayesian world and ushered in the modern era in 1990. They discovered Gibbs sampling, a tool with its roots in engineering and image restoration. Aided by improvements and availability in computing power, Gibbs sampling replaced analytical derivation of marginals from a joint with large numbers of draws sampled by the computer. Together with the Metropolis-Hastings algorithm unearthed from statistical physics, Gibbs sampling solved Bayesians' problems and became known collectively as Markov chain Monte Carlo. Thanks to MCMC, models were no longer too complex for marginalization of joint posteriors to create regression tables. MCMC revolutionized Bayesian inference, released decades of frustration, and led to countless Bayesian applications and publications.

Finally, the postmodern era began in the early 21st century when researchers realized

the full potential of MCMC beyond the estimation of previously inestimable models. It gradually became apparent that Bayesian stochastic simulation could also be exploited for enhanced purposes, such as model checking and model comparison. The consequence of this realization was a number of tools designed to extend the reach of MCMC, such as poster predictive checks, Hamiltonian Monte Carlo, Bayes Factor calculations, and Bayesian nonparametrics. As a result, it is now possible to not only easily produce Bayesian results, but also to extend the Bayesian paradigm and its application far beyond model estimation alone. This makes the Bayesian inferential process extraordinarily useful in political science and international relations.

Note

1 Our thanks to the methodology reading group at American University: Le Bao, Ryan DeTamble, Michael Heseltine, Daisy Muibu, Abhishek Regmi, Samantha Senn, Rui Wang, Kumail Wasif, Morten Wendelbo.

REFERENCES

Abramowitz, Milton and Stegun, Irene A. (eds.) (1972) *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Mineola: Dover Publications.

Bayes, Thomes (1763) An Essay Towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society of London*, 53: 370–418.

Bernardo, José M. (1979) Reference Posterior Distributions for Bayesian Inference (with Discussion), *Journal of the Royal Statistical Society, Series B (Methodological)*, 41(2): 113–147.

Birnbaum, Allan (1962) On the Foundations of Statistical Inference, *Journal of the American Statistical Association*, 57(298): 269–306.

Blackwell, David and MacQueen, James B. (1973) Ferguson Distributions via Polya Urn Schemes, *The Annals of Statistics*, 1(2): 353–355.

Box, George E.P. and Tiao, George C. (1973) *Bayesian Inference in Statistical Analysis*. New York: John Wiley & Sons.

Brooks, Stephen P., Dellaportas, Petros and Roberts, Gareth O. (1997) An Approach to Diagnosing Total Variation Convergence of MCMC Algorithms, *Journal of Computational and Graphical Statistics*, 6(3): 251–265.

Brooks, Stephen P. and Roberts, Gareth O. (1999) On Quantile Estimation and Markov Chain Monte Carlo Convergence, *Biometrika*, 86(3): 710–717.

Chen, Ming-Hui and Schmeiser, Bruce (1993) Performance of the Gibbs, Hit-and-Run and Metropolis Samplers, *Journal of Computational and Graphical Statistics*, 2(3): 251–272.

Chib, Siddhartha (1995) Marginal Likelihood from the Gibbs Output, *Journal of the American Statistical Association*, 90(432): 1313–1321.

Chib, Siddhartha and Jeliazkov, Ivan (2001) Marginal Likelihood from the Metropolis-Hastings Output, *Journal of the American Statistical Association*, 96(453): 270–281.

Copas, John (1969) Compound Decisions and Empirical Bayes, *Journal of the Royal Statistical Society, Series B*, 31: 397–423.

Duane, Simon, Kennedy, Anthony D., Pendleton, Brian J. and Roweth, Duncan (1987) Hybrid Monte Carlo, *Physics Letters B*, 195(2): 216–222.

Escobar, Michael D. and West, Mike (1995) Bayesian Density Estimation and Inference Using Mixtures, *Journal of the American Statistical Association*, 90(430): 577–588.

Evans, Stephen (1994) Discussion of the Paper by Spiegelhalter, Freedman, and Parmar, *Journal of the Royal Statistical Society, Series A*, 157: 395.

Ferguson, Thomas (1973) A Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, 1(2): 209–230.

Fisher, Ronald A. (1925a) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, Ronald A. (1925b) Theory of Statistical Estimation, *Proceedings of the Cambridge Philosophical Society*, 22: 700–725.

Gamerman, Dani and Lopes, Hedibert F. (2006) *Markov Chain Monte Carlo*, 2nd edition. New York: Chapman & Hall.

Gelfand, Alan E. and Smith, Adrian F.M. (1990) Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85(410): 398–409.

Gelman, Andrew and Shalizi, Cosma R. (2013) Philosophy and the Practice of Bayesian Statistics, *British Journal of Mathematical and Statistical Psychology*, 66: 8–38.

Geman, Stuart and Geman, Donald (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741.

Gilks, Walter R., Richardson Sylvia and Spiegelhalter, David J. (1996) *Markov Chain Monte Carlo In Practice*. New York: Chapman & Hall/CRC.

Gill, Jeff (1999) The Insignificance of Null Hypothesis Significance Testing, *Political Research Quarterly*, 52(3): 647–74.

Gill, Jeff (2014) *Bayesian Methods for the Social and Behavioral Sciences*. New York: Chapman & Hall/CRC.

Gill, Jeff and Casella, George (2009) Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation, *Journal of the American Statistical Association*, 104(486): 453–454.

Gill, Jeff and Torres, Michelle (2019) *Generalized Linear Models: A Unified Approach*, 2nd edition. Thousand Oaks: Sage Publications.

Gill, Jeff and Waterman, Richard (2004) Solidary and Functional Costs: Explaining the Presidential Appointment Contradiction, *Journal of Public Administration Research and Theory*, 14(4): 547–569.

Gill, Jeff and Womack, Andrew (2013) The Multilevel Model Framework. Edited by: Marc A. Scott, Jeffrey S. Simonoff and Brian D. Marx, *The Sage Handbook of Multilevel Modeling*. Thousand Oaks: Sage Publications. pp. 3–20.

Jain, Sonia and Neal, Radford M. (2004) A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model, *Journal of Computational and Graphical Statistics*, 13(1): 158–182.

Jeffreys, Harold (1983) *Theory of Probability*. Oxford: Clarendon Press.

Korwar, Ramesh M. and Hollander, Myles (1973) Contributions to the Theory of Dirichlet Processes, *The Annals of Statistics*, 1(4): 705–711.

Kyung, Minjung, Gill, Jeff and Casella, George (2010) Estimation in Dirichlet Random Effects Models, *The Annals of Statistics*, 38(2): 979–1009.

Kyung, Minjung, Gill, Jeff and Casella, George (2011) New Findings from Terrorism Data: Dirichlet Process Random Effects Models for Latent Groups, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 60(5): 701–721.

Kyung, Minjung, Gill, Jeff and Casella, George (2012) Sampling Schemes for Generalized Linear Dirichlet Process Random Effects Models. With Discussion and Rejoinder, *Statistical Methods and Applications*, 20(3): 259–290.

Laplace, Pierre-Simon (1774) Mémoire sur la Probabilité des Causes par le Évènemens, *Mémoires de l'Académie Royale des Sciences Presentés par Divers Savans*, 6: 621–656.

Laplace, Pierre-Simon (1781) Mémoire sur la Probabilités, *Mémoires de l'Académie Royale des Sciences de Paris*, 1778: 227–332.

Leimkuhler, Benedict and Reich, Sebastian (2005) *Simulating Hamiltonian Dynamics*. Cambridge: Cambridge University Press.

Liu, Jun S. (1996) Nonparametric Hierarchical Bayes via Sequential Imputations, *The Annals of Statistics*, 24(3): 911–930.

Lo, Albert Y. (1984) On a Class of Bayesian Nonparametric Estimates: I. Density Estimates, *The Annals of Statistics*, 12(1): 351–357.

MacEachern, Steven N. and Müller, Peter (1998) Estimating Mixture of Dirichlet Process Models, *Journal of Computational and Graphical Statistics*, 7(2): 223–238.

Mackenzie, G. Calvin and Light, Paul (1987) *Presidential Appointees, 1964–1984, ICPSR Study 8458*. Ann Arbor, Michigan: Inter-University Consortium for Political and Social Research.

McCulloch, Charles E., Searle, Shayle R. and Neuhaus, John M. (2008) *Generalized, Linear, and Mixed Models*, 2nd edition. New York: John Wiley & Sons.

Metropolis, Nicholas, Rosenbluth, Arianna W., Rosenbluth, Marshall N., Teller, Augusta H. and Teller, Edward (1953) Equation of State Calculations by Fast Computing Machines,

Journal of Chemical Physics, 21(1087): 1087–1091.

Meyer, Kenneth, Hall, Glen and Offin, Daniel C. (1992) *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, 2nd edition. New York: Springer-Verlag.

Morris, Carl N. (1983) Parametric Empirical Bayes Inference: Theory and Applications, *Journal of the American Statistical Association*, 78(381): 47–55.

Neal, Radford M. (1993) Probabilistic Inference Using Markov Chain Monte Carlo Methods, *Technical Report CRG-TR-93-1*, Dept. of Computer Science, University of Toronto.

Neal, Radford M. (2011) MCMC Using Hamiltonian Dynamics. Edited by: Steve Brooks, Andrew Gelman, Galin Jones and Xiao-Li Meng, *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press. pp. 113–162.

Neal, Radford M. M. (2003) Slice Sampling, *Annals of Statistics*, 31(3): 705–767.

O'Hagan, Anthony (1994) *Bayesian Inference*, Volume 2 (Part 2 of Kendall's Advanced Theory of Statistics). London: Edward Arnold.

Polson, Nicholas (1996) Convergence of Markov Chain Monte Carlo Algorithm. Edited by: James O. Berger, José M. Bernardo, Alexander P. Dawid, Dennis V. Lindley and Adrian F. M. Smith, *Bayesian Statistics 5*. Oxford: Oxford University Press. pp. 297–321.

Robert, Christian and Casella, George (2004) *Monte Carlo Statistical Methods*, 2nd edition. New York: Springer-Verlag.

Robert, Christian and Casella George (2011) A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data, *Statistical Science*, 26(1): 102–115.

Robert, Christian and Mengersen, Kerrie L. (1999) Reparameterization Issues in Mixture Estimation and Their Bearings on the Gibbs Sampler, *Computational Statistics and Data Analysis*, 29(3): 325–343.

Roberts, Gareth O. and Rosenthal, Jeffrey S. (1998) Two Convergence Properties of Hybrid Samplers, *Annals of Applied Probability*, 8(2): 397–407.

Roberts, Gareth O. and Tweedie, Richard L. (1996) Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms, *Biometrika*, 83(1): 95–110.

Sethuraman, Jayaram (1994) A Constructive Definition of Dirichlet Priors, *Statistica Sinica*, 4(2): 639–650.

Strevens, Michael (2006) The Bayesian Approach to the Philosophy of Science. Edited by: Donald M. Borchert, *Macmillan Encyclopedia of Philosophy*, 2nd edition. Carmel: Pearson. pp. 495–502.

Teh, Yee W., Jordan, Michael I., Beal, Matthew J. and Blei, David M. (2006). Hierarchical Dirichlet Process, *Journal of the American Statistical Association*, 101(476): 1566–1581.

Zellner, Arnold (1971) *Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.