Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo

Author(s): Simon Jackman

Source: *American Journal of Political Science*, Vol. 44, No. 2 (Apr., 2000), pp. 375-404
Published by: Midwest Political Science Association
Stable URL: http://www.jstor.org/stable/2669318
Accessed: 25-07-2017 17:44 UTC

Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo

Simon Jackman Stanford University

Bayesian statistics have made great strides in recent years, developing a class of methods for estimation and inference via stochastic simulation known as Markov Chain Monte Carlo (MCMC) methods. MCMC constitutes a revolution in statistical practice with effects beginning to be felt in the social sciences: models long consigned to the "too hard" basket are now within reach of quantitative researchers. I review the statistical pedigree of MCMC and the underlying statistical concepts. I demonstrate some of the strengths and weaknesses of MCMC and offer practical suggestions for using MCMC in social-science settings. Simple, illustrative examples include a probit model of voter turnout and a linear regression for time-series data with autoregressive disturbances. I conclude with a more challenging application, a multinomial probit model, to showcase the power of MCMC methods.

> Bayesianism has obviously come a long way. It used to be that you could tell a Bayesian by his tendency to hold meetings in isolated parts of Spain and his obsession with coherence, self-interrogation, and other manifestations of paranoia. Things have changed. . . .
>
> *P. Clifford (1993, 53)*

Markov Chain Monte Carlo (MCMC) methods are probably the most exciting development in statistics within the last ten years. The techniques comprising MCMC are extraordinarily general, and their use has dramatically reshaped the way applied statisticians go about their work. Models long thought to be in the "too hard" basket are now well within the reach of quantitative researchers. In short, MCMC constitutes a revolution in statistical practice, with effects just beginning to be felt within the social sciences.

Consider a simple illustrative example that I will revisit in the pages that follow. Models for binary responses—typically, logit or probit models—are part of any quantitative social scientist's tool kit. Logit/probit models present no computational difficulties, and save for any degeneracies specific to a given data set, maximum likelihood estimates are easily obtained. The popularity and ease of MLE makes it easy to forget that underlying binary-response models is a latent regression function: $y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$, with the observed y_i equal to 0 when $y_i^* < 0$, and y_i when $y_i^* \geq 0$. Note that this is a garden-variety regression model, which we could estimate by least squares, *if we knew* y_i^*. Since we don't observe y_i^*, running this regression is impossible, and so we form a likelihood function for the observed binary responses; e.g., $\Pr(y_i = 1) = \Pr(y_i^* \geq 0) = \Pr(\varepsilon_i \geq -\mathbf{x}_i\boldsymbol{\beta})$, etc.

Simon Jackman is Assistant Professor of Political Science, Stanford University, Stanford, CA 94305-2044 (jackman@stanford.edu).

I thank Mike Alvarez, Larry Bartels, Neal Beck, Brad Carlin, John Freeman, John Jackson, Brad Jones, Eric Lawrence, Gary King, Robert Kohn, Walter Mebane, Rob McCulloch, Nicholas Polson, Douglas Rivers, Peter Rossi, Alastair Smith, Martin Tanner, Wendy Tam, and Bruce Western for advice, comments, and useful discussion. Students and colleagues at the ICPSR Quantitative Methods Summer School (1998 and 1999) also provided useful feedback. I acknowledge support from the Stanford Institute for the Quantitative Study of Society (and its Director, Norman Nie) and from the Reshaping Australian Institutions Project, Research School of the Social Sciences, Australian National University. Errors and omissions remain my own responsibility.

American Journal of Political Science, Vol. 44, No. 2, April 2000, Pp. 369–398

But to get a taste for MCMC methods, think of the latent y_i^* as missing data. Making imputations for the y_i^* is easy, given an approximation for $\boldsymbol{\beta}$, the covariates $\mathbf{x}_i$, and that the observed y_i tell us which side of 0 we will locate each corresponding y_i^*. With these imputations we have a "complete" data set for running the regression of y_i^* on $\mathbf{x}_i$, which updates the estimate of $\boldsymbol{\beta}$. The new estimate of $\boldsymbol{\beta}$ can be used to generate new imputations for y_i^*, which in turn allow us to update the estimate of $\boldsymbol{\beta}$. Repeating this procedure generates a sequence of estimates of $\boldsymbol{\beta}$ that converges to the MLE of $\boldsymbol{\beta}$.

This procedure—zig-zagging between estimating parameters and making imputations for missing data—is a simple example of a broad family of statistical techniques. As stated, the procedure described above is an example of the *EM* algorithm. But if at each stage of the algorithm we were to make imputations for y_i^* and generate our estimate of $\boldsymbol{\beta}$ by *sampling* from the appropriate distributions,[1] we have a simple example of a Markov Chain Monte Carlo method.

Although this introductory example might seem unremarkable, MCMC has an astonishingly broad range of applications. The underlying ideas are relatively straightforward to grasp, and once understood, unlock estimation problems long considered intractable or impossible. But despite the explosion of MCMC methods in the statistics literature,[2] most social scientists remain unaware of the usefulness of these methods. While there are a number of excellent summaries of MCMC methods in the statistical literature—for instance, I rely heavily on the taxonomy presented in Tanner (1996)—many of these treatments are all written at a level inaccessible to many social scientists, or are motivated with relatively unfamiliar applications.[3] My aim here is to remedy that deficiency.

MCMC has a distinctly Bayesian heritage and is associated with a resurgence in Bayesian statistics, prompting the humorous remarks in the epigraph. To be sure, the use of MCMC requires familiarization with some of the basics of Bayesian statistics. However, it is not necessary to commit to a Bayesian philosophical position in order to employ MCMC methods. When using MCMC, researchers often employ diffuse priors (which result in posterior densities that are overwhelmingly shaped by the data) or priors that have the effect of identifying parameters that would otherwise be inestimable. Accordingly, statisticians have accepted MCMC with surprisingly little controversy, given MCMC's Bayesian underpinnings. In short, one need not take sides in the centuries old "holy war" pitting frequentists against subjectivists in order to exploit the power of MCMC.[4]

[1]What is meant by an "appropriate" distribution will be made clearer below. For now, a useful simplification is to consider sampling from a distributions that have each y_i^* and $\hat{\boldsymbol{\beta}}$ as their respective means, such that on average, we still make the "correct" imputations, but the sampling accounts for our uncertainty in each set of quantities.

[2]By the mid-1990s Gelfand (1997) could claim that "several hundred papers" dealing with MCMC methods had appeared since 1990.

[3]Other general treatments include Gelman et al. (1995), Gilks, Richardson, and Spiegelhalter (1996), and Gamerman (1997). Albert and Chib (1996) is a good overview with econometric examples similar to those I employ here.

[4]For an introduction to this long-standing debate from a political science perspective, see Western and Jackman (1994).

Statistical Preliminaries

Most statistical inference in the social sciences is driven by probability models relating observed data, $\mathbf{y}$, to unknown parameters, $\boldsymbol{\theta}$. A simple example involves modeling Normal data: e.g., $y_i \sim N(\mu, \sigma^2)$, $\forall\ i = 1, \ldots, n$. The familiar linear regression model follows by replacing μ with $\mathbf{x}_i\boldsymbol{\beta}$. Generically, we can write these probability models as $\mathbf{y} \sim f(\mathbf{y}|\boldsymbol{\theta})$.

Likelihood and Frequentist Inference

The *likelihood function* summarizes the information about $\boldsymbol{\theta}$ in $\mathbf{y}$, defined as any function of $\boldsymbol{\theta}$ proportional to $f(\mathbf{y}|\boldsymbol{\theta})$ (Tanner 1996, 14):

$$L(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}). \tag{1}$$

Both the frequentist and Bayesian approaches to statistical inference exploit the likelihood function. Frequentist inference treats $\boldsymbol{\theta}$ as fixed but unknown and sample-based estimates of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, as random (since repeated sampling, if undertaken, could yield different values of $\hat{\boldsymbol{\theta}}$). Frequentists use the likelihood function to evaluate the plausibility of the other $\hat{\boldsymbol{\theta}}$ that might (hypothetically) result from repeated sampling, relative to the observed sample estimate $\hat{\boldsymbol{\theta}}$. Neyman-Pearson type inferential procedures such as likelihood ratio tests follow fairly directly from this perspective (e.g., Edwards 1992; Bernardo and Smith 1994, 450–455). This approach to statistical inference has been championed within political science by King (1989).

Bayesian Inference

Bayesian inference takes $\hat{\boldsymbol{\theta}}$ as fixed, conditional on the observed data y, and $\boldsymbol{\theta}$ as random. The Bayesian is interested in making posterior probability statements about $\boldsymbol{\theta}$ (i.e., "posterior" in the sense of "after" observing the data). The likelihood function summarizes the sample information about $\boldsymbol{\theta}$, providing an essential ingredient in Bayesian statistics. Recall Bayes' theorem,

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{2}$$

where $p(\boldsymbol{\theta})$ is the *prior* on $\boldsymbol{\theta}$ (which characterizes knowledge or beliefs about $\boldsymbol{\theta}$ before seeing the data), and $p(\boldsymbol{\theta}|\mathbf{y})$ is the *posterior density* of $\boldsymbol{\theta}$. Given the definition of the likelihood function in Equation 1, Bayes' Rule can be rewritten as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})\, L(\boldsymbol{\theta}|\mathbf{y}), \tag{3}$$

or, in words, a posterior density is proportional to the product of a prior distribution and a likelihood. Inference about $\boldsymbol{\theta}$ follows from inspection of the likelihood (for frequentists) or the posterior density (for Bayesians). Note immediately that when prior information is diffuse, $p(\boldsymbol{\theta})$ approximates a constant, and in turn the posterior density is proportional to the likelihood; i.e., with uninformative priors, Bayesian and classical procedures yield equivalent inferences about $\boldsymbol{\theta}$.

Point Estimates vs. Posterior Densities

In the Bayesian framework inference involves communicating features of the posterior distribution of $\boldsymbol{\theta}$. For example, a Bayesian might report the mean or the mode of a posterior density, along with some measure of dispersion (perhaps quantiles or highest density regions), or perhaps even a graphical summary of the posterior (a histogram or density estimate). MCMC techniques tend to inherit this property of Bayesian analyses. Whereas procedures for conventional statistical inference focus attention on point estimates of parameters and their standard errors, MCMC methods seek to characterize a posterior distribution for parameters. Of course, sometimes it will be convenient to summarize a posterior in terms of its mean and standard deviation, say, for posteriors that are approximately Normal, and hence symmetric about the mean. But in some cases the mean and standard deviation may be misleading posterior summaries, say when the posterior is asymmetric or has multiple modes, and/or an asymptotically-valid Normal approximation is unrealistic. So while MLEs and least squares estimates are single numbers (i.e., solutions to well-defined optimization problems), MCMC methods produce *samples* from the joint posterior density of model parameters that are then summarized for the purposes of inference.

By *sampling* rather than *optimizing*, MCMC can make estimation and inference simpler for both Bayesians *and* frequentists. Substantively interesting statistical models can give rise to complex likelihood functions, having either lots of parameters or a computationally intractable functional form, or both. Even without complexity in the likelihood itself, sometimes the derivatives of the likelihood with respect to the unknown parameters are prohibitively expensive to derive, program, or compute. Maximization algorithms may reach terminal solutions extremely slowly or not at all, say because there are many parameters in the likelihood function, and/or because the likelihood function is highly nonlinear in the parameters. In other cases the likelihood will be known *a priori* not to have a unique maximum, as is the case in unconstrained finite mixture models. In yet another class of cases, the researcher may want to estimate not just parameters, but the values of missing data points as well, complicating the optimization problem substantially (at first glance, missing data problems can appear to involve estimating many more parameters than data points). And in explicitly Bayesian setups, sometimes the form of the joint posterior density for all the model parameters may be extremely complicated, even if the likelihood can be calculated relatively simply.

The Data Augmentation Principle: Simplifying Estimation via Conditioning

Over the last twenty years or so, a number of related techniques have been developed for dealing with tough maximization or approximation problems, such as those discussed above. These techniques are a useful point of departure for considering MCMC. A key development is the *EM* algorithm, typically credited to Dempster, Laird, and Rubin (1977),[5] which bears some useful resemblances to the primary MCMC technique, Gibbs sampling. Tanner neatly summarizes the common foundation of these approaches:

> ... rather than performing a complicated maximization or simulation, one augments the observed data with "stuff" (latent data) which simplifies the calculation and subsequently performs a series of simple

[5]Many scholars remark that *EM* was foreshadowed in numerous places in the statistics literature. See Titterington, Smith, and Makov (1985, 84) or McLachlan and Krishnan (1997, 34–34) for summaries.

> maximizations or simulations. This "stuff" can be the "missing" data or parameter values. The principle of data augmentation can then be stated as follows: Augment the observed data Y with latent data Z so that the augmented posterior distribution $p(\boldsymbol{\theta}|Y, Z)$ is "simple." Make use of this simplicity in maximizing/marginalizing/calculating/sampling the observed posterior $p(\boldsymbol{\theta}|Y)$. (1996, 38)

Underlying this intuition is the following *posterior identity*:

$$p(\boldsymbol{\theta}|Y) = \int_Z p(\boldsymbol{\theta}|Y, Z)p(Z|Y)dZ, \quad (4)$$

where Y and Z are observed and latent data, respectively. The critical idea is that while $p(\boldsymbol{\theta}|Y)$ may be difficult to work with, if we could condition on Z as well, the posterior density or likelihood for $\boldsymbol{\theta}$ would be much easier to evaluate. This is well and good, but in implementing this strategy one needs to come up with values of Z. The integration over the predictive density for Z, $p(Z|Y)$, averages over more or less likely values of Z. Performing this integration by Monte Carlo methods is the second "MC" in the MCMC acronym (see the following section).

Lest there be any confusion, the "augmentation" referred to here is not as suspicious as it might first sound. There is no "double counting" of the data or any other statistical sleight of hand. "Augmenting the observed data with latent data" is simply a convenient mechanism for estimation. In the examples below "latent data" turns out to be quantities that are presumed to exist in specific models, but are unobserved by the analyst (e.g., disturbances, or utilities in a discrete choice problem). The model structure and current guesses about the model parameters generate the conditional expectations or conditional distributions for these quantities, which are then used to update estimates of the model parameters. In short, "data augmentation" is simply a clever way of exploiting model assumptions and the observed data so as to estimate parameters.

Monte Carlo

Many readers will be familiar with Monte Carlo simulations in deriving the repeated sampling characteristics of a statistic. Similarly, the use of a computer to repeatedly sample and average allows us to escape some of the more thorny mathematical expressions routinely encountered in applied Bayesian statistics. Integrations of the sort in the posterior identity are a typical example. Consider the expression

$$J(y) = \int f(y|x)g(x)dx = E[f(y|x)],$$

i.e., the expected value of some function of y, conditional on x. Note that $f(y)$ could be an identity, in which case $J(y) = E(y|x)$. If $g(x)$ is a probability density from which we can generate random samples, then $J(y)$ can be approximated by *Monte Carlo integration*: i.e.,

$$\hat{J}(y) = \frac{1}{n}\sum_{i=1}^{n} f(y|x_i),$$

where $x_1, \ldots, x_n \overset{\text{iid}}{\sim} g(x)$ are samples from the density of x. Importantly, the approximation $\hat{J}(y)$ grows more accurate as $n \to \infty$. If it is computationally inexpensive to sample from $g(x)$, evaluate $f(y|x_i)$, and store the results, then it is possible to obtain arbitrarily precise evaluations of integrals of the sorts given above, by setting n to a large number (Geweke 1989; Tanner 1996, 51).

Generalizing this method yields a *sample* from $J(y)$. For example, recall the posterior identity in Equation 4, $p(\boldsymbol{\theta}|Y) = \int p(\boldsymbol{\theta}|Y, Z)p(Z|Y)dZ$. A sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta}|Y)$ can be generated by (1) drawing $Z^{(i)}$ from $p(Z|Y)$; (2) drawing $\boldsymbol{\theta}^{(i)}$ from $p(\boldsymbol{\theta}|Y, Z^{(i)})$, $i = 1, \ldots, n$.

Summary

To conclude this introduction, Table 1 summarizes the distinctions between estimation and inference by the method of maximum likelihood, the *EM* algorithm, and MCMC. The *EM* algorithm emerges as a (pedagogically) useful intermediate method between MLE and MCMC methods. Like MLE, *EM* provides point estimates as a solution to an optimization problem, but exploits the posterior identity in Equation 4. MCMC also exploits the posterior identity, but in the context of *sampling* from marginal posterior densities.

EM

The *EM* algorithm is *not* a MCMC method, since it does not involve sampling. Rather, the *EM* algorithm is an optimizer, a method for computing and finding local maxima of likelihoods (or, from a Bayesian perspective, finding the mode of a posterior density). The *EM* algorithm is traditionally employed when the researcher has missing data to worry about in addition to parameters to estimate. The missing data are a problem in that they are *nonignorable*, meaning that simply dropping the observations with missing or incomplete data from the analysis

TABLE 1 Summary and Comparison: Three Methods for Estimation and Inference

Procedure	Output	Inference
MLE: optimization of likelihood function $L(\boldsymbol{\theta}\|\mathbf{y}) \propto f(\mathbf{y}\|\boldsymbol{\theta})$.	point estimate: $\hat{\boldsymbol{\theta}}_{\text{MLE}}$	$\text{var}(\hat{\boldsymbol{\theta}}_{\text{MLE}}) \approx -\left[\left.\frac{\partial^2 L(\boldsymbol{\theta}\|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\|_{\hat{\boldsymbol{\theta}}_{\text{MLE}}}\right]^{-1}$
EM: Let $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)}) = \int \ln[p(\boldsymbol{\theta}\|Z,Y)]p(Z\|\boldsymbol{\theta}^{(t)},Y)dZ$, 1. *E* step: compute $Z^{(t)} = E(Z\|\boldsymbol{\theta}^{(t)},Y)$ 2. *M* step: $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \ln[p\boldsymbol{\theta}\|Z^{(t)},Y)]$	point estimate: $\hat{\boldsymbol{\theta}}_{\text{MLE}}$	additional computation required
MCMC: Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_j)'$. 1. Sample $\boldsymbol{\theta}_1^{(t+1)}$ from $p(\boldsymbol{\theta}_1\|\boldsymbol{\theta}_2^{(t)},\boldsymbol{\theta}_3^{(t)},\ldots,\boldsymbol{\theta}_j^{(t)},Y)$. 2. Sample $\boldsymbol{\theta}_2^{(t+1)}$ from $p(\boldsymbol{\theta}_2\|\boldsymbol{\theta}_1^{(t+1)},\boldsymbol{\theta}_3^{(t)},\ldots,\boldsymbol{\theta}_j^{(t)},Y)$. . . . *j*. Sample $\boldsymbol{\theta}_j^{(t+1)}$ from $p(\boldsymbol{\theta}_j\|\boldsymbol{\theta}_1^{(t+1)},\boldsymbol{\theta}_2^{(t+1)},\ldots,\boldsymbol{\theta}_{j-1}^{(t+1)},Y)$.	sampled values: $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(n)},\ldots,\boldsymbol{\theta}^{(T)}$	calculate confidence intervals from observed quantiles of sampled $\boldsymbol{\theta}^{(n)},\ldots,\boldsymbol{\theta}^{(T)}$.

will bias the parameter estimates and other quantities of interest yielded by the model (Little & Rubin 1987; Gelman et al. 1995, 199). The augmentation in the *EM* algorithm consists of an imputation for the missing data (Z, in the context of Equation 4 above), which yields a *complete* data set with which to calculate the posterior density $p(\boldsymbol{\theta}|Y)$ or evaluate the log-likelihood $L(\boldsymbol{\theta}|Y)$.

Applying the identity in Equation 4, we obtain the following function Q, the log-posterior density of $\boldsymbol{\theta}$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \int \ln[p(\boldsymbol{\theta}|Z, Y)]p(Z|\boldsymbol{\theta}^{(t)}, Y)dZ, \qquad (5)$$

or, in words, the log-posterior for $\boldsymbol{\theta}$ (or the log-likelihood, given an uninformative prior for $\boldsymbol{\theta}$) is formed by averaging over imputations for Z, which are in turn generated using iteration t's "best guess" for $\boldsymbol{\theta}$, denoted as $\boldsymbol{\theta}^{(t)}$.

In the *E* or *expectation* step we average over possible values of the missing data Z, using the expected value of Z so as to to evaluate the log-posterior or log-likelihood for the parameters of substantive interest, $\boldsymbol{\theta}$. In the *M* step of the algorithm, the Q function is *maximized* with respect to $\boldsymbol{\theta}$ to yield the the next iteration's estimate, $\boldsymbol{\theta}^{(t+1)}$. The algorithm is iterated until convergence in the log-likelihood or parameters.

Under a wide set of conditions, the *EM* algorithm yields an "*EM* sequence" $\langle\boldsymbol{\theta}^{(t)}\rangle$ that is monotonically increasing in the incomplete-data likelihood function. Furthermore, if a likelihood function $L(\boldsymbol{\theta})$ is unimodal, with $\boldsymbol{\theta}^*$ being the only stationary point, then subject to some continuity assumptions, the *EM* sequence converges to the unique maximizer $\boldsymbol{\theta}^*$ of $L(\boldsymbol{\theta})$ (i.e., the MLE of $\boldsymbol{\theta}$); proofs and statements of the necessary regularity conditions can be found in the statistical literature (e.g., McLachlan and Krishnan 1997, Chapter 3).

Example: Probit Model for Binary Data

Consider a probit model for a binary outcome, $y_i \in \{0, 1\}$, $i = 1, \ldots, n$. As discussed in the introduction, we relate the observed binary outcome to covariates via the latent regression function

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \qquad (6)$$

where $\mathbf{x}_i$ is a row vector of observations on k independent variables, $\boldsymbol{\beta}$ is a column vector of parameters to be estimated, $y_i^* \in \mathbb{R}$ is a latent dependent variable, observed only in terms of its sign, i.e.,

$$y_i = \begin{cases} 0, & \text{if } y_i^* < 0, \\ 1, & \text{if } y_i^* \geq 0, \end{cases}$$

and ε_i is a zero mean stochastic disturbance, identically and independently distributed for all i. For probit, we will assume $f(\varepsilon_i) = N(0,1) \equiv \phi()$, the standard normal density, normalized to have unit variance (recall that the regression parameters $\boldsymbol{\beta}$ are identified only up to the scale factor σ, and so setting $\sigma = 1$ is a convenient

normalization with no substantive implications). This model yields a relatively simple log-likelihood function that can be easily maximized with respect to the parameters $\boldsymbol{\beta}$. However, for expository purposes, consider estimating $\boldsymbol{\beta}$ using the *EM* algorithm, treating the latent $\mathbf{y}^*$ as missing data.

Each y_i^* is known only in terms of its sign (given by the corresponding y_i), but we can use the current estimate of $\boldsymbol{\beta}$ and other model assumptions to make an imputation for each y_i^*, the *conditional expectation* of each y_i^*; conditional on that imputation we then choose $\hat{\boldsymbol{\beta}}$ so as to *maximize* the complete-data log-likelihood, thereby updating our estimate of $\boldsymbol{\beta}$. These two steps—(1) calculating the expected value of $\mathbf{y}^*$ conditional on $\hat{\boldsymbol{\beta}}^{(t)}$ and the observed data; and (2) updating $\hat{\boldsymbol{\beta}}^{(t)}$—comprise the *E* (expectation) and *M* (maximization) steps of the *EM* algorithm at iteration *t*, for this problem.

The imputation for $\mathbf{y}^*$ is given by

$$\begin{aligned} E(y_i^*|y_i,\mathbf{x}_i,\hat{\boldsymbol{\beta}}^{(t)}) &\equiv y_i^{*(t)} \\ &= E[(\mathbf{x}_i\boldsymbol{\beta}+\varepsilon_i)|y_i,\mathbf{x}_i,\hat{\boldsymbol{\beta}}^{(t)}] = \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(t)} + M_i, \end{aligned}$$

where

$$M_i = E(\varepsilon_i|y_i,\mathbf{x}_i,\hat{\boldsymbol{\beta}}^{(t)}) = \begin{cases} -\phi_i/\Phi_i & \text{if } y_i = 0, \\ \phi_i/(1-\Phi_i) & \text{if } y_i = 1, \end{cases}$$

and where $\phi_i = \phi(-\mathbf{x}_i\hat{\boldsymbol{\beta}}^{(t)})$ is the Normal probability density function, and $\Phi_i = \Phi(-\mathbf{x}_i\hat{\boldsymbol{\beta}}^{(t)})$ is the Normal cumulative distribution function, each evaluated at $-\mathbf{x}_i\hat{\boldsymbol{\beta}}^{(t)}$ (e.g., Johnson, Kotz, and Balakrishnan 1994, 156). Armed with the imputed y_i^*, we update the estimate of $\boldsymbol{\beta}$ by choosing the value of $\boldsymbol{\beta}$ that maximizes the complete-data log-likelihood, simply by running a regression of the imputed values for $\mathbf{y}^*$ on the covariates $\mathbf{X}$:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(t+1)} &= \arg\min_{\boldsymbol{\beta}}\left(\frac{1}{2}\sum_{i=1}^{n}\left[E\left(y_i^*\middle|y_i,\mathbf{x}_i,\hat{\boldsymbol{\beta}}^{(t)}\right)-\mathbf{x}_i\boldsymbol{\beta}\right]^2\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}^*|\mathbf{y},\mathbf{X},\hat{\boldsymbol{\beta}}^{(t)}). \end{aligned}$$

The algorithm is iterated until convergence in the complete-data log-likelihood and/or the parameters. At convergence, we require estimates of the parameter's standard errors in order to perform inference, and in general, the *EM* algorithm will not supply these as a matter of course. Thus we might consider the *EM* algorithm as an estimation tool, but not a tool for inference.

I implemented this algorithm for a probit model, using a random subset of 3,000 observations from Nagler's (1994) data on voter turnout, from the 1984 Current Population Survey; predictor variables are education, age, the number of days registration closes before the election, whether or not a gubernatorial election took place in the respondent's state, and whether the respondent lives in the South. Starting values come from an OLS regression of the observed binary responses on the covariates, and after thirty iterations of the *EM* algorithm the log-likelihood was increasing by steps of less than 10^{-9}.

Figure 1 shows the iterative history of the *EM* algorithm for the log-likelihood, two parameters, and the estimated value of the latent dependent variable for the 1,000th observation. The algorithm converges quite quickly in this case, and after a few iterations has done most of its work, moving away from the OLS starting values towards the maximum likelihood estimates.

Example: Linear Regression with AR(1) Disturbances (Presidential Approval)

Consider a regression set-up with a pattern of first-order auto-correlation among the regression's disturbances:

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + u_t, \qquad u_t = \rho u_{t-1} + e_t,$$

where $\mathbf{y}$ is a T by 1 vector of observations on the dependent variable, $\mathbf{X}$ is a T by k matrix of explanatory variables, $\boldsymbol{\beta}$ is a k by 1 vector of parameters to be estimated, ρ is a scalar to be estimated ($|\rho| < 1$), $e_t \sim N(0,\sigma^2)$, $\forall\ t$, and $t = 1, \ldots, T$ indexes the observations. Given normality, the log-likelihood for a linear regression with stationary AR(1) disturbances is

$$\begin{aligned} &\ln L(\boldsymbol{\beta},\rho,\sigma^2|\mathbf{y},\mathbf{X}) \\ &\quad = -\frac{T}{2}(\ln(2\pi)+\ln\sigma^2) + \frac{\ln(1-\rho^2)}{2} - \frac{\mathbf{u}^{*\prime}\mathbf{u}^*}{2\sigma^2}, \end{aligned} \tag{7}$$

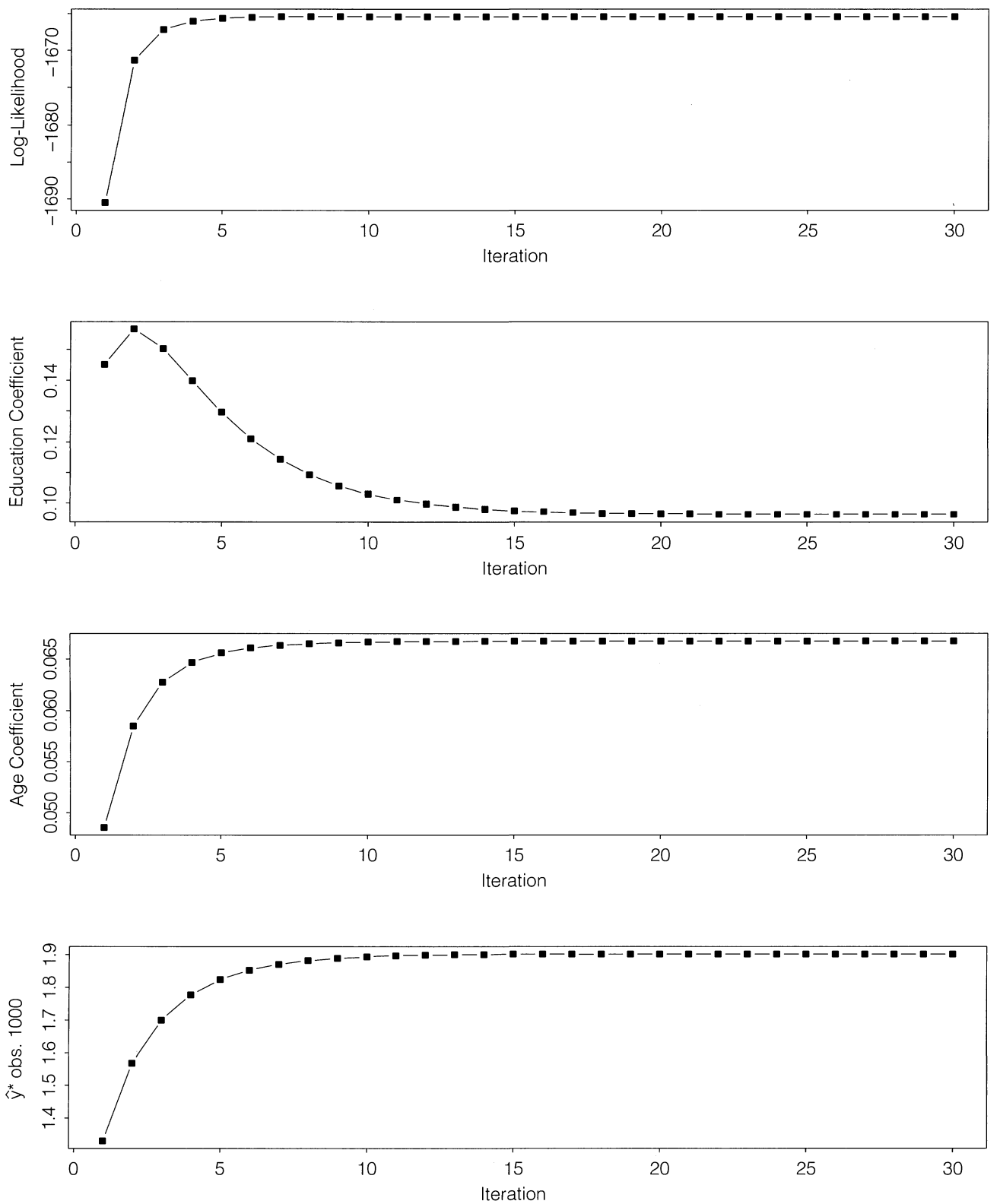
where $\mathbf{u}^* = \mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}$, and

$$\mathbf{y}^* = \begin{bmatrix} \sqrt{1-\rho^2}\,y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}, \quad \mathbf{X}^* = \begin{bmatrix} \sqrt{1-\rho^2}\,\mathbf{x}_1 \\ \mathbf{x}_2 - \rho\mathbf{x}_1 \\ \mathbf{x}_3 - \rho\mathbf{x}_2 \\ \vdots \\ \mathbf{x}_T - \rho\mathbf{x}_{T-1} \end{bmatrix}$$

are the familiar transformations that retain the first observation (Prais and Winsten 1954).

To motivate the application of the *EM* algorithm in this context, consider ρ as "missing data" in the context of estimating $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)'$; alternatively, think of the transformed "white-noise" disturbances $\mathbf{u}^*$ as missing

FIGURE 1 Iterative History of EM Algorithm, Probit Model of Voter Turnout



data, but dependent on the observed data via ρ. In any event, with ρ unknown, evaluating the log-likelihood in Equation 7 is clearly problematic. With ρ treated as "information" to be imputed, the Q function or posterior identity appropriate in this context is

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}) = \int_{-1}^{1} \ln p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\rho)p(\rho|\boldsymbol{\theta}^{(i)},\mathbf{y},\mathbf{X})d\rho, \quad (8)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)'$ and the limits of integration follow from the assumption of stationarity. The posterior identity shows that the estimation problem here has been broken into two components; (1) averaging over the uncertainty as to the value of ρ, conditional on the data and other parameters, and (2) using the estimated value of ρ to transform the data and then make inferences about the other parameters. The *EM* algorithm's two steps implement each of these tasks in turn. The *E* step yields imputations for the transformed data $\mathbf{y}^*$ and $\mathbf{X}^*$ conditional on the observed data and the current estimates of $\boldsymbol{\beta}$ and ρ. The *M* step consists of finding estimates of the model parameters that maximize Equation 7 conditional on the imputation for ρ and the resulting imputations for $\mathbf{y}^*$ and $\mathbf{X}^*$.

Consider each step. At the end of iteration i, estimates of $\boldsymbol{\beta}$ and σ^2 are $\hat{\boldsymbol{\beta}}^{(i)}$ and $\hat{\sigma}^{2(i)}$, respectively. Conditional on these estimates and the data, the predictive-density for $\rho^{(i+1)}$ is

$$p(\rho^{(i+1)}|\hat{\boldsymbol{\beta}}^{(i)},\hat{\sigma}^{2(i)},\mathbf{y},\mathbf{X}) = \prod_{t=2}^{T}\left[(2\pi\sigma^{2(i)})^{-1/2}\exp\left(\frac{(u_t^{(i)}-\rho u_{t-1}^{(i)})^2}{2\hat{\sigma}^{2(i)}}\right)\right]$$

given $e_t = u_t - \rho u_{t-1}$ and $e_t \sim N(0,\sigma^2)\ \forall\ t$, and where $u_t^{(i)} = y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}^{(i)}$. Taking logs, differentiating and rearranging shows that the estimate of ρ that maximizes this predictive density is

$$\hat{\rho}^{(i+1)} = \frac{\sum_{t=2}^{T} u_t^{(i)}u_{t-1}^{(i)}}{\sum_{t=2}^{T}\left(u_{t-1}^{(i)}\right)^2}, \quad (9)$$

Notice that $\hat{\rho}^{(i+1)}$ is just the coefficient from the regression of $u_t^{(i)}$ on $u_{t-1}^{(i)}$, without a constant. With this estimate of ρ, the *M* step of the *EM* algorithm is quite simple as well. Maximizing the log-likelihood in Equation 7 with respect to $\boldsymbol{\beta}$ and σ^2 (now that ρ has been estimated or "imputed") is achieved via the regression of $\mathbf{y}^*$ on $\mathbf{X}^*$, where $\hat{\rho}^{(i+1)}$ is used in forming the transformed variables, as shown above. This regression yields $\hat{\boldsymbol{\beta}}^{(i+1)} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^*$, while a maximum-likelihood estimate of $\sigma^{2(i+1)}$ is $(\mathbf{u}^{*\prime}\mathbf{u}^*)/T$, where $\mathbf{u}^* = \mathbf{y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}^{(i+1)}$.

Iterating this algorithm provides maximum-likelihood estimates of the parameters. Readers familiar with the linear regression model with first-order autoregressive errors will see that this application of the *EM* algorithm is simply the Cochrane and Orcutt (1949) iterative method of estimating a regression with AR($\cdot$) errors. Each step in the *EM* algorithm here involves two regressions—a regression to obtain $\hat{\rho}$, and the other to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$—and successive iterations of the algorithm consist of "zigzagging" between these two regressions, to use Hamilton's (1994, 224) description.

I implemented this algorithm for a model of using monthly approval ratings for President Reagan, over his two terms in office ($T = 96$); covariates are the inflation rate and unemployment levels. I started the algorithm with starting values given by OLS with $\rho = 0$. After nine iterations the log-likelihood was changing by less than 10^{-8}, and estimates of the parameters had stabilized. The algorithm converges extremely quickly, moving rapidly away from the OLS starting values to within a small neighborhood of the maximum likelihood estimates. In particular, the coefficient for unemployment changes sign after just one iteration. This rapid convergence follows given that just ρ—a scalar—is the missing "stuff" in this context; contrast the probit example where the entire n by 1 vector of $\mathbf{y}^*$ must be imputed at each iteration.

Other Applications

The *EM* algorithm is perhaps most commonly recognized as a technique for imputing missing data (e.g., King et al. 1998). Indeed, the applications presented above all turn on an expansive notion of what constitutes missing data: the lesson of these applications is that it can be useful to treat unknown parameters or latent variables as missing data. For instance, Watson and Engle (1983) survey several models for econometric time series, representable in state-space form; they show that the *EM* algorithm is well-suited to estimating these models, treating the latent state vector as missing data, in conjunction with a Kalman filter. Shumway and Stoffer (1982) is an early study of this type. The so-called Hamilton (1990) model for Markov-switching time series is also relatively easy to estimate using the *EM* algorithm, treating the latent regime probabilities as missing data. Censored or truncated data examples (of which the probit example above is a special case) are among the earliest examples of the use of the *EM* algorithm; Dempster, Laird, and Rubin (1977, 15) provide citations that pre-date their introduction of the "*EM*" acronym.

Summary

To reiterate, while computing and evaluating likelihoods is "the name of the statistical game," in various situations this is more or less difficult, either because the expression for the likelihood contains a great many unknown quantities (i.e., missing data and model parameters). But if the *EM* algorithm finds a maximum of the log-likelihood with respect to unknown parameters in the presence of missing data, why can't it do the same where the data are known, but the parameters are unknown or "missing"? That is, why not make the Z term in the posterior identity a vector of parameters instead of missing data?

This extension of the *EM* algorithm has yielded some tremendous dividends in specific applications. To restate the driving idea here, missing data and unknown parameters are both instances of *information*—"stuff," in Tanner's phrase—required to evaluate a likelihood. Blurring the distinction between "data" and "parameters" turns out to be very powerful and underlies MCMC methods.

Gibbs Sampling

Gibbs sampling—the workhorse MCMC method—differs from the *EM* algorithm in two critical respects. First, instead of calculating expected values of the "missing stuff," the Gibbs sampler *samples* from the conditional distributions for each quantity. That is, the Gibbs sampler performs the integration in the posterior identity (Equation 4) by Monte Carlo methods. Second, the Gibbs sampler does not distinguish between the parameters of direct substantive interest (e.g., regression coefficients) and the nuisance parameters (e.g., missing data points); all are considered random quantities and can be stacked into a single parameter vector $\boldsymbol{\theta}$. As a practical matter, it is convenient to partition $\boldsymbol{\theta}$ into d blocks or subvectors (possibly scalars), $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d)$. Iteration t of the Gibbs sampler starts with $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_d^{(t)})$ and makes the transition to $\boldsymbol{\theta}^{(t+1)}$ via the following scheme:

1. Sample $\boldsymbol{\theta}_1^{(t+1)}$ from $p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_d^{(t)}, \mathbf{Y})$.
2. Sample $\boldsymbol{\theta}_2^{(t+1)}$ from $p(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_d^{(t)}, \mathbf{Y})$.

⋮ ⋮

d. Sample $\boldsymbol{\theta}_d^{(t+1)}$ from $p(\boldsymbol{\theta}_d | \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_{d-1}^{(t+1)}, \mathbf{Y})$.

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 in to a series of lower-dimensional conditional densities, circumventing the "curse of [high] dimensionality" (Gelfand 1997, 284). In turn this is driven by the fact (well known to Bayesians) that "conditional [densities] determine marginals" (Casella and George 1992, 170–171).

The sequence of sampled vectors produced by this scheme, $\langle \boldsymbol{\theta}^{(t)} \rangle = \{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots\}$, form a Markov chain, hence the first "MC" in the MCMC acronym. More specifically, under a wide set of conditions, the sampled vector $\boldsymbol{\theta}^{(t)}$ is the state vector of a convergent Markov chain that has the posterior distribution $p(\boldsymbol{\theta}|Y)$ as the chain's "invariant," "stationary," or "limiting" distribution.[6] Thus, the output of the Gibbs sampler, $\boldsymbol{\theta}^{(t)}$, converges in distribution to the target posterior density as $t \to \infty$. Or more simply, when the Markov chain has been run for a sufficiently lengthy "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 (e.g., the proportion of sampled $\boldsymbol{\theta}$ that lie above or below zero, the observed data log-likelihood, residuals in a regression setting, or the percent cases correctly classified in a qualitative dependent variable context).

Critical to the utility of MCMC is that the samples forming the Markov chain are successively better approximations of the target posterior density $p(\boldsymbol{\theta}|Y)$. What drives this is the particular form of the transition probabilities governing the Markov chain, rather than the fact that the series of sampled $\boldsymbol{\theta}$ form a Markov chain per sé (for more on this point, see "Generalizations," below). As Gelman et al. (1995, 323) point out, there are many ways to sequentially explore parameter spaces that need not be Markovian (e.g., genetic algorithms, simulated annealing), and from a broader perspective, MCMC methods

[6] "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, 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) is that "the chain must be able to reach all interesting parts of the state-space" (Tierney 1996, 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 regions" in a finite number of transitions, since failing to do so means 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 counter-examples can be easily constructed (e.g., Gamerman 1997, 124).

are part of a family of methods that constitute "random tours" of parameter spaces (Fishman 1996, chapter 5). The virtue of the Markov property is that convergence results for Markov chains can be applied with relative ease, given the transition probabilities inherent in Gibbs sampling and other forms of MCMC.

Genesis

The Gibbs sampler is commonly attributed to Geman and Geman (1984) who used the technique to study image restoration; i.e., the joint distribution of the contents of a field of pixels (pixel values) is usually a complicated high dimensional density, but can be tractably dealt with by using conditional distributions, where the conditioning is on the contents of neighboring groups of pixels. Also dealing with a spatial setting, Besag (1974) showed that if a joint distribution of d components is positive over its entire domain, then that joint distribution is uniquely determined by the d conditional distributions (taking each component in turn and conditioning on the remaining $d-1$ components). This insight drives the Gibbs sampler, although in conventional statistical settings the relationships between elements of $\boldsymbol{\theta}$ are imposed by model assumptions, rather than by physical or spatial structure. In the spatial setting considered by Geman and Geman (1984; a Markov random field), the result of Besag (1974) holds if each of the d conditional distributions is a *Gibbs distribution*, and hence Geman and Geman gave the algorithm the name "Gibbs sampling."[7] Image reconstruction and other spatial settings remain an active area of development and application of MCMC (Besag and Green 1993; Green 1996; Smith and Roberts 1993, 18–19), although obviously application of the Gibbs sampler is by no means restricted to Gibbs distributions or statistical inference for spatial processes.

Gelfand and Smith (1990) are generally credited with bringing MCMC to the statistical mainstream. The data augmentation algorithm of Tanner and Wong (1987) amounts to a special case of Gibbs sampling, as is Rubin's (1987) work on multiple imputations for missing data.

[7] The "Gibbs" referred to here is Josiah Willard Gibbs (1839–1903), an American theoretical physicist and chemist, and one of the great figures of nineteenth–century science. Gibbs is credited with founding statistical mechanics via the application of principles of thermodynamics. Several other quantities and functions in statistical mechanics bear his name, most notably "Gibbs free energy" (the lowest energy state of a chemical system in a thermodynamic equilibrium with given pressure and temperature).

Generalizations

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; a useful explanation of the Metropolis-Hastings algorithm and practical tips for its implementation appears in Chib and Greenberg (1995).

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, 324, 326):

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

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

   which taps the plausibility of the candidate point $\boldsymbol{\theta}^*$ relative to the current value $\boldsymbol{\theta}^{(t-1)}$.
3. Set

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

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 $p(\boldsymbol{\theta}|y)$ as its invariant distribution. The power of the Metropolis-Hastings method stems from the fact that the proposal distribution *can have any form* and the invariant distribution of the resulting Markov chain will still be the desired posterior distribution, $p(\boldsymbol{\theta}|y)$; for proofs see Gilks, Richardson, and Spiegelhalter (1996) and the cites therein. Gibbs sampling is a special case of the Metropolis-Hastings algorithm in the sense that each component of $\boldsymbol{\theta}$ is updated sequentially, and the implicit jumping distributions are simply the conditional densities $p(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j}^{(t-1)}, y)$; this means that $r = 1$ and each candidate point is always accepted.

The Metropolis-Hastings algorithm is often used in conjunction with a Gibbs sampler for those components of $\boldsymbol{\theta}$ that have conditional distributions that can be evaluated, but can not be sampled from directly, typically because the distribution is known only up to a scale factor. The Metropolis-Hastings algorithm ensures that

MCMC algorithms can still be constructed for these cases. All that is required is that the analyst have some approximating density from which it is possible to sample and then be able to able to evaluate the ratio r with the sampled candidate point: this supplies the ingredients necessary to implement the Metropolis-Hastings algorithm. Examples of the use of Metropolis-Hastings in applications of MCMC include:

- estimating autoregressive parameters (Chib and Greenberg 1994; Chib and Greenberg 1995, 333 and following): Given an AR(p) process with Gaussian white noise, and with the only prior information being that the process is stationary, the resulting posterior density for the autoregressive parameters is not a standard density and known only up to a constant of proportionality.
- estimating a logit model with a generalized link function (e.g., Carlin and Louis 1996, 176 and following): Given the model

$$\Pr[y_i = 1] = \left[\frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta})}\right]^m$$

  the resulting conditional distributions for $\boldsymbol{\beta}$ and m do not have closed form expressions and are only known up to constants of proportionality.
- estimating precinct-level proportions in an "ecological" or "cross-level" model (King, Rosen, and Tanner 1999): With precincts $i = 1, \ldots, n$, consider a series of precinct-specific "two-by-two" cross-tabulations of T_i by X_i (say, turnout by race, respectively), for which we observe only the marginal distributions on T_i and X_i. This yields the identity

$$T_i = X_i\beta_i^b + (1 - X_i)\beta_i^w,$$

with $\beta_i^j \in [0,1]$ the unobserved cell entries (the observed T_i and X_i usually narrow the bounds on the β_i^j). King, Rosen, and Tanner assume the following three level *hierarchical* model: (a) binomial distributions for the *number* turning out in precinct i; (b) $\beta_i^b \sim \text{Beta}(c_b, d_b)$, $\beta_i^w \sim \text{Beta}(c_w, d_w)$ (i.e., independent beta distributions as priors on the β_i parameters); (c) c_b, d_b, c_w, and d_w each have a diffuse exponential prior, with mean 2. The posterior distribution (or log-likelihood, given diffuse priors) implied by this model is complex, and MCMC is an attractive way to recover the posterior densities of the model parameters. However, the conditional distributions for the unknown parameters are nonstandard and known only up to constants of proportionality, thus requiring the use of Metropolis-Hastings methods.

Assessing Convergence

Recall the general results described in this section: MCMC samplers will get to the desired posterior density for a very wide class of models, even though it may take a long time to get there. Determining how long is "sufficiently long" in particular settings is an ongoing topic of research (e.g., Rosenthal 1995; Polson 1996; Roberts 1996). Tierney (1997, 397) notes that "universally useful, reliable [convergence] diagnostics do not exist, and cannot exist," given the problem-specific Markov chains generated by MCMC.

A large and growing literature deals with techniques for monitoring convergence of MCMC sequences; Cowles and Carlin (1996) provide a comprehensive review of thirteen diagnostics. As Cowles and Carlin point out, the chief difficulty in diagnosing convergence is that MCMC produces *samples* from distributions, rather than the value of a function being optimized (refer to Table 1, above):

> Worse yet, the Markov nature of the algorithm means that members of this sample will generally be *correlated* with each other, slowing the algorithm in its attempt to sample from the entire stationary [posterior] distribution and muddying the determination of appropriate Monte Carlo variances for estimates of model characteristics based on the output. . . . such high correlations, both within the output for a single model parameter (*autocorrelations*) and across parameters (*cross-correlations*) are not uncommon, caused, for example, by a poor choice of parameterization or perhaps overparameterization. The latter situation can of course lead to "ridges" in the posterior or likelihood surface, long the bane of familiar statistical optimization algorithms (1996, 883–884).

Graphical inspection of MCMC sequences is critically important in assessing problems with convergence. In conjunction with formal, analytic diagnostics, so-called "trace plots" of the iterative history of MCMC sequences help researchers identify some common problems with convergence. In addition, to "slow-mixing" of the Markov chain, multi-modal posterior distributions are sometimes obvious from inspecting a trace plot. Slow mixing and multi-modal posteriors are not fatal in and of themselves—the theoretical results guaranteeing convergence to the posterior distribution apply to a wide range of circumstances—but they do mean that a long MCMC sequence may be needed to explore all regions of the parameter space with positive posterior probability.

A popular convergence diagnostic is based on Geweke's (1992) observation that for some function of a scalar output of the MCMC sampler, say $g(\boldsymbol{\theta})$, the spectral density of the time series $\{g(\boldsymbol{\theta})^{(t)}\}$ can be used to estimate the asymptotic variance of an estimate of the average of the time series. In turn then, this permits a comparison of two stages of the Markov chain (say "early" with n_A iterations and "late" based on the last n_B iterations), which yield estimates $\bar{g}(\boldsymbol{\theta})_A$ and $\bar{g}(\boldsymbol{\theta})_B$. The difference of these means divided by the asymptotic standard error of the difference tends to a standard normal distribution as $n \to \infty$ (holding n_A/n and n_B/n constant and $n_A + n_B < n$). Cowles and Carlin (1996, 866) discuss the strengths and weaknesses of this diagnostic. In particular, it is unclear how large n_A and n_B should be, relative to n, although Geweke suggested $n_A = .1n$ and $n_B = .5n$.

Just as practitioners are routinely advised to try different starting values with maximization routines, Gelman and Rubin (1992) recommend starting the Gibbs sampler with overdispersed starting points,[8] especially when working with a posterior distribution reasonably thought to be multi-modal. In practice this is best done by running several Gibbs samplers in parallel. Given output from parallel MCMC samplers, a simple test statistic can be formed by comparing within-sequence and between-sequence variation in each scalar component of $\boldsymbol{\theta}$ (Gelman and Rubin 1992).

Formally, Gelman and Rubin's convergence diagnostic is based on an estimate of the marginal posterior variance of some scalar estimand of interest

$$\widehat{\text{var}}^{+}(\psi|y) \frac{n-1}{n} W + \frac{1}{n} B,$$

where W is the (average) within-chain variance and B is the between-chain variance, for some scalar of interest ψ, conditional on observed data y. As $n \to \infty$ (i.e., we generate longer MCMC sequences), the contribution of the between-chain variation gets smaller, since it picks up weight $1/n$ in contributing to $\widehat{\text{var}}^{+}(\psi|y)$. Simultaneously, the within-chain variance increasingly dominates this term with additional iterations. Thus this estimate of the marginal posterior variance is an *overestimate* of the true marginal posterior variance for any finite length chain (hence the "+" superscript). Accordingly, Gelman and Rubin propose the following statistic as a convergence diagnostic:

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{\text{var}}^{+}(\psi|y)}{W}}.$$

This quantity declines to 1 as $n \to \infty$ and can be interpreted as the "potential scale reduction" that might result from continuing to run the MCMC sampler. Given streams of output from parallel Gibbs samplers, this statistic can be calculated after a prespecified number of iterations; Gelman et al. (1995, 332) suggest that values of $\sqrt{\hat{R}}$ below 1.2 are "acceptable," but any determination of convergence will vary from data set to data set.[9]

While the "parallel sequences" recommendation of Gelman and Rubin is widely endorsed in the statistical literature, many authors point out that there is an inevitable tradeoff between one long run of a MCMC sampler versus several shorter runs. Consequently, a consensus position lies somewhere between the "one long chain" and "shorter multiple chains" positions (e.g., Cowles and Carlin 1996, 903). Of course, "more is better," both in terms of the number of sequences run and the length of each sequence, given the theoretical results that guarantee that MCMC samplers will eventually reach the target posterior distribution in most settings. Convergence may be slow in a specific context, and so we might prefer devoting computer resources to one long MCMC run versus several shorter runs. Given that computational power is cheap, and getting cheaper, the "more is better" advice is increasingly easy to implement, as some of the examples below make clear. But most importantly, there is no substitute for a clear understanding of the model parameterization and data being passed to a MCMC sampler and how they might possibly impede convergence.

For instance, note that the components $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d$ of $\boldsymbol{\theta}$ can themselves be vectors of parameters. Collecting interdependent parameters in the same subvector of $\boldsymbol{\theta}$ is often an efficient strategy for speeding the convergence of a MCMC sequence. When groups of parameters are interdependent, sampling from their joint posterior density speeds up the convergence of the Gibbs sampler (e.g., sampling a vector of regression coefficients from their multivariate Normal joint density, rather than sampling each coefficient from its univariate Normal marginal density, conditional on the other coefficients).

[8] By overdispersed it is meant that the variance among the different starting points should be greater than that thought to exist in the target distribution. At the same time, the starting values should not be "wildly inaccurate" (Gelman and Rubin 1992, 458–459).

[9] This is one of the more simple versions of the Gelman and Rubin convergence diagnostic; more complicated versions and generalizations appear in the statistical literature (e.g., Brooks and Gelman 1998).

Examples

The number of applications of MCMC methods has increased dramatically since 1990, although they are just now making their way into the social sciences. Here I illustrate some simple uses of Gibbs sampling with the examples introduced earlier.

Example: Probit Model for Binary Data[10]

Recall the probit model introduced in the earlier *EM* section. At this stage *prior distributions* for the probit coefficients $\boldsymbol{\beta}$ are required, reflecting the Bayesian underpinnings of MCMC. A multivariate Normal prior on $\boldsymbol{\beta}$ is flexible and convenient and takes the general form $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_{\text{prior}}, \mathbf{B}_{\text{prior}})$, where $\boldsymbol{\beta}_{\text{prior}}$ is a k by 1 vector of prior means and $\mathbf{B}_{\text{prior}}$ is a k by k prior covariance matrix, reflecting the researcher's prior uncertainty regarding $\boldsymbol{\beta}$.

The Gibbs sampler requires expressions for the conditional distributions of all random quantities in the probit model, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{y}^*)'$. Albert and Chib (1993) show that these are

$$y_i^* | (y_i = 0, \mathbf{x}_i, \boldsymbol{\beta}) \sim N(\mathbf{x}_i\boldsymbol{\beta}, 1) I(y_i^* < 0) \quad \text{(trunc. Normal)} \tag{10}$$

$$y_i^* | (y_i = 1, \mathbf{x}_i, \boldsymbol{\beta}) \sim N(\mathbf{x}_i\boldsymbol{\beta}, 1) I(y_i^* \geq 0) \quad \text{(trunc. Normal)} \tag{11}$$

$$\boldsymbol{\beta} | \mathbf{y}^*, \mathbf{X}, \mathbf{y} \sim N(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{B}}), \tag{12}$$

where

$$\tilde{\boldsymbol{\beta}} = (\mathbf{B}_{\text{prior}}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{B}_{\text{prior}}^{-1}\boldsymbol{\beta}_{\text{prior}} + \mathbf{X}'\mathbf{y}^*)$$

$$\tilde{\mathbf{B}} = (\mathbf{B}_{\text{prior}}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

These last two expressions are simply those for the posterior mean and posterior covariance of regression parameters; the posterior mean $\tilde{\boldsymbol{\beta}}$ is the matrix-weighted average of the estimate of $\boldsymbol{\beta}$ from the data and the prior mean, where the matrix weights are the respective "precision matrices" (inverted covariance matrices) of the prior and the data. Note that with an uninformative prior $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$ and $\tilde{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}$ (i.e., the posterior corresponds to the results obtained by simply running a regression of $\mathbf{y}^*$ on $\mathbf{X}$).

[10]Albert and Chib (1993) considered the application of the Gibbs sampler in the context of models for binary and polychotomous outcomes. McCulloch and Rossi (1994) and Chib and Greenberg (1997) consider the application of MCMC to the multinomial probit model. Johnson and Albert (1999) is a treatment of binary response models, largely from the perspective of MCMC.

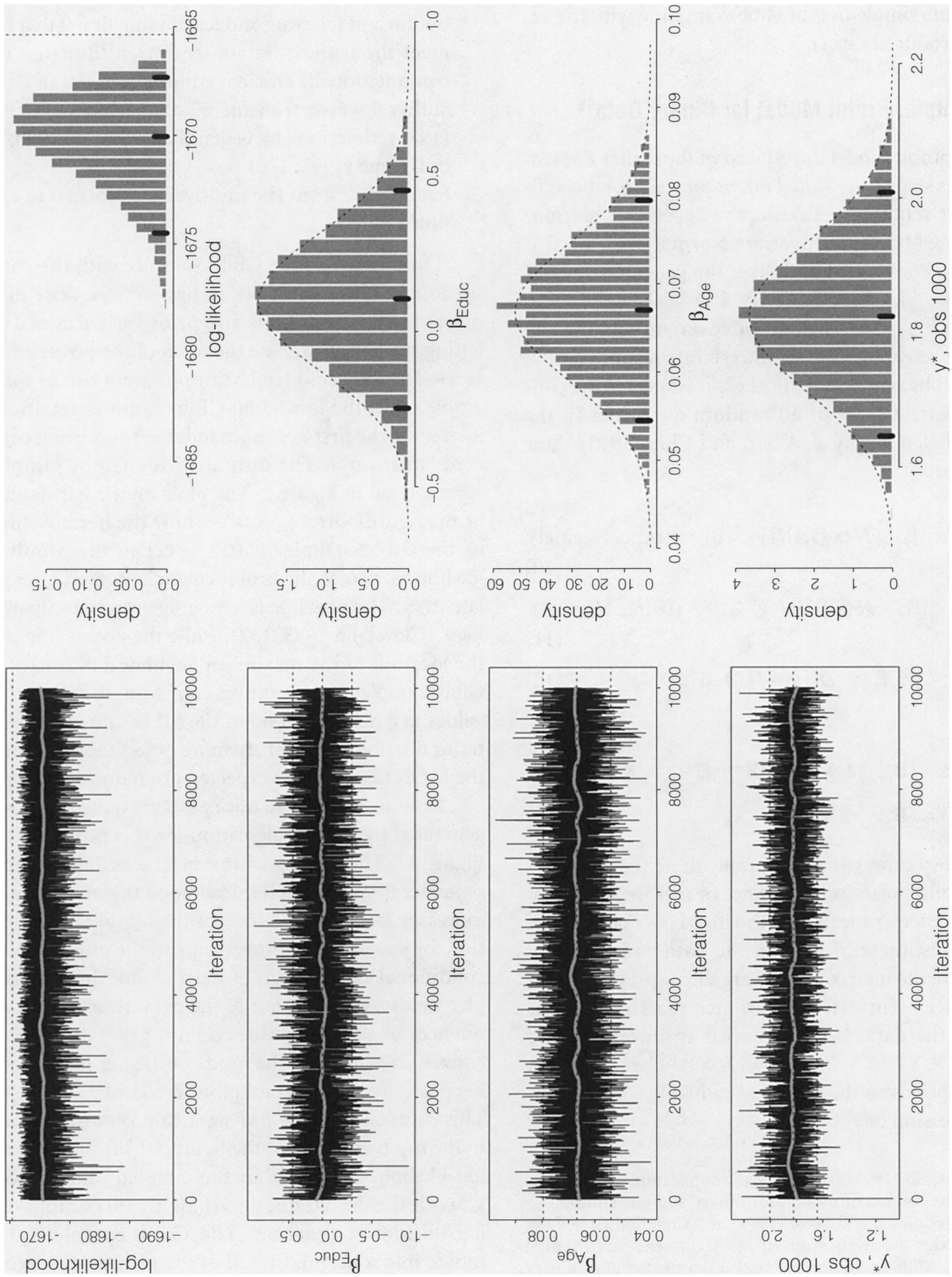
With these conditional distributions the Gibbs sampler (at iteration t) consists of the following steps:

1. Sample $y_i^{*(t)}$ from respective truncated Normals in Equations 10 and 11. This is easily accomplished either by sampling from (untruncated) Normal distribution and rejecting and resampling draws that fail to meet the truncation constraints, although a more computationally efficient strategy is to use the "probability inverse transform" algorithm of Devroye (1986), described by Gelfand and Smith (1990, 977) or Greene (1997, 179).
2. Sample $\boldsymbol{\beta}^{(t)}$ from the multivariate Normal in Equation 12.

I implemented the Gibbs sampler with the data on turnout considered in the earlier section. I use diffuse priors for $\boldsymbol{\beta}$ (zero means and prior variances of 1000), letting the data dominate the form of the posterior densities. Ten thousand Gibbs samples were run in this example, with the last 5,000 Gibbs samples retained for inference (the first five thousand iterations being considered "burn-in"). The output of the Gibbs sampler is summarized in Figure 2. The plots on the left-hand side of the figure—"trace plots"—show the iterative history of the Gibbs sampler with respect to the probit log-likelihood, two of the probit coefficients, and one of the latent y_i^*. The gray line is a moving average estimated by loess (Cleveland 1993, 152), while the dotted line marks the location of the maximum likelihood estimates. The Gibbs sampler quickly moves away from the OLS starting values to a neighborhood of the MLEs and recovers posterior distributions that are more or less those implied by the MLEs (as would be expected given my diffuse prior).

Note also that these trace plots are quite unlike those generated by the *EM* algorithm for this problem (recall Figure 1). The *EM* algorithm produces a deterministic sequence in the probit log-likelihood that monotonically increases towards the mode of the log-likelihood function, by successive improvements in the estimates of the conditional expectations of the $\mathbf{y}^*$. Since the Gibbs sampler yields a random tour of the parameter space, the sequences of sampled values do not generate monotone convergence towards the mode of the log-likelihood (or log posterior), as the trace plots in Figure 2 demonstrate. This contrast with the *EM* algorithm is most prominent in the top two panels of the figure, which summarize the log-likelihoods implied by the sampled values of $\boldsymbol{\beta}$ and $\mathbf{y}^*$, with the dotted lines marking the maximum of the log-likelihood function. The Gibbs sampler quickly moves into a neighborhood of the parameter space that supports the maximum, but never *exactly* attains it; this

FIGURE 2 Output of Gibbs Sampler, Probit Model of Voter Turnout



The left-hand panels show the iterative history of the Gibbs sampler for the designated quantities; the dotted lines indicate the location of the MLE and its 90 percent confidence interval, and the thicker grey line indicates a moving average (estimated by loess). The right hand panels show the posterior density of each quantity as a histogram, using the last 5,000 iterations of the Gibbs sampler; the dotted lines indicate the asymptotic Normal density implied by the MLEs; the tick marks on the horizontal axis indicate the 5th, 50th, and 95th percentiles of the Gibbs samples.

TABLE 2 Comparison of MLEs and Gibbs Sampler Output, Probit Model of Voter Turnout

	MLE	MCMC
Intercept	−2.32 (.56) [−3.24, −1.40]	−2.34 — [−3.26, −1.43]
Education	.096 (.22) [−.26, .45]	.11 — [−.25, .46]
Education^2	.021 (.022) [−.015, .057]	.020 — [−.016, .056]
Age	.067 (.008) [.054, .079]	.067 — [.054, .079]
Age^2	−.00047 (.00008) [−.00061, −.00034]	−.00047 — [−.00061, −.00034]
South	−.094 (.061) [−.19, .007]	−.095 — [−.19, .006]
Gubernatorial Election	.065 (.066) [−.044, .17]	.064 — [−.046, .17]
Closing Day	−.021 (.020) [−.053, .012]	−.020 — [−.053, .012]
Education × *Closing Day*	.0063 (.0081) [−.0071, .020]	.0061 — [−.0073, .020]
Education^2 × *Closing Day*	−.00061 (.00082) [−.0020, .00074]	−.00059 — [−.0020, .00076]

Standard errors appear in parentheses for the MLEs. For the Gibbs sampler output, the mean of the last 5,000 samples is reported as the point estimate, no standard error is reported, and a 90 percent confidence interval is reported in brackets; the 90 percent confidence interval implied by the MLEs point estimate and standard error (assuming asymptotic Normality) is also reported in brackets.

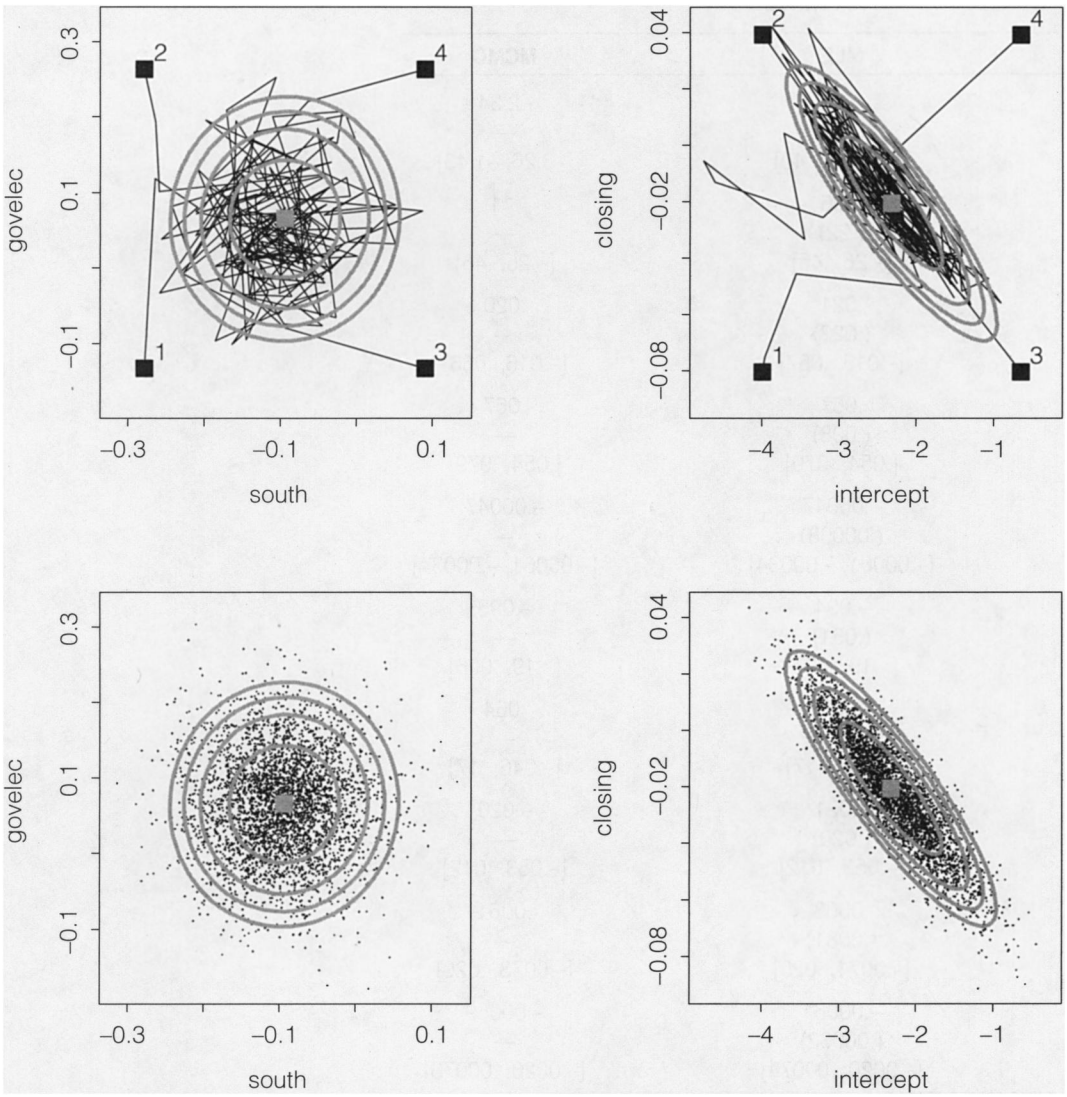
would only happen if the Gibbs sampler happened to sample $\boldsymbol{\beta} = \boldsymbol{\beta}_{\text{MLE}}$. Rather, the Gibbs sampler randomly explores the parameter space in a neighborhood of the MLEs,[11] with the occasional departure into regions with relatively low log-likelihood (i.e., the downward spikes in the log-likelihood trace plot).

Table 2 provides an additional comparison of the MLEs and the output of the Gibbs sampler for this example. I summarize the output of the Gibbs sampler with the mean of the last 5,000 samples of the 10,000 sample run, and use the interval from the 5th to the 95th percentiles of these samples as an estimate of the 90 percent confidence interval around each posterior mean. Clearly, the MLEs and the posterior means are extremely close to one another, and any differences between the two sets of point estimates are extremely small relative to the MLE standard errors, or the confidence intervals on the posterior means.

I also use this example to illustrate the Gelman and Rubin convergence diagnostic. Figure 3 show the two-dimensional trace plots that result from running four

[11]Strictly speaking the Gibbs sampler is exploring the parameter space in a neighborhood of the mode of the joint log-posterior density, but given my diffuse priors, this corresponds to that region of the parameter space supporting the maximum of the log-likelihood.

FIGURE 3 Parallel Gibbs Samplers, Probit Example



Four parallel Gibbs samplers were started from points ± 3 standard errors from the maximum likelihood estimates. The top two panels show the paths for the first fifty iterations, along with the maximum likelihood estimate (gray square) and likelihood contours. The lower two panels show the Gibbs samples from the last 1,000 iterations. Note the strong correlation between the intercept and the registration closing day coefficient.

parallel Gibbs samplers for 2,000 iterations each. For expository purposes, starting points were chosen by adding plus or minus three standard errors to the maximum likelihood estimates for the indicated parameters; obviously in a "real" application the MLEs may not be known. The Gibbs sampler quickly moves towards the MLEs in all cases, and the parallel traces quickly overlap one another, confirming that the Gibbs sampler converges quickly for this example.

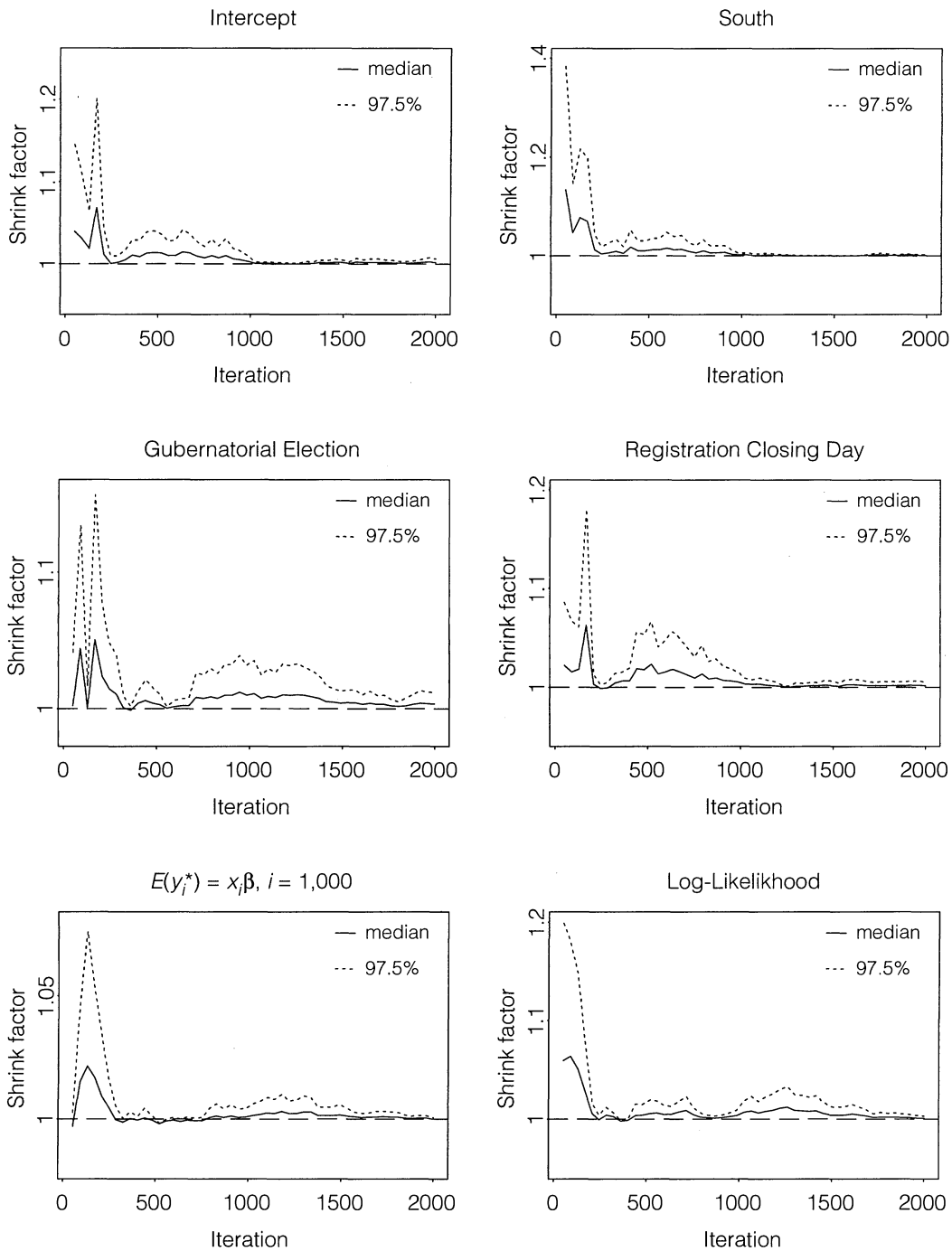
In Figure 4, I show the iterative history of the Gelman and Rubin (1992) test statistic, calculated using the output of the four chains with respect to the indicated quantities. The shrink factors are all comfortably within the ranges suggested by Gelman and Rubin after 1,000 iterations, suggesting that we may validly regard the MCMC sampler to have converged on the target distribution by this stage. For details on the generation of these diagnostic plots, see the Software section, below.

Example: Linear Regression with AR(1) Disturbances (Presidential Approval)[12]

Using the notation from page 374, the inferential problem here is to obtain the posterior density of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \rho)'$. As

[12]This example closely follows Chib's (1993) analysis of the more general AR(p) case. Chib and Greenberg (1994) consider the ARMA(p,q) case.

FIGURE 4 Gelman and Rubin Shrink Factors, Probit Example



The Gelman and Rubin test statistic is calculated over the course of the iterations and plotted as a trace plot. The shrink factors all quickly fall towards 1 for the indicated quantities, suggesting that the MCMC sequence has converged on the posterior density.

in the probit example, informative priors are not necessary, but can be readily implemented here (I choose to use diffuse priors). As is often the case in the Bayesian analysis of linear regression models, it is convenient[13] to factor the joint prior distribution for θ as $p(\beta, \sigma^2, \rho) = p(\beta|\sigma^2)p(\sigma^2)p(\rho)$, i.e., (β, σ^2) is *a priori* independent of

[13]This particular factorization of the prior on β and σ^2 is not necessary, but fairly standard in the Bayesian literature; see Western and Jackman (1994) for an alternative, simpler, parameterization. Here I follow the Normal-inverse-Gamma parameterization as it appears in Chib (1993).

the auto-regressive parameter ρ. Flexible specifications of each piece of the prior distribution are

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2 &\sim N(\boldsymbol{\beta}_0, \sigma^2 \mathbf{A}_0^{-1}), \\ \sigma^2 &\sim \text{Inverse-}\Gamma\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right), \\ \rho &\propto N(\rho_0, R_0^{-1}) I(\rho \in (-1,1)), \end{aligned} \tag{13}$$

i.e., a Normal-inverse-Gamma prior for $(\boldsymbol{\beta}, \sigma^2)$ and a truncated Normal prior on ρ so as to ensure stationarity.[14] Uninformative, diffuse priors can be set if the hyperparameters in Equation 13 are set to $\boldsymbol{\beta}_0 = 0, \mathbf{A}_0 - c\mathbf{I}_k$ (where c is an arbitrarily large, positive, scalar and $\mathbf{I}_k$ is an identity matrix of size k, the number of columns in $\mathbf{X}$), $\nu_0 = -k$, $\delta_0 = 0$, $\rho_o = 0$, and $R_0 = 0$.

Implementing the Gibbs sampler here involves iterating over the following steps:

1. Sample $\boldsymbol{\beta}^{(i+1)}$ from $p(\boldsymbol{\beta}|\sigma^{2(i)}, \rho^{(i)}, \mathbf{y}, \mathbf{X})$
2. Sample $\sigma^{2(i+1)}$ from $p(\sigma^2|\boldsymbol{\beta}^{(i+1)}, \rho^{(i)}, \mathbf{y}, \mathbf{X})$
3. Sample $\rho^{(i+1)}$ from $p(\rho|\boldsymbol{\beta}^{(i+1)}, \sigma^{2(i+1)}, \mathbf{y}, \mathbf{X})$.

In step 1, the conditioning is on the current estimates of σ^2 and ρ, so $\mathbf{y}$ and $\mathbf{X}$ can be transformed to form $\mathbf{y}^*$ and $\mathbf{X}^*$, respectively. This step makes the posterior density of $\boldsymbol{\beta}$ simple to obtain. The disturbances from the regression of $\mathbf{y}^*$ on $\mathbf{X}^*$ are distributed iid Normal with mean zero under the stated assumptions; given the multivariate Normal prior and a Normal likelihood, textbook results on the Bayesian analysis of the Normal regression model apply (e.g., Leamer 1978):

$$\boldsymbol{\beta}|\sigma^2, \rho, \mathbf{y}, \mathbf{X} \sim N(\tilde{\boldsymbol{\beta}}, \sigma^2 \tilde{\mathbf{A}}^{-1}), \tag{14}$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{A}_0 + \mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}(\mathbf{A}_0\boldsymbol{\beta}_0 + \mathbf{X}^{*\prime}\mathbf{y}^*)$ and $\tilde{\mathbf{A}} = (\mathbf{A}_0 + \mathbf{X}^{*\prime}\mathbf{X}^*)$. Sampling from this k dimensional Normal distribution is easily done.

Step 2 is also easy. The conditional distribution for σ^2 is

$$\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \rho \sim \text{Inverse-}\Gamma\left(\frac{n + \nu_0 + k}{2}, \frac{\delta_0 + Q_{\boldsymbol{\beta}} + \mathbf{e}^{*\prime}\mathbf{e}^*}{2}\right) \tag{15}$$

where $Q_{\boldsymbol{\beta}} = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\mathbf{A}_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ and $\mathbf{e}^* = (\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})$. Note that in the actual implementation of the Gibbs sampler, $\mathbf{b}$ in Equation 15 would be replaced by the sampled value from step 1.

[14] The parameters (or "hyperparameters") in the inverse-Γ over σ^2, ν_0 and δ_0, can be interpreted as a prior number of observations parameter and a prior sums of squares, respectively.

Step 3 is also easily implemented. Conditional on the data, $\boldsymbol{\beta}$, σ^2, and the stated assumptions, ρ has a normal marginal posterior density, subject to any truncations due if the stationarity assumption is to be strictly imposed. Net of the complication imposed by stationarity, the updating involved in this setup is simply that for the Bayesian analysis of a bivariate regression:

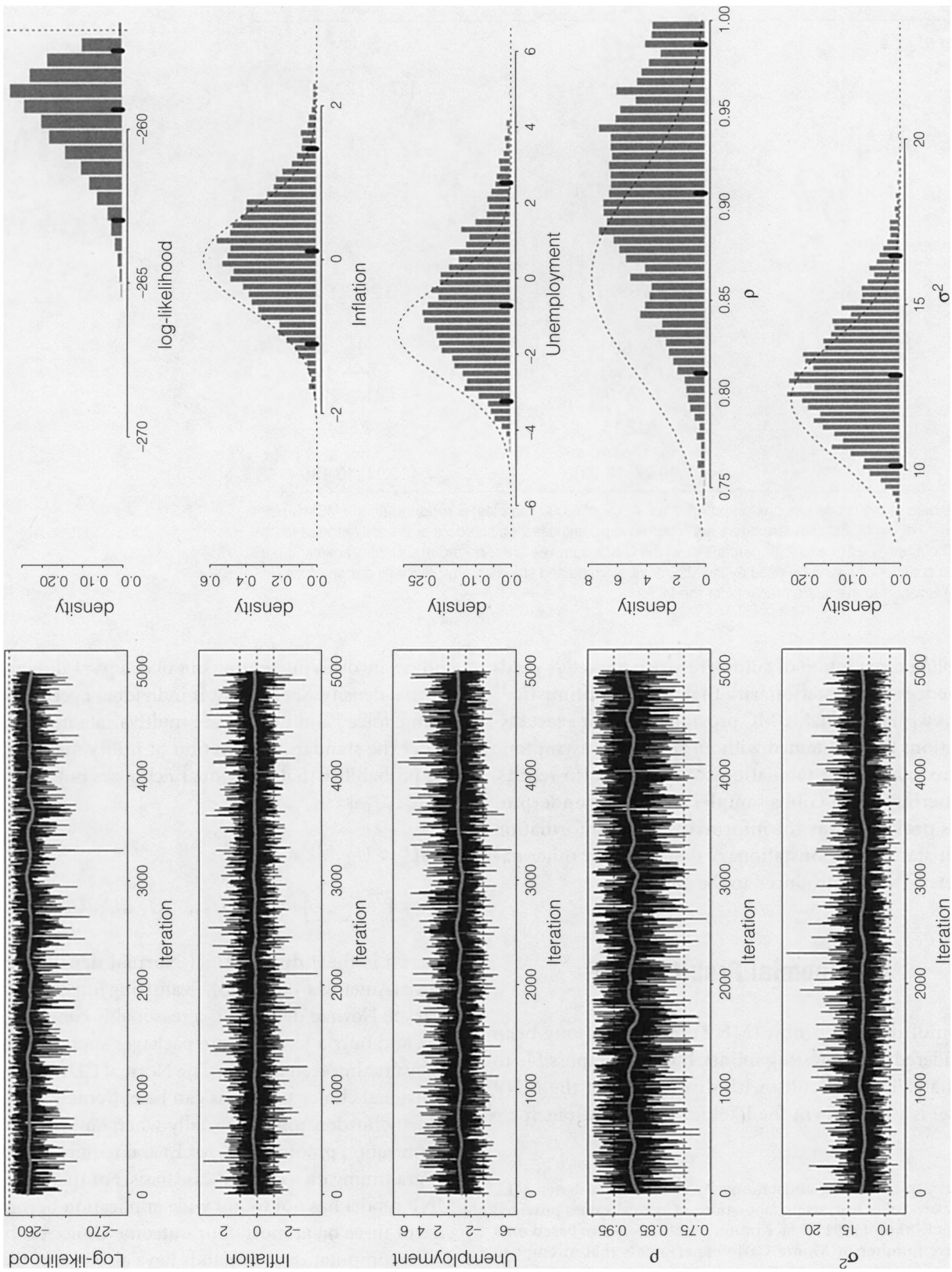
$$\rho|\boldsymbol{\beta}, \sigma^2, \mathbf{y}, \mathbf{X} \propto N(\tilde{\rho}, \tilde{R}^{-1}) I_{\rho \in (-1,1)}, \tag{16}$$

where $\tilde{\rho} = \tilde{R}^{-1}\left(R_0\rho_0 + \sigma^{-2}\sum_{t=2}^{T} u_t u_{t-1}\right)$ and $\tilde{R} = R_0 + \sigma^{-2}\sum_{t=2}^{T} u_t^2$. Draws from the normal density that lie outside the $(-1,1)$ interval are rejected, and the iterations continue with a draw within the interval. A useful check on stationarity is to note the proportion of draws that fail to meet this constraint.

I implemented the Gibbs sampler using the data on Reagan's monthly approval ratings introduced earlier. The Gibbs sampler uses noninformative priors, and 5,000 iterations, with the first 1,000 discarded as burn-in. Figure 5 summarizes the output of the Gibbs sampler and contrasts the asymptotic Normal approximations for the maximum-likelihood estimates. Table 3 also compares the MLEs with the output of the Gibbs sampler. With my diffuse priors, the output from the Gibbs sampler are reasonably close to the MLEs. Interesting differences arise due to the Gibbs sampler exploring the exact posterior density of ρ, which has much more probability mass between .9 and 1.0 than is implied by the (asymptotically-valid) Normal density associated with the MLEs. The Gibbs sampler attempted to sample $\rho > 1$ in less than 4 percent of the 5,000 iterations, and forcing the Gibbs sampler to sample ρ from the stationarity interval $(-1,1)$ helps account for the the skewed shape of the posterior for ρ. Notwithstanding this constraint, it is interesting to see how much probability mass piles up in the region between the MLE of ρ and 1.0, underlining just how poor the asymptotically-valid Normal approximation can be for autocorrelation parameters in finite samples. In turn, the unemployment parameter is nudged away from the MLEs, and we also obtain a slightly higher estimate of the white-noise variance σ^2 than the MLE.

This example highlights one of the strengths of MCMC methods: one can use Monte Carlo methods to obtain arbitrarily precise approximations to the posterior density of a quantity of interest (or the marginal log-likelihood, given a flat prior). For example, it is well known that the standard Normal approximation for the

FIGURE 5 Gibbs Sampler Output for Regression Model of Reagan Approval with AR(1) Disturbances



Trace plots appear in the left-hand panels; histograms using the last 4,000 iterations appear in the right-hand panels. See Figure 2 for further details.

TABLE 3 Comparison of MLEs and Gibbs Sampler Output, Regression Model of Reagan Approval with AR(I) Disturbances

	MLE	MCMC
Intercept	63.92	59.28
	(10.72)	—
	[46.29, 81.55]	[27.62, 80.45]
Inflation	−.0064	.11
	(.63)	—
	[−1.04, 1.03]	[−1.11, 1.44]
Unemployment	−1.38	−.72
	(1.31)	—
	[−3.53, .77]	[−3.24, 2.53]
ρ	.86	.91
	(.053)	—
	[.77, .95]	[.81, .99]
σ^2	12.15	12.86
	(1.75)	—
	[9.27, 15.03]	[10.11, 16.49]

The median of the Gibbs sampler output (the last 4,000 of 5,000 samples) is reported as the MCMC point estimate. For the MLEs, standard errors are reported in parentheses; no standard errors are reported for the MCMC output. The 5th and 95th percentiles of the Gibbs samples are reported in square brackets; the 95 percent confidence interval implied by the MLE point estimate and standard error (assuming asymptotic Normality) is reported in square brackets for the MLEs.

sampling distribution of autoregressive parameters yields inferences about stationarity that are too optimistic.[15] But as we have seen, MCMC provides a way for exact distributions to be obtained without relying on asymptotic approximations or tabulations of Monte Carlo results. Furthermore, the Gibbs sampler's Bayesian underpinnings provide a way for informative prior information about stationary/nonstationary dynamics (or other parameters) to be introduced to the analysis.

Multinomial Probit Model

The multinomial probit (MNP) model has long been considered an interesting model, but nigh impossible to estimate. The difficulty with direct MLE of the MNP model is well known; the likelihood calculations for p choices involves integrating out over a $p-1$ dimensional normal density. Specifically, if individual i gets utility U_{ij} from choice j and utilities are multivariate normal, then under the standard assumption of utility maximization the probability that individual i chooses outcome j, $j \in \{1, \ldots, J\}$ is

$$P(U_{ij} > U_{ik}, \forall k \neq j) = \int_{-\infty}^{\infty}\int_{-\infty}^{U_{ij}} \ldots \int_{-\infty}^{U_{ij}} f(U_1, U_2, \ldots, U_j) dU_1 dU_2 \ldots dU_j$$

where f is the J-dimensional Normal density defined above (Amemiya 1985, 308). Evaluating integrals of multivariate Normal densities is a reasonably complex problem, and only a few software packages support routines for evaluating even the bivariate Normal CDF. Higher dimensional choice problems can be extremely computationally burdensome, especially when embedded in an optimization problem (i.e., MLE) and require specialized programming on a case-by-case basis. For this reason the MNP model has not found wide application beyond the case of three or, at most, four outcome choice problems. These computational demands have also meant that the MNP model has been "off-limits" for the overwhelming majority of political scientists.

[15]Contrast diagnosing nonstationarity using normal-based MLE or Dickey-Fuller tests using tabulations of critical values provided by MacKinnon (1991). MacKinnon's tabulations are based on a massive number of Monte Carlo experiments that attempt to mimic different conditions likely to be encountered in applied settings; with MCMC researchers can obtain the posterior distribution of ρ conditional on *their* data.

MCMC makes the MNP problem much more tractable. Here I implement the MCMC sampler of McCulloch and Rossi (1994), as amended by Mc Culloch, Polson, and Rossi (1998). Assume individual i ($i = 1, \ldots, n$) chooses outcome $j \in \{0, \ldots, p-1\}$ if $z_{ij} \geq \max(\mathbf{z}_i)$, where $\mathbf{z}_i = \mathbf{R}_i\boldsymbol{\beta} + \mathbf{u}_i$, $\mathbf{u}_i \sim N(0, \mathbf{V})$, $\mathbf{R}_i$ is a p by k matrix of regressors, and $\boldsymbol{\beta}$ a k by 1 vector of parameters to be estimated. Here regressors vary over choices and individuals, but the effects of each regressor on utilities are constant over choices (and individuals). Alternative specifications can be easily handled. The analyst does not observe the latent utilities in $\mathbf{z}_i$, but does observe the choices made by each individual. Given the assumption of utility maximization, this is equivalent to observing the index of the largest element in each $\mathbf{z}_i$.

Identification

As is well known, the MNP model is unidentified as stated, since any location shift of the latent utilities or scale shift are observationally equivalent. To solve the location invariance problem, difference p – 1 of the choice utilities with respect to a baseline choice by defining $\mathbf{w}_i = (w_{i1}, \ldots, w_{i,p-1})'$ with $w_{ij} = z_{ij} - z_{ip}$. Similarly define $\mathbf{X}_i$ as a $(p-1)$ by k matrix of observations on independent variables, obtained by subtracting the pth row from the first $p-1$ rows of $\mathbf{R}_i$. These normalizations yield a $p-1$ dimensional model:

$$\mathbf{w}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_i \sim N(0,\Sigma), \forall\ i \tag{17}$$

$$y_i = \begin{cases} 0 & \text{if } \max(\mathbf{w}_i) < 0 \\ j & \text{if } \max(\mathbf{w}_i) = w_{ij} > 0 \end{cases} \tag{18}$$

Scale invariance is also a problem. For any for any constant $c > 0$, $\tilde{y}_i = h(c\mathbf{w}_i)$ is observationally indistinguishable from $y_i = h(\mathbf{w}_i)$, where $h(\,)$ is the assignment rule in Equation 18. Put differently, the distribution of $\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \Sigma$ is the same as the distribution of $\mathbf{y}|\mathbf{X}, c\boldsymbol{\beta}, c^2\Sigma$ (e.g., McCulloch and Rossi 1994, 209). Defining σ_{ij} as the ij-th element of Σ, then a common solution to the scale invariance problem is to set σ_{11} to 1.

Gibbs Sampling Algorithm

The inferential problem here is to find the posterior density $p(\boldsymbol{\beta}, \Sigma|\mathbf{y}, \mathbf{X})$. McCulloch and Rossi (1994) and Chib and Greenberg (1997) use Gibbs sampling in this context, exploiting the fact that conditional on the latent random utilities the problem reduces to a fairly standard Bayesian multivariate regression model. In this way the troublesome (if not impossible) integrations involved in calculating the $\Pr(y_{ij} = 1)$ are avoided. The Gibbs sampler requires the following conditional distributions:

$$\boldsymbol{\beta} \mid \Sigma, \mathbf{W}, \mathbf{X}$$
$$\Sigma \mid \boldsymbol{\beta}, \mathbf{W}, \mathbf{X}$$
$$\mathbf{W} \mid \boldsymbol{\beta}, \Sigma, \mathbf{X}$$

Note that W is the n by $p-1$ matrix of latent utilities, and Σ is the $p-1$ by $p-1$ variance-covariance matrix for the $\boldsymbol{\epsilon}_i$. I consider each of these conditional distributions in turn.

Let the prior for $\boldsymbol{\beta}$ be $N(\bar{\boldsymbol{\beta}}, \mathrm{A}^{-1})$. Then the conditional distribution for $\boldsymbol{\beta}$ is $N(\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\mathrm{X}'\mathrm{G}\mathrm{X} + \mathrm{A})^{-1}$, $\hat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(\mathrm{X}'\mathrm{G}\mathrm{W} + \mathrm{A}\bar{\boldsymbol{\beta}})$ and $\mathrm{G} = \Sigma^{-1} \otimes \mathbf{I}_n$. That is, with a diffuse prior we are essentially estimating a system of seemingly unrelated regressions (e.g., Judge et al. 1988, 450).

The prior and the conditional distribution for Σ is complicated by the identifying constraint $\sigma_{11} = 1$. McCulloch, Polson, and Rossi (1998) note that usual approaches for specifying priors over covariance matrices are not appropriate in this instance and propose an alternative approach. Partition $\boldsymbol{\epsilon}_i$ as $(\nu_i, \boldsymbol{\omega}_i)$, where $\nu_i = \varepsilon_{i1}$ and $\boldsymbol{\omega}_i = (\varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{i,p-1})'$. Obviously the joint distribution of $(\nu, \boldsymbol{\omega})$ is the joint distribution of $\boldsymbol{\epsilon}$, which is $N(\mathbf{0}, \Sigma)$ (suppressing the i subscript). McCulloch, Polson, and Rossi's (1998) contribution is to note that the joint distribution of $(\nu, \boldsymbol{\omega})$ can be factored as the marginal distribution of ν and the conditional distribution $\boldsymbol{\omega}|\nu$. If $\boldsymbol{\gamma} = E(\nu\boldsymbol{\omega})$ ($a\ p-2$ column vector) and $\Sigma_{\omega} = E(\boldsymbol{\omega}\boldsymbol{\omega}')$ (a $p-2$ by $p-2$ matrix), then $\nu \sim N(0, \sigma_{11})$ and $\boldsymbol{\omega}|\nu \sim N((\boldsymbol{\gamma}/\sigma_{11})\nu, \Sigma_{\omega} - \boldsymbol{\gamma}\boldsymbol{\gamma}'/\sigma_{11})$. Now let $\boldsymbol{\Phi} = \Sigma_{\omega} - \boldsymbol{\gamma}\boldsymbol{\gamma}'/\sigma_{11}$. Just as there is a correspondence between and $\boldsymbol{\epsilon}$ and ν and $\boldsymbol{\omega}$, there is a correspondence between Σ and $(\sigma_{11}, \boldsymbol{\gamma}, \boldsymbol{\Phi})$:

$$\Sigma = \begin{bmatrix} 1 & \boldsymbol{\gamma}' \\ \boldsymbol{\gamma} & \boldsymbol{\Phi} + \boldsymbol{\gamma}\boldsymbol{\gamma}' \end{bmatrix}, \tag{19}$$

recalling the identification constraint $\sigma_{11} = 1$. A prior for $\{\Sigma|\sigma_{11} = 1\}$ is now given by priors on $\boldsymbol{\gamma}$ and $\boldsymbol{\Phi}$:

$$\boldsymbol{\gamma} \sim N(\bar{\boldsymbol{\gamma}}, \mathrm{B})$$
$$\boldsymbol{\Phi}^{-1} \sim \text{Wishart}\ (\kappa, \mathrm{C})$$

McCulloch, Polson, and Rossi (1998) suggest specifications for these prior parameters that yield relatively diffuse prior distributions.

With these priors it is reasonably straightforward to obtain expressions for the conditional distributions for $\boldsymbol{\gamma}$ and $\boldsymbol{\Phi}$. At a given iteration of the Gibbs sampler we sample from these conditional distributions and combine

the sampled $\boldsymbol{\gamma}$ and $\boldsymbol{\Phi}$ as in Equation 19 to give a sample from the conditional distribution for $\{\Sigma|\sigma_{11} = 1\}$.

Given $\boldsymbol{\beta}^{(t)}$ and $\mathrm{W}^{(t)}$ (at iteration t of the Gibbs sampler) we also have $\boldsymbol{\epsilon}_i^{(t)}$ and thus $(\nu_i^{(t)}, \boldsymbol{\omega}_i^{(t)})$. The quantities $\boldsymbol{\gamma}$ and $\boldsymbol{\Phi}$ are the parameters obtained from a (multivariate) regression of $\boldsymbol{\omega}$ on ν; at iteration t this regression is

$$\boldsymbol{\omega}_i^{(t)} = \nu_i^{(t)}\boldsymbol{\gamma}' + \boldsymbol{\eta}_i, \boldsymbol{\eta}_i \sim N(0, \boldsymbol{\Phi}).$$

Given an estimate of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}$, and the Wishart prior, the conditional distribution for $\boldsymbol{\Phi}^{-1}$ is

$$\text{Wishart}\,(\kappa + n, \mathrm{C} + (\boldsymbol{\omega}^{(t)} - \boldsymbol{\nu}^{(t)}\hat{\boldsymbol{\gamma}}')'(\boldsymbol{\omega}^{(t)} - \boldsymbol{\nu}^{(t)}\hat{\boldsymbol{\gamma}}'))$$

from which we can sample $\boldsymbol{\Phi}^{(t)}$. Now given $\boldsymbol{\Phi}^{(t)}$, McCulloch, Polson, and Rossi (1998, 19) show that the conditional distribution for $\boldsymbol{\gamma}$ is

$$N(\mathrm{A}_{\boldsymbol{\gamma}}\,(\mathrm{vec}(\boldsymbol{\Phi}^{-1}\boldsymbol{\omega}'\boldsymbol{\nu}) + \mathrm{B}\boldsymbol{\gamma}), \mathrm{A}_{\boldsymbol{\gamma}}),$$

where $\mathrm{A}_{\boldsymbol{\gamma}} = (\boldsymbol{\nu}\text{¢}\boldsymbol{\nu}\boldsymbol{\Phi}^{-1} + \mathrm{B})^{-1}$, $\boldsymbol{\nu} = \boldsymbol{\nu}^{(t)}$ (a n by 1 column vector containing $\boldsymbol{\epsilon}_1^{(t)}$, $\boldsymbol{\omega} = \boldsymbol{\omega}^{(t)}$ (a n by $p-2$ matrix containing $\boldsymbol{\epsilon}_2^{(t)}, \boldsymbol{\epsilon}_3^{(t)}, \ldots, \boldsymbol{\epsilon}_{p-1}^{(t)}$, and $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{(t)}$ (a $p-2$ by $p-2$ matrix). Sampling from this distribution yields $\boldsymbol{\gamma}^{(t)}$.

The sampled quantities $\boldsymbol{\gamma}^{(t)}$ and $\boldsymbol{\Phi}^{(t)}$ define $\Sigma^{(t)}$, following Equation 19. In short, the problem of sampling from the conditional distribution for $\{\Sigma|\sigma_{11} = 1\}$ has been broken down into two sampling problems.

Finally, the conditional distribution of w_i is a truncated $(p-1)$-dimensional Normal, where the truncation points follow from the fact that if the jth choice is observed for individual i (i.e., $y_{ij} = 1$) then $w_{ij} > \max(\mathrm{w}_{i,-j}, 0)$. Conversely, if the jth choice is not observed for individual i (i.e., $y_{ij} = 0$) then $w_{ij} < \max(\mathrm{w}_{i,-j}, 0)$, where $\mathrm{w}_{i,-j}$ is the $p-2$ vector of elements of w_i excluding w_{ij}. Following Albert and Chib (1993) I sample repeatedly from the (untruncated) $(p-1)$-variate normal until a draw satisfies the constraint implied by the observed choice.

Advantages over MLE

Just a handful of MNP applications have been presented in the political science literature. Directly attacking anything beyond a three choice problem is simply infeasible using maximum likelihood estimation, and even for the three choice problem, it is difficult to obtain precise estimates of the off-diagonal elements of Σ.

For instance, McCulloch and Rossi (1994) report a series of experiments comparing the sampling distributions of MLE estimates of a three-dimensional problem with those produced by their Gibbs sampling approach. The findings of this exercise are striking. With as many as 1000 observations *per parameter* there is pronounced skewness in the sampling distributions of the error variance-covariance parameters, suggesting that "asymptotic theory may be of little use for the MNP model" (MCulloch and Rossi 1994, 219). That is, attacking the MNP model with MLE is not just difficult, but relying on asymptotic normality in making inferences about these error-variance and covariance parameters probably involves a huge leap of faith. Part of the problem here is due to the normalization employed to identify the MNP model, which means that bounded functions of variance parameters are actually estimated, such as variance ratios and correlations. Given that there is not much information about these variance and covariance parameters even in a large sample, it is the "boundedness" of the estimated parameters that stops asymptotic normality from kicking in (McCulloch and Rossi 1994, 221–222). These problems are avoided with the Gibbs sampler's arbitrarily precise approximations to the posterior densities.

Application: Vote Choice in the 1992 U.S. Presidential Election

I reanalyze the Bush-Clinton-Perot 1992 vote choice problem with a MNP model. Previous attempts to estimate this three-way choice problem using MNP have met with only limited success (e.g., Alvarez and Nagler 1995; Lacy and Burden 1999). I employ the identification constraints sketched above, with the Perot choice (P) as the "baseline" outcome. Thus Σ reduces to a 2×2 matrix, and with σ_{11} set to 1, there is just one variance and a covariance to estimate (σ_{22} and σ_{12}, respectively). For the Bush (B) and Clinton (C) outcomes, I also employ a model with a simple set of covariates:

$$\begin{aligned} U_{iC} - U_{iP} &= \boldsymbol{\beta}_{10}[\text{Common Covariates}]_i' + \boldsymbol{\beta}_{11}\text{DemPID}_i \\ &\quad + \boldsymbol{\beta}_{12}\,|\,\text{Clinton}_i - R_i| + \varepsilon_{iC} \\ U_{iB} - U_{iP} &= \boldsymbol{\beta}_{20}[\text{Common Covariates}]_i' + \boldsymbol{\beta}_{21}\text{RepPID}_i \\ &\quad + \boldsymbol{\beta}_{22}\,|\,\text{Bush}_i - R_i| + \varepsilon_{iB} \end{aligned}$$

$$\begin{bmatrix} \varepsilon_{iC} \\ \varepsilon_{iB} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right).$$

where R_i is the respondent's self-location on a unidimensional measure of ideology, Bush_i and Clinton_i are measures of the respondents' estimate of the ideological location of Bush and Clinton, respectively, and DemPID and RepPID are indicators for Democratic and Republican

party identification, respectively. Note that the ideological distance and partisanship measures do not appear in each equation, corresponding to parameter restrictions suggested by Keane (1992) as being necessary to avoid what he describes as the "fragile identification" problem in MNP models. The "common covariates" are intercepts for each equation, a binary indicator for whether the respondent felt the national economy was worse, a binary indicator for whether the respondent opposed government-provided health care, and a binary indicator coded 1 for female respondents, and 0 otherwise. Data come from the 1992 National Election Studies, as used by Alvarez and Nagler (1995).

For $\boldsymbol{\beta}$ I use a diffuse $N(\mathbf{0}, 1000 \cdot \mathbf{I})$ prior. For the error variances, recall that for the three choice problem I have here Σ reduces to a 2×2 matrix and so $\boldsymbol{\gamma}$ and $\boldsymbol{\Phi}$ reduce to scalars. Following McCulloch, Polson, and Rossi (1998), my prior for γ is relatively tight $N(0,.25)$, and my prior for Φ is scaled inverse-χ^2, with degrees of freedom parameter $\kappa = 20$, and scale parameter $C = 13.5$. Of more direct substantive interpretation is what these choices imply for the prior on the unrestricted variance parameter $\sigma_{22} = \Phi + \gamma^2$ and the correlation $\rho = \gamma / \sqrt{\Phi + \gamma^2}$. These choices result in a prior over the correlation parameter with a mode at 0, but with probability mass relatively uniform between –.5 and .5, and falling off towards –1 and 1. For the unrestricted variance σ_{22} these choices imply a reasonably tight χ^2-shaped prior, with a mode at around 1.0, but with virtually zero probability mass above 3.0. In this way I am being reasonably agnostic as to the size of the error variance for the Bush-Perot utility comparison, relative to the fixed value of 1.0 for the error variance in the Clinton-Perot utility comparison.

The Gibbs sampling scheme described above was run for 100,000 iterations. For the regression parameters, the sampled values appear to randomly oscillate around their posterior modes, while the two free elements of Σ exhibit some over-time dependencies in their iterative histories. That is, the Gibbs sampler is efficiently exploring the parameter space supporting the posterior for $\boldsymbol{\beta}$, but slowly meanders through the parameter space for $\sigma_{12} = \gamma$ and $\sigma_{22} = \Phi + \gamma\gamma'$. This is not fatal, but suggests that a large number of Gibbs samples is required to ensure that the posterior densities for these parameters are being thoroughly explored.

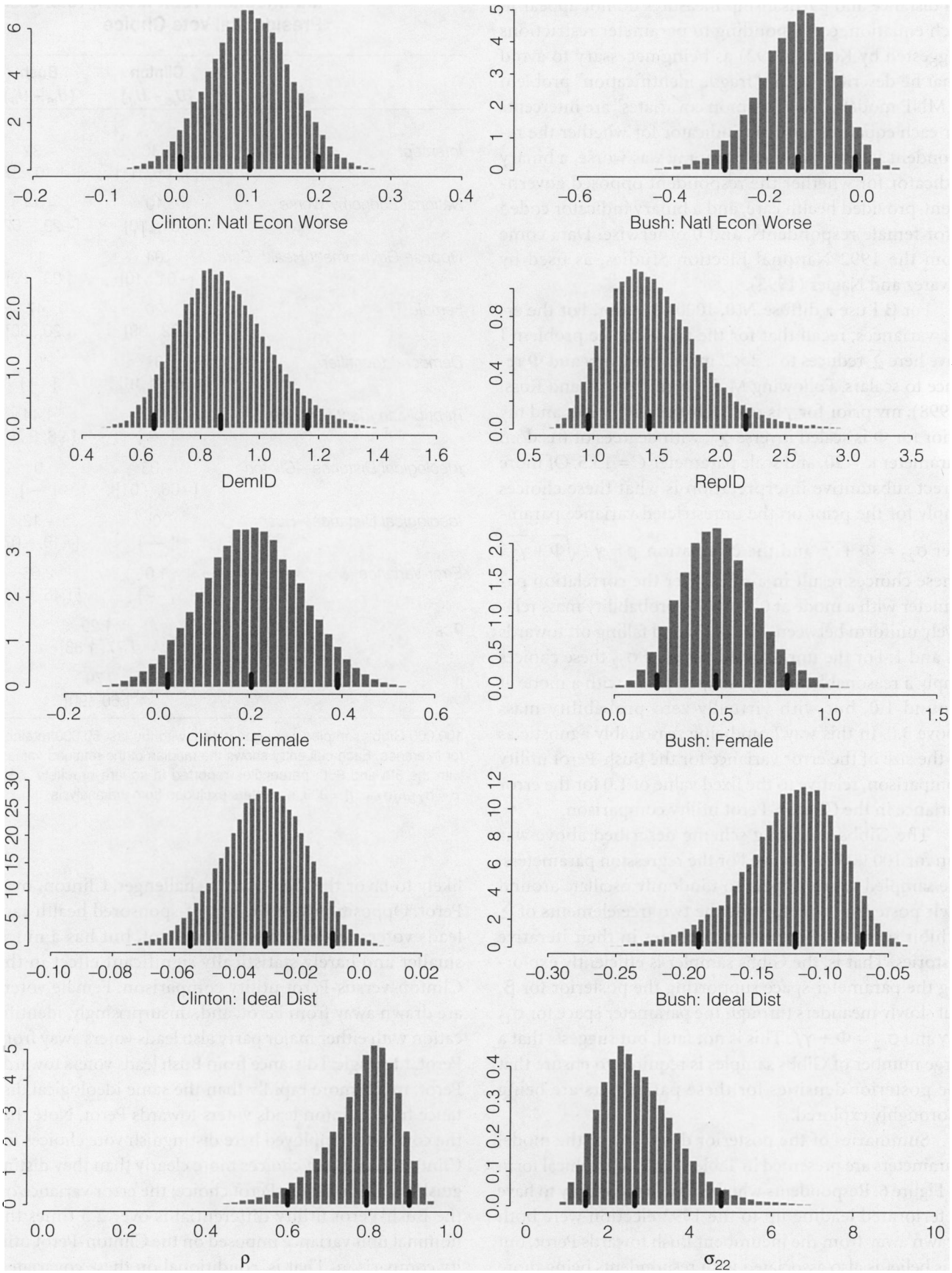
Summaries of the posterior densities for the model parameters are presented in Table 4, and in graphical form in Figure 6. Respondents who believe the economy to have deteriorated leading up to the 1992 election were both drawn away from the incumbent Bush towards Perot, but this belief is also associated with respondents being more likely to favor the Democratic challenger, Clinton, over Perot. Opposition to government-sponsored health care leads voters towards Bush over Perot, but has a much smaller and barely statistically significant effect in the Clinton-versus-Perot utility comparison. Female voters are drawn away from Perot, and, unsurprisingly, identification with either major party also leads voters away from Perot. Ideological distance from Bush leads voters towards Perot, much more rapidly than the same ideological distance from Clinton leads voters towards Perot. Note that the covariates employed here distinguish vote choices for Clinton from Perot choices more clearly than they distinguish the Bush-versus-Perot choice; the error variance on the Bush-Perot utility differential is over 2.5 times the nominal unit variance imposed on the Clinton-Perot utility comparison. That is, conditional on these covariates,

TABLE 4 Summary of Posterior Densities, Multinomial Probit Model, 1992 U.S. Presidential Vote Choice

	Clinton ($U_{iC} - U_{iP}$)	Bush ($U_{iB} - U_{iP}$)
Intercept	–.46 [–1.06, .11]	–.32 [–1.10, .37]
National Economy Worse	.10 [.01, .20]	–.14 [–.29, –.02]
Oppose Government Health Care	.04 [–.01, .10]	.11 [.03, .20]
Female	.20 [.02, .39]	.47 [.20, .80]
Democrat Identifier	.87 [.64, 1.16]	0 [—]
Republican Identifier	0 [—]	1.44 [.96, 2.20]
Ideological Distance—Clinton	–.03 [–.06, –.01]	0 [—]
Ideological Distance—Bush	0 [—]	–.12 [–.19, –.07]
Error Variance	1.0 [—]	2.66 [1.45, 4.78]
σ_{CB}	1.29 [.77, 1.88]	
ρ	.79 [.60, .90]	

100,000 Gibbs samples were generated, with the last 50,000 retained for inference. Each cell entry shows the median of the retained values, with the 5th and 95th percentiles reported in square brackets. n.b., $\rho = \sigma_{CB}/\sqrt{\sigma_{BB}}$; n = 909, nonvoters excluded from the analysis.

FIGURE 6 Density Estimates, Output of Gibbs Sampler for the MNP Problem



The density estimates summarize the output of the last half of the 100,000 Gibbs samples. The tick marks on the horizontal axis indicate the location of the 5th, 50th, and 95th percentiles.

the Bush-Perot alternatives are less distinguishable than the Clinton-Perot alternatives, which sits well with many understandings of the 1992 election. This is in part cause and consequence of the way ideological distance works differently in the two utility comparisons: small perceived ideological differences with Bush were associated with large impacts on the Bush-Perot utility comparison (relative to the Clinton-Perot comparison), suggesting that voters were drawing fine distinctions in distinguishing the two candidates, or at least relative to the ideological distinctions made in distinguishing Clinton and Perot. Of course, a more complete explanation of the politics of 1992 would include abstention as an option, but since my primary goal here is methodological exposition, I reserve that analysis for another paper.

Other Political Science Applications

Finally, I briefly list some published political science applications of MCMC, several of which have appeared in the pages of the *AJPS*.

I already mentioned King, Rosen, and Tanner's (1999) use of Metropolis-Hastings methods as an example of a hierarchical model. Western (1998) uses the Gibbs sampler to estimate a two-level hierarchical model of GDP in the OECD with pooled cross-sectional time-series data; the hierarchical structure of the model makes the country-specific effects of covariates conditional on (time-invariant and country-specific) levels of labor organization. Hierarchical models are an obvious and frequent area of application for MCMC; I refer readers to the summaries in Western (1998) and King, Rosen, and Tanner (1999) and the treatments in the statistical literature: e.g., Gelman et al. (1995, chapters 5, 13–14), Carlin and Louis (1996, chapter 7), and the book length treatment by Draper (1999).

Smith (1999) uses MCMC to estimate an ordered-probit model in a strategic choice setting (crisis escalation in international relations), subject to the assumption that observed crisis behavior (e.g., escalation and acquiescence) provides strategically censored insights as to each player's true underlying level of resolve. Smith terms this model a Strategically Censored Discrete Choice model.

Quinn, Martin, and Whitford (1999) use MCMC to estimate multinomial probit models for vote choices in both the United Kingdom and the Netherlands, using a slightly different parameterization to the one I employ here, due to Chib and Greenberg (1997). They compare the MNP model to the multinomial logit model and conclude that the choice as to which model may not simply be a theoretical concern, but will depend on the structure of the choice problem faced by voters, and the types of covariates available to the analyst.

Gelman and King (1990) is an early application of the Gibbs sampler and the first application of MCMC in a political science setting.[16] They use a three-component mixture model to decompose the distribution of vote shares across districts, treating the district-specific mixing probabilities as missing data to be estimated using either an *EM* algorithm or a Gibbs sampler. This approach allows Gelman and King to estimate seats-votes curves and quantities such as the bias and responsiveness of an electoral system election-by-election, and without covariates. Jackman (1994) generalizes this model and the methodology for cases where malapportionment is an issue (e.g., Australian jurisdictions).

Lastly, an interesting use of MCMC is in robust statistics. Data with outliers can often be conveniently modeled using t distributions with low degrees of freedom, $\nu > 2$ (recall that the t distribution tends towards the Normal as the degrees of freedom increases). MCMC allows the degrees of freedom parameter to be estimated conditional on the data, rather than (arbitrarily) set in advance by the analyst (e.g., Gelman et al. 1995, 357 and following). This approach is discussed by King and Katz (1999), in their model of vote shares in British House of Commons constituencies (using multivariate t distributions with unknown degrees of freedom parameter).

Software

Since MCMC methods are computationally intensive, no survey is complete without a discussion of software. *BUGS* — Bayesian inference Using Gibbs Sampling—is a general purpose package for Gibbs sampling, written by Spiegelhalter et al. (1997). The software relies on two key features: (1) a parser, that reads model statements input by the user, deduces the form of the model, and compiles code for sampling from the conditional distributions needed to implement the Gibbs sampler; (2) an algorithm for adaptive rejection sampling (Gilks 1992), that permits sampling from log-concave conditional distributions, thus allowing the software to work in a wide class of situations (e.g., Gilks and Wild 1992; Spiegelhalter et al. 1997, Table 3).

An extremely useful feature of *BUGS* is that it handles missing data "on-the-fly." Recall that from the perspective

[16] Indeed, this application predates the introduction of the terms "MCMC" or "Gibbs sampler" into the statistical mainstream, let alone the political science mainstream.

of the Gibbs sampler, missing data is simply another random quantity embedded in the model; if we have a conditional distribution for the missing quantities then we can apply either the Gibbs sampler (sampling directly from the conditional distribution) or a Metropolis-Hasting method (sampling from a candidate density). In many situations, the form of the conditional distribution

$$p(\text{missing data} \mid \text{observed data, parameters})$$

will be an explicit part of the model; this is always true in a regression-type setting when the dependent variable y contains missing data, since $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ is required for specifying the likelihood of the observed data, quite apart from dealing with any missing data problem. Accordingly, when encountering missing data on a dependent variable in a regression-type setting, *BUGS* will make "multiple imputations" automatically, with no special flagging of the problem required by the user.

BUGS is currently free and can be downloaded from http://www.mrc-bsu.cam.ac.uk/bugs. A Windows version, *WinBUGS*, lets users specify models via a graphical user-interface, drastically reducing the programming skill-level required in order to exploit MCMC. An extensive set of well-documented examples accompany the software, largely drawn from biostatistics. A set of Splus functions, *BOA*—**B**ayesian **O**utput **A**nalysis program—provides trace plots, summary statistics, and convergence diagnostics via a menu driven interface and is also freely available, from http://www.pmeh.uiowa.edu/BOA. A web-based summary of covergence diagnostics and software is at http://www.ensae.fr/crest/statistique/robert/McDiag/.

In the Appendix I present *WinBUGS* code for the binary probit example I presented in the Examples section. All other examples and graphs are generated using problem-specific code in *Splus* or *GAUSS*; all code, data, and supporting documentation is available from the *AJPS* web site (http://psweb.shs.ohio-state.edu/ajps/) as well as the author's web site http://tamarama.stanford.edu/simon.

Conclusion

At the time of writing, the social sciences stand poised to exploit the power of MCMC. A number of impediments stand in the way, which are steadily being overcome. First, by their very nature, MCMC methods are computationally intensive, relying on random sampling from conditional distributions to generate a random tour of the parameter space for all random quantities (parameters and missing data). Second, the desirable statistical properties of MCMC are asymptotic: subject to some regularity conditions, the approximation to the posterior density improves as the number of iterations increases, though exactly how many iterations is "enough" is very difficult to pin down in advance (recall my advice above that "more is better" in this regard). All this means that for problems with lots of (interdependent) parameters or data points (or both), the computational burdens of MCMC are still not trivial by the standards of the late 1990s. Offsetting this is the fact that (a) MCMC is the subject of a tremendous amount of statistical research, looking for ways to speed convergence; (b) computing resources for social-scientists continue to get faster and cheaper.

Prior to the release of *BUGS* and *WinBUGS*, MCMC was a do-it-yourself affair, requiring a high level of statistical and programming expertise, often requiring the user to get "close to the machine" and code problem-specific routines in *C* or *Fortran*. The advice of one prominent Bayesian statistician to me in 1995 was that acquiring skills in these programming languages is the entry price one has to pay in order to use MCMC; what one loses in programming time and debugging, one hopes to make up with faster execution and more iterations per unit of time. Given that the desirable properties of MCMC are asymptotic, the ability to generate a large number of iterations is paramount, and so in many cases the time spent developing code in *C* or *Fortran* could be well spent.[17] The arrival and ongoing development of *BUGS* and especially *WinBUGS* has largely made the statistician's advice redundant and goes a long way towards putting MCMC in the toolkit of methodologically literate social scientists.

Finally, I remind readers that for purposes of exposition, I have demonstrated how the methods work with a simple set of examples relatively familiar to social scientists, before turning to the more interesting multinomial

[17]For truly pathbreaking work, coding in a high-level language is probably inescapable: the problem may not be able to be leveraged into *BUGS* (e.g., the multinomial probit model I estimate), the particular combination of priors and data yield nonstandard posteriors and hence the use of Metropolis-Hastings methods, or the properties of the problem-specific Markov chain may not be well understood in advance. In these situations, hundreds of thousands of iterations of the MCMC sampler may be required, if not more. For instance, the King, Rosen, and Tanner (1999) treatment of the ecological inference problem has all these features and was implemented using *Fortran* on an extremely powerful *UNIX* workstation. Perhaps all this is to say that pathbreaking work is hard and that not everyone will do it!

probit model. The simplicity of my expository examples should not deceive readers: MCMC methods are not simply computer-intensive mechanisms for replicating maximum likelihood estimates! Rather, MCMC allows us to employ models previously considered impossible to estimate. The applications I cite throughout this paper span the frontiers of political methodology: e.g., the multinomial probit model, solutions to the ecological inference problem, hierarchical or "multi-level" models, or imputations for missing data. Without doubt, MCMC is opening up new methodological terrain to social scientists.

As with any new methodological technique, mistakes will be made. And, as always, there will be no substitute for working through a model parameterization or conducting exploratory data analysis, before throwing MCMC at a problem. Subject to this caveat, I anticipate that the impact of MCMC methods on quantitative social science in the next ten years will be as impressive as their impact in statistics over the last ten years.

Manuscript submitted December 3, 1998.
Final manuscript received August 17, 1999.

Appendix
WinBugs Programs for Turnout Example

These programs and accompanying files containing data and priors are available from http://psweb.sbs.ohio-state.edu/ajps/ and http://tamarama.stanford.edu/simon. The usefulness of *WinBUGS* is apparent by comparing these programs with the substantially longer programs necessary to implement the required Gibbs samplers in Splus, also available from my web site.

The following program implements a logit model for the turnout data. The last line of the program specifies the prior for $\boldsymbol{\beta}$; I employ a vague prior by setting the mean vector $b0$ to a null vector and the precision matrix B to .001 times an identity matrix. Missing data on the binary response y would present no problem in this instance; *WinBUGS* would simply sample to create multiple imputations for any missing values encountered. Note also the calculation of the log-likelihood; this is not necessary in order to implement the model, but demonstrates that quantities of interest can be calculated, monitored, and output by the program.

```
#############################################################################
## turnout model, by logit,                                                ##
## uses random 3,000 obs subset of Nagler's original data set              ##
##                                                                         ##
## simon jackman, dept of political science, stanford university           ##
#############################################################################
model{
  for (i in 1:N){                        ## loop over observations
    y[i] ~ dbern(p[i]);                  ## binary outcome, Bernoulli trial
    logit(p[i]) <- ystar[i];             ## logit link
    ystar[i] <- beta[1]                  ## regression structure for covariates
              + beta[21*educ[i]
              + beta[31*(educ[i]*educ[i])
              + beta[41*age[i]
              + beta[5]*(age[i]*age[i])
              + beta[61*south [i]
              + beta[7]*govelec[i]
              + beta[8]*closing[i]
              + beta[9]*(closing[i]*educ[i])
              + beta[10]*(educ[i]*educ[i]*closing[i]);

    llh[i] <- y[i]*log(p[i]) + (1-y[i])*log(1-p[i]);   ## llh contributions
  }

  sumllh <- sum(llh[]);                  ## sum of log-likelihood contributions

  ## priors
  beta[1:10] ~ dmnorm(b0[ ] , B[ , ]) ;  ## multivariate Normal prior
}
```

The next program implements a probit model, but without a probit link function. The use of the $I(\ ,\ )$ construct directs the software to use truncated Normal sampling, as described in the text. The observed binary outcomes, y, only enter the program in selecting which set of truncation bounds to use when sampling the latent variable `ystar`. From the perspective of the software, it is the truncated Normal sampling that makes this a probit model: the probit link function is only used in order to generate auxiliary quantities of interest (predicted probabilities for specific data points and the log-likelihood).

```
##############################################################################
## turnout model, probit, via truncated Normal sampling                     ##
## y's determining the truncation points                                     ##
## this avoids the link to y via Bernoulli sampling and a link function      ##
## use the link function "ex-post" to extract predicted probabilities        ##
##                                                                           ##
## simon jackman, dept of political science, stanford university             ##
##############################################################################
model{
  for (i in 1:N){                          ## loop over observations
     mu[i] <- beta[1]
            + beta[2]*educ [i]
            + beta[3]*(educ[i]*educ[i])
            + beta[4]*age[i]
            + beta[5]*(age[i]*age [i])
            + beta[6]*south[i]
            + beta[7]*govelec[i]
            + beta[8]*closing[i]
            + beta[9]*(closing[i]*educ[i])
            + beta[10]*(educ[i]*educ[i]*closing[i]);

    ## truncated normal sampling
    ystar[i] ~ dnorm(mu[i],1)I(lo[y[i]+1],up[y[i]+1]);

    probit(p[i]) <- ystar[i];              ## probs, as probit link
    llh[i] <- y[i]*log(p[i]) + (1-y[i]*log(1-p[i]);

  }
  ## truncation points
  lo[1] <- -50; lo[2] <- 0;               ## ystar | y=0 ~ N(xb,1)I(-50,0)
  up[1] <-   0; up[2] <- 50;              ## ystar | y=1 ~ N(xb,1)I(0,50)

  sumllh <- sum(llh[ ]);                  ## sum log-likelihood contributions
  ## priors
  beta[1:10] ~ dmnorm(b0[ ] , B[ , ]) ; ## multivariate Normal prior
}
```

References

Albert, James A., and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–679.

Albert, James A. and Siddhartha Chib. 1996. "Computation in Bayesian Econometrics: An Introduction to Markov Chain Monte Carlo." In *Advances in Econometrics*, ed. T. Fomby and R. Carter Hill. Vol. 11A. Greenwich, Conn.: JAI Press.

Alvarez, R. Michael and Jonathan Nagler. 1995. "Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election." *American Journal of Political Science* 39:714–744.

Amemiya, Takeshi. 1985. *Advanced Econometrics.* Cambridge: Harvard University Press.

Bernardo, José, and Adrian F. M. Smith. 1994. *Bayesian Theory.* Chichester: Wiley.

Besag, J. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion)." *Journal of the Royal Statistical Society, Series B* 41:143–168.

Besag, J., and P. J. Green. 1993. "Spatial Statistics and Bayesian Computation (with Discussion)." *Journal of the Royal Statistical Society, Series B* 55:25–37.

Brooks, Stephen P., and Andrew Gelman. 1998. "Alternative Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics* 7:434–455.

Carlin, Bradley P., and Thomas A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis.* London: Champman and Hall.

Casella, George, and Edward I. George. 1992. "Explaining the Gibbs Sampler." *The American Statistician* 46:167–174.

Chib, Siddhartha. 1993. "Bayes Regression with Autoregressive Errors: a Gibbs Sampling Approach." *Journal of Econometrics* 58:275–294.

Chib, Siddhartha, and Edward Greenberg. 1994. "Bayes Inference in Regression Models with ARMA (p,q) errors." *Journal of Econometrics* 64:188–206.

Chib, Siddhartha, and Edward Greenberg. 1995. "Understanding the Metropolis-Hastings Algorithm." *The American Statistician* 49:327–335.

Chib, Siddhartha, and Edward Greenberg. 1997. "Analysis of Multivariate Probit Models." *Biometrika* 85:347–361.

Cleveland, William S. 1993. *Visualizing Data.* Summit, N.J.: Hobart Press.

Clifford, P. 1993. "Discussion on the Meeting on the Gibbs Sampler and Other Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society, Series B* 55:53–102.

Cochrane, D., and G.H. Orcutt. 1949. "Application of Least Squares Relationships Containing Autocorrelated Error Terms." *Journal of the American Statistical Association* 44:32–61.

Cowles, Mary Kathryn, and Bradley P. Carlin. 1996. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of the American Statistical Association* 91:883–904.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the *EM* Algorithm." *Journal of the Royal Statistical Society, Series B* 39:1–38.

Devroye, Luc. 1986. *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

Draper, David. 1999. *Bayesian Hierarchical Modeling.* Bath, UK: Department of Mathematical Sciences, University of Bath. In preparation. http://www.bath.ac.uk/~masdd.

Edwards, A. W. F. 1992. *Likelihood* Expanded Edition. Baltimore: Johns Hopkins.

Fishman, George S. 1996. *Monte Carlo: Concepts, Algorithms, and Applications.* New York: Springer.

Gamerman, Dani. 1997. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* London: Champman and Hall.

Gelfand, Alan E. 1997. "Gibbs Sampling." In *Encyclopedia of the Statistical Sciences*, ed. Samuel Kotz, Campbell B. Read, and David L. Banks. Vol. Update Volume 1 New York: Wiley pp. 283–292.

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

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Sciences* 7:457–472.

Gelman, Andrew, and Gary King. 1990. "Estimating the Consequences of Electoral Redistricting." *Journal of the American Statistical Association* 85:274–282.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis.* London: Chapman and Hall.

Geman, S., and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–741.

Geweke, J. 1989. "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica* 57:1317–1339.

Geweke, J. 1992. "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments (with Discussion)." In *Bayesian Statistics 4*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford: Oxford University Press.

Gilks, W. R., and P. Wild. 1992. "Adaptive Rejection Sampling for Gibbs Sampling." *Applied Statistics* 41:337–348.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. "Introducing Markov Chain Monte Carlo." In *Markov Chain Monte Carlo in Practice*, ed. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. London: Chapman and Hall.

Gilks, Walter R. 1992. "Derivative-Free Adaptive Rejection Sampling for Gibbs sampling." In *Bayesian Statistics 4*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford: Clarendon.

Green, P. J. 1996. "MCMC in Image Analysis." In *Markov chain Monte Carlo in practice*, ed. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. London: Chapman, and Hall.

Greene, William H. 1997. *Econometric Analysis.* 3rd ed. New York: Prentice-Hall.

Hamilton, James D. 1990. "Analysis of Time Series Subject to Changes in Regime." *Journal of Econometrics* 45:39–70.

Hamilton, James D. 1994. *Time Series Analysis.* Princeton: Princeton University Press.

Hastings, W. K. 1970. "Monte Carlo Sampling Methods Using Markov Chains, and their applications." *Biometrika* 57:97–109.

Jackman, Simon. 1994. "Measuring Electoral Bias: Australia, 1949–1993." *British Journal of Political Science* 24:319–357.

Johnson, Norman L., Samuel Kotz, and N. Balakrishnan. 1994. *Continuous Univariate Distributions.* Vol. 1. 2nd ed. New York: Wiley.

Johnson, Vance E., and James H. Albert. 1999. *Ordinal Data Modeling.* New York: Springer-Verlag.

Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee. 1988. *Introduction to the Theory, and Practice of Econometrics.* 2nd ed. New York: Wiley.

Keane, Michael P. 1992. "A Note on Identification in the Multinomial Probit Model." *Journal of Economics, and Business Statistics* 10:193–200.

King, Gary. 1989. *Unifying Political Methodology.* New York: Cambridge University Press.

King, Gary, James Honaker, Anne Joesph, and Kenneth Scheve. 1998. "Listwise Deletion is Evil: What to Do About Missing Data in Political Science." Unpublished manuscript, Harvard University (http://GKing.Harvard.Edu/preprints.shtml).

King, Gary, and Jonathan Katz. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review* 93:15–32.

King, Gary, Ori Rosen, and Martin A. Tanner. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference." *Sociological Methods, and Research.* 28:61–90.

Lacey, Dean, and Barry Burden. 1999. "The Vote-Stealing and Turnout Effects of Ross Perot in the 1992 U.S. Presidential Election." *American Journal of Political Science* 43:233–255.

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

Little, Roderick J. A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York: Wiley.

MacKinnon, James. 1991. "Critical Values for Cointegration Tests." In *Long-Run Economic Relationships: Readings in Cointegration*, ed. R. F. Engle, and C. W. J. Granger. Oxford: Oxford University Press.

McCulloch, Robert E., Nicholas G. Polson, and Peter E. Rossi. 1998. "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters." Unpublished manuscript. University of Chicago.

McCulloch, Robert E., and Peter E. Rossi. 1994. "An Exact Likelihood Analysis of the Multinomial Probit Model." *Journal of Econometrics* 64:207–240.

McLachlan, Geoffrey J., and Thriyambakam Krishnan. 1997. *The EM Algorithm and Extensions.* New York: Wiley.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. "Equations of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21:1087–1091.

Nagler, Jonathan. 1994. "Scobit: An Alternative Estimator to Logit, and Probit." *American Journal of Political Science* 38:230–255.

Polson, Nicholas G. 1996. "Convergence of Markov Chain Monte Carlo Algorithms." In *Bayesian Statistics 5*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F .M. Smith. Oxford: Oxford University Press.

Prais, S. J., and C. B. Winsten. 1954. *Trend Estimators and Serial Correlation.* Chicago: Cowles Commission.

Quinn, Kevin M., Andrew D. Martin, and, Andrew B. Whitford. 1999. "Voter Choice in Multi-Party Democracies: A Test of Competing Theories and Models." *American Journal of Political Science* 43:1231–1247.

Roberts, Gareth O. 1996. "Markov Chain Concepts Related to Sampling Algorithms." In *Markov Chain Monte Carlo in Practice*, ed. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. London: Chapman and Hall.

Rosenthal, J. 1995. "Rates of Convergence for Gibbs Sampling for Variance Component Models." *Annals of Statistics* 23:740–761.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Shumway, R. H., and D. S. Stoffer. 1982. "An Approach to Time Series Smoothing, and Forecasting Using the EM Algorithm." *Journal of Time Series Analysis* 3:253–264.

Smith, Adrian F.M., and G.O. Roberts. 1993. "Bayesian Computation via the Gibbs Sampler, and Related Markov Chain Monte Carlo Methods (with Discussion)." *Journal of the Royal Statistical Society, Series B* 55:3–23.

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

Spiegelhalter, David J., Andrew Thomas, Nicky Best, and Wally R. Gilks. 1997. *BUGS 0.6: Bayesian inference using Gibbs sampling.* Cambridge, UK: MRC Biostatistics Unit.

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

Tanner, Martin, and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82:528–540.

Tierney, Luke. 1996. "Introduction to General State-Space Markov Chain Theory." In *Markov Chain Monte Carlo in Practice*, ed. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. London: Chapman, and Hall.

Tierney, Luke. 1997. "Markov Chain Monte Carlo Algorithms." In *Encyclopedia of the Statistical Sciences*, ed. Samuel Kotz, Campbell B. Read, and David L. Banks. Vol. 1 (Update). New York: Wiley.

Titterington, D. M, Adrian F. M. Smith, and U. E. Makov. 1985. *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Watson, Mark W., and Robert F. Engle. 1983. "Alternative Algorithms for the Estimation of Dynamic Factor, MIMIC and Varying Coefficient Regression Models." *Journal of Econometrics* 23:385–400.

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

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