Gibbs Sampling

Alan E. Gelfand

Journal of the American Statistical Association, Vol. 95, No. 452. (Dec., 2000), pp. 1300-1304.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28200012%2995%3A452%3C1300%3AGS%3E2.0.CO%3B2-J

Journal of the American Statistical Association is currently published by American Statistical Association.

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, New York: Wiley.

Wahba, G. (1977), "A Survey of Some Smoothing Problems and the Method of Generalized Cross-Validation for Solving Them," in *Applications of Statistics*, ed. P. R. Krisnaiah, Amsterdam: North-Holland, pp. 507–523.

——— (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Yao, Q., and Tong, H. (1994), "On Prediction and Chaos in Stochastic Systems," *Philosophical Transactions of the Journal of the Royal Statistical Society*, Ser. A, 348, 357–369.

Zhang, H. P., and Singer, B. (1999), *Recursive Partitioning in the Health Sciences*, New York: Springer-Verlag.

Gibbs Sampling

Alan E. GELFAND

1. INTRODUCTION

During the course of the 1990s, the technology generally referred to as Markov chain Monte Carlo (MCMC) has revolutionized the way statistical models are fitted and, in the process, dramatically revised the scope of models which can be entertained.

This vignette focuses on the Gibbs sampler. I provide a review of its origins and its crossover into the mainstream statistical literature. I then attempt an assessment of the impact of Gibbs sampling on the research community, on both statisticians and subject area scientists. Finally, I offer some thoughts on where the technology is headed and what needs to be done as we move into the next millennium. The perspective is, obviously, mine, and I apologize in advance for any major omissions. In this regard, my reference list is modest, and again there may be some glaring omissions. I present little technical discussion, as by now detailed presentations are readily available in the literature. The books of Carlin and Louis (2000), Gelman, Carlin, Stern, and Rubin (1995), Robert and Casella (1999), and Tanner (1993) are good places to start. Also, within the world of MCMC, I adopt an informal definition of a Gibbs sampler. Whereas some writers describe "Metropolis steps within Gibbs sampling," others assert that the blockwise updating implicit in a Gibbs sampler is a special case of a "block-at-a-time" Metropolis–Hastings algorithm. For me, the crucial issue is replacement of the sampling of a high-dimensional vector with sampling of lower-dimensional component blocks, thus breaking the so-called curse of dimensionality.

In Section 2 I briefly review what the Gibbs sampler is, how it is implemented, and how it is used to provide inference. With regard to Gibbs sampling, Section 3 asks the question "How did it make its way into the mainstream of statistics?" Section 4 asks "What has been the impact?" Finally, Section 5 asks "Where are we going?" Here, speculation beyond the next decade seems fanciful.

2. WHAT IS GIBBS SAMPLING?

Gibbs sampling is a simulation tool for obtaining samples from a nonnormalized joint density function. Ipso facto, such samples may be "marginalized," providing samples from the marginal distributions associated with the joint density.

2.1 Motivation

The difficulty in obtaining marginal distributions from a nonnormalized joint density lies in integration. Suppose, for example, that $\boldsymbol{\theta}$ is a $p\times 1$ vector and $f(\boldsymbol{\theta})$ is a nonnormalized joint density for $\boldsymbol{\theta}$ with respect to Lebesgue measure. Normalizing f entails calculating $\int f(\boldsymbol{\theta})\,d\boldsymbol{\theta}$. To marginalize, say for $\boldsymbol{\theta}_i$, requires $h(\boldsymbol{\theta}_i) = \int f(\boldsymbol{\theta})\,d\boldsymbol{\theta}_{(i)} / \int f(\boldsymbol{\theta})\,d\boldsymbol{\theta}$, where $\boldsymbol{\theta}_{(i)}$ denotes all components of $\boldsymbol{\theta}$ save $\boldsymbol{\theta}_i$. Integration is also needed to obtain a marginal expectation or find the distribution of a function of $\boldsymbol{\theta}$. When p is large, such integration is analytically infeasible (the curse of dimensionality). Gibbs sampling offers a Monte Carlo approach.

The most prominent application has been for inference within a Bayesian framework. Here models are specified as a joint density for the observations, say $\mathbf{Y}$, and the model unknowns, say $\boldsymbol{\theta}$, in the form $h(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. In a Bayesian setting, the observed realizations of $\mathbf{Y}$ are viewed as fixed, and inference proceeds from the posterior density of $\boldsymbol{\theta}, \pi(\boldsymbol{\theta}|\mathbf{Y}) \propto h(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \equiv f(\boldsymbol{\theta})$, suppressing the fixed $\mathbf{Y}$. So $f(\boldsymbol{\theta})$ is a nonnormalized joint density, and Bayesian inference requires its marginals and expectations, as earlier. If the prior, $\pi(\boldsymbol{\theta})$, is set to 1 and if $h(\mathbf{Y}|\boldsymbol{\theta})$ is integrable over $\boldsymbol{\theta}$, then the likelihood becomes a nonnormalized density. If marginal likelihoods are of interest, then we have the previous integration problem.

2.2 Monte Carlo Sampling and Integration

Simulation-based approaches for investigating the nonnormalized density $f(\boldsymbol{\theta})$ appeal to the duality between population and sample. In particular, if we can generate arbitrarily many observations from $h(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) / \int f(\boldsymbol{\theta})$, so-called *Monte Carlo sampling*, then we can learn about any feature of $h(\boldsymbol{\theta})$ using the corresponding feature of the sample. Noniterative strategies for carrying out such sampling usually involve identification of an importance sampling density, $g(\boldsymbol{\theta})$ (see, e.g., Geweke 1989; West 1992). Given a sample from $g(\boldsymbol{\theta})$, we convert it to a sample from $h(\boldsymbol{\theta})$,

Alan E. Gelfand is Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269. (E-mail: *alan@stat.uconn.edu*). His work was supported in part by National Science Foundation grant DMS 96-25383.

by resampling, as done by Rubin (1988) and Smith and Gelfand (1992). If one only needs to compute expectations under $h(\boldsymbol{\theta})$, this can be done directly with samples from $g(\boldsymbol{\theta})$ (see, e.g., Ripley 1987) and is referred to as *Monte Carlo integration.* Noniterative Monte Carlo methods become infeasible for many high-dimensional models of interest.

Iterative Monte Carlo methods enable us to avoid the curse of dimensionality by sampling low-dimensional subsets of the components of $\boldsymbol{\theta}$. The idea is to create a Markov process whose stationary distribution is $h(\boldsymbol{\theta})$. This seems an unlikely strategy, but, perhaps surprisingly, there are infinities of ways to do this. Then, suppose that $P(\boldsymbol{\theta} \to A)$ is the transition kernel of a Markov chain with stationary distribution $h(\boldsymbol{\theta})$. (Here $P(\boldsymbol{\theta} \to A)$ denotes the probability that $\boldsymbol{\theta}^{(t+1)} \in A$ given $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$.) If $h^{(0)}(\boldsymbol{\theta})$ is a density that provides starting values for the chain, then, with $\boldsymbol{\theta}^{(0)} \sim h^{(0)}(\boldsymbol{\theta})$, using $P(\boldsymbol{\theta} \to A)$, we can develop a trajectory (sample path) of the chain $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(t)}, \ldots$ If t is large enough (i.e., after a sufficiently long "burn-in" period), then $\boldsymbol{\theta}^{(t)}$ is approximately distributed according to $\mathbf{h}(\theta)$.

A bit more formally, suppose that $P(\boldsymbol{\theta} \to A)$ admits a transition density, $p(\eta|\boldsymbol{\theta})$, with respect to Lebesgue measure. Then π is an invariant density for p if $\int \pi(\boldsymbol{\theta}) p(\eta|\boldsymbol{\theta})\, d\boldsymbol{\theta} = \pi(\eta)$. In other words, if $\boldsymbol{\theta}^{(t)} \sim \pi$, then $\boldsymbol{\theta}^{(t+1)} \sim \pi$. Also, Γ is a limiting (stationary, equilibrium) distribution for p if $\lim_{t\to\infty} P(\boldsymbol{\theta}^{(t)} \in A | \boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}) = \Gamma(A)$ (and thus $\lim_{t\to\infty} P(\boldsymbol{\theta}^{(t)} \in A) = \Gamma(A)$). The crucial result is that if $p(\eta|\boldsymbol{\theta})$ is aperiodic and irreducible and if π is a (proper) invariant distribution of p, then π is the unique invariant distribution; that is, π is the limiting distribution. A careful theoretical discussion of general MCMC algorithms with references was given by Tierney (1994). Also highly recommended is the set of three Royal Statistical Society papers in 1993 by Besag and Green (1993), Gilks et al. (1993), and Smith and Roberts (1993), together with the ensuing discussion, as well as an article by Besag, Green, Higdon, and Mengersen (1996), again with discussion.

2.3 The Gibbs Sampler

The Gibbs sampler was introduced as a MCMC tool in the context of image restoration by Geman and Geman (1984). Gelfand and Smith (1990) offered the Gibbs sampler as a very general approach for fitting statistical models, extending the applicability of the work of Geman and Geman and also broadening the substitution sampling ideas that Tanner and Wong (1987) proposed under the name of data augmentation.

Suppose that we partition $\boldsymbol{\theta}$ into r blocks; that is, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r)$. If the current state of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_r^{(t)})$, then suppose that we make the transition to $\boldsymbol{\theta}^{(t+1)}$ as follows:

$$
\begin{aligned}
&\text{draw } \boldsymbol{\theta}_1^{(t+1)} \text{ from } h(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_r^{(t)}),\\
&\text{draw } \boldsymbol{\theta}_2^{(t+1)} \text{ from } \mathbf{h}(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(t+1)}, \ldots, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_r^{(t)}),\\
&\qquad\vdots\\
&\text{draw } \boldsymbol{\theta}_r^{(t+1)} \text{ from } h(\boldsymbol{\theta}_r | \boldsymbol{\theta}_1^{(t+1)}, \ldots, \boldsymbol{\theta}_{r-1}^{(t+1)}).
\end{aligned}
$$

The distributions $h(\boldsymbol{\theta}_i | \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \ldots, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_r)$ are referred to as the full, or complete, conditional distributions, and the process of updating each of the r blocks as indicated updates the entire vector $\boldsymbol{\theta}$, producing one complete iteration of the Gibbs sampler. Sampling of $\boldsymbol{\theta}$ has been replaced by sampling of lower-dimensional blocks of components of $\boldsymbol{\theta}$.

2.4 How To Sample the $\boldsymbol{\theta}_i$

Conceptually, the Gibbs sampler emerges as a rather straightforward algorithmic procedure. One aspect of the art of implementation is efficient sampling of the full conditional distributions. Here there are many possibilities. Often, for some of the $\boldsymbol{\theta}_i$, the form of the prior specification will be conjugate with the form in the likelihood, so that the full conditional distribution for $\boldsymbol{\theta}_i$ will be a "posterior" updating of a standard prior. Note that even if this were the case for every $\boldsymbol{\theta}_i$, $f(\boldsymbol{\theta})$ itself need not be a standard distribution; conjugacy may be more useful for Gibbs sampling than for analytical investigation of the entire posterior.

When $f(\boldsymbol{\theta}_i | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots \boldsymbol{\theta}_r)$ is nonstandard, we might consider the rejection method, as discussed by Devroye (1986) and Ripley (1987); the weighted bootstrap, as discussed by Smith and Gelfand (1992); the ratio-of-uniforms method, as described by Wakefield, Gelfand, and Smith (1992); approximate cdf inversion when $\boldsymbol{\theta}_i$ is univariate, such as the griddy Gibbs sampler, as discussed by Ritter and Tanner (1992); adaptive rejection sampling, as often the full conditional density for $\boldsymbol{\theta}_i$ is log concave, in which case the usual rejection method may be adaptively improved in a computationally cheap fashion, as described by Gilks and Wild (1992); and Metropolis-within-Gibbs. For the last, the Metropolis (or Hastings–Metropolis) algorithms—which, in principle, enable simultaneous updating of the entire vector $\boldsymbol{\theta}$ (Chib and Greenberg 1995; Tierney 1994)—are usually more conveniently used within the Gibbs sampler for updating some of the $\boldsymbol{\theta}_i$, typically those with the least tractable full-conditional densities.

The important message here is that no single procedure dominates the others for all applications. The form of $h(\boldsymbol{\theta})$ determines which method is most suitable for a given $\boldsymbol{\theta}_i$.

2.5 Convergence

Considerable theoretical work has been done on establishing the convergence of the Gibbs sampler for particular applications, but perhaps the simplest conditions have been given by Smith and Roberts (1993). If $f(\boldsymbol{\theta})$ is lower semicontinuous at 0, if $\int f(\boldsymbol{\theta})\, d\boldsymbol{\theta}_i$ is locally bounded for each i, and if the support of f is connected, then the Gibbs sampler algorithm converges. In practice, a range of diagnostic tools is applied to the output of one or more sampled chains. Cowles and Carlin (1994) and Brooks and Roberts (1998) provided comparative reviews of the convergence diagnostics literature. Also, see the related discussions in the hierarchical models vignette by Hobert and the MCMC vignette by Cappé and Robert. (In principle, convergence can never be assessed using such output, as comparison can be made only between different iterations of one chain or

between different observed chains, but never with the true stationary distribution.)

2.6 Inference Using the Output of the Gibbs Sampler

The retained output from the Gibbs sampler will be a set of $\boldsymbol{\theta}_j^*, j = 1, 2, \ldots, B$, assumed to be approximately iid from $h = f/\int f$. If independently started parallel chains are used, then observations from different chains are independent but observations within a given chain are dependent. "Thinning" of the output stream (i.e., taking every kth iteration, perhaps after a burn-in period) yields approximately independent observations within the chain, for k sufficiently large. Evidently, the choice of k hinges on the autocorrelation in the chain. Hence sample autocorrelation functions are often computed to assess the dependence. Given $\{\boldsymbol{\theta}_j^*\}$, for a specified feature of h we compute the corresponding feature of the sample. Because B can be made arbitrarily large, inference using $\{\boldsymbol{\theta}_j^*\}$ can be made arbitrarily accurate.

3. HOW DID IT HAPPEN?

The Gibbs sampler was not developed by statisticians. For at least the past half-century, scientists (primarily physicists and applied mathematicians) have sought to simulate the behavior of complex probabilistic models formulated to approximate the behavior of physical, chemical, and biological processes. Such processes were typically characterized by regular lattice structure and the joint probability distribution of the variables at the lattice points was provided through local specification; That is, the full conditional density $h(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j, j = 1, 2, \ldots, r, j \neq i)$ was reduced to $h(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j \in N_i)$, where N_i is a set of neighbors of location i. But then an obvious question is whether the set of densities $h(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j \in N_i)$, a so-called Markov random field (MRF) specification, uniquely determines $h(\boldsymbol{\theta})$. Geman and Geman (1984) argued that if each full conditional distribution is a so-called Gibbs density, the answer is yes and, in fact, that this provides an equivalent definition of a MRF. The fact that each $\boldsymbol{\theta}_i$ is updated by making a draw from a Gibbs distribution motivated them to refer to the entire updating scheme as Gibbs sampling.

The Gibbs sampler is, arguably, better suited for handling simulation from a posterior distribution. As noted by Gelfand and Smith (1990), $h(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j, j \neq i) \propto f(\boldsymbol{\theta})$, where $f(\boldsymbol{\theta})$ is viewed as a function of $\boldsymbol{\theta}_i$ with all other arguments fixed. Hence we always know (at least up to normalization) the full conditional densities needed to implement the Gibbs sampler. The Gibbs sampler can also be used to investigate *conditional* distributions associated with $f(\boldsymbol{\theta})$, as done by Gelfand and Smith (1991). It is also well suited to the case where $f(\boldsymbol{\theta})$ arises as the restriction of a joint density to a set S (see Gelfand, Smith, and Lee 1992).

The 1990s have brought unimaginable availability of inexpensive high-speed computing. Such computing capability was blossoming at the time of Gelfand and Smith's 1990 article. The former fueled considerable experimentation with the latter, in the process demonstrating its broad practical viability. Concurrently, the increasing computing possibilities were spurring interest in a broad range of complex modeling specifications, including generalized linear mixed models, time series and dynamic models, nonparametric and semiparametric models (particularly for censored survival data), and longitudinal and spatial data models. These could all be straightforwardly fitted as Bayesian models using Gibbs sampling.

4. WHAT HAS BEEN THE IMPACT?

Previously, within the statistical community, Bayesians, though confident in the unification and coherence that their paradigm provides, were frustrated by the computational limitations described in Section 2.1, which restricted them to "toy" problems. Though progress was made with numerical integration approaches, analytic approximation methods, and noniterative simulation strategies, fitting the rich classes of hierarchical models that provide the real inferential benefits of the paradigm (e.g., smoothing, borrowing strength, accurate interval estimates) was generally beyond the capability of these tools. The Gibbs sampler provided Bayesians with a tool to fit models previously inaccessible to classical workers. The tables were turned; if one specified a likelihood and prior, the Gibbs sampler was ready to go!

The ensuing fallout has by and large been predictable. Practitioners and subject matter researchers, seeking to explore more realistic models for their data, have enthusiastically embraced the Gibbs sampler, and Bayesians, stimulated by such receptiveness, have eagerly sought collaborative research opportunities. An astonishing proliferation of articles using MCMC model fitting has resulted. On the other hand, classical theoreticians and methodologists, perhaps feeling somewhat threatened, find intellectual vapidity in the entire enterprise; "another Gibbs sampler paper" is a familiar retort.

Though not all statisticians participate, an ideological divide, perhaps stronger than in the past, has emerged. Bayesians will argue that with a full model specification, full inference is available. And the inference is "exact" (although an enormous amount of sampling from the posterior may be required to achieve it!), avoiding the uncertainty associated with asymptotic inference. Frequentists will raise familiar concerns with prior specifications and with inference performance under experimental replication. They also will feel uncomfortable with the black box, nonanalytic nature of the Gibbs sampler. Rather than "random" estimates, they may prefer explicit expressions that permit analytic investigation.

Moreover, Gibbs sampling, as a model fitting and data-analytic technology is fraught with the risk for abuse. MCMC methods are frequently stretched to models more complex than the data can hope to support. Inadequate investigation of convergence in high-dimensional settings is often the norm, improper posteriors surface periodically in the literature, and inference is rarely externally checked.

5. WHERE ARE WE GOING?

At this point, the Gibbs sampler and MCMC in general are well accepted and utilized for data analysis. Its use in the applied sector will continue to grow. Nonetheless, in the statistical community the frenzy over Gibbs sampling has passed, the field is now relatively stable, and future direction can be assessed. I begin with a list of "tricks of the trade," items still requiring further clarification:

- Model fitting should proceed from simplest to hardest, with fitting of simpler models providing possible starting values, mode searching, and proposal densities for harder models.
- Attention to parameterization is crucial. Given the futility of "transformation to uncorrelatedness," automatic approaches, such as that of Gelfand, Sahu, and Carlin (1995a,b), are needed. Strategies for nonlinear models are even more valuable.
- Latent and auxiliary variables are valuable devices but effective usage requires appreciation of the trade-off between simplified sampling and increased model dimension.
- When can one use the output associated with a component of θ that appears to have converged? For instance, population-level parameters, which are often of primary interest, typically converge more rapidly than individual-level parameters.
- Blocking is recognized as being helpful in handling correlation in the posterior but what are appropriate blocking strategies for hierarchical models?
- Often "hard to fit" parameters are fixed to improve the convergence behavior of a Gibbs sampler. Is an associated sensitivity analysis adequate in such cases?
- Because harder models are usually weakly identified, informative priors are typically required to obtain well behaved Gibbs samplers. How does one use the data to develop these priors and to specify them as weakly as possible?
- Good starting values are required to run multiple chains. How does one obtain "overdispersed" starting values?

With the broad range of models that can now be explored using Gibbs sampling, one naturally must address questions of model determination. Strategies that conveniently piggyback onto the output of Gibbs samplers are of particular interest as ad hoc screening and checking procedures. In this regard, see Gelfand and Ghosh (1998) and Spiegelhalter, Best, and Carlin (1998) for model choice approaches and Gelman, Meng, and Stern (1995) for model adequacy ideas.

Finally, with regard to software development, the BUGS package (Spiegelhalter, Thomas, Best, and Gilks 1995) at this point, is general and reliable enough (with no current competition) to be used both for research and teaching. CODA (Best, Cowles, and Vines 1995) is a convenient add-on to implement a medley of convergence diagnostics. The future will likely bring specialized packages to accommodate specific classes of models, such as time series and dynamic models. However, fitting cutting edge models will always require tinkering and tuning (and possibly specialized algorithms), placing it beyond extant software. But the latter can often fit simpler models before exploring harder ones and can be used to check individual code.

As for hardware, it is a given that increasingly faster machines with larger and larger storage will evolve, making feasible the execution of enormous numbers of iterations for high-dimensional models within realistic run times, diminishing convergence concerns. However, one also would expect that more capable multiprocessor machines will be challenged by bigger datasets and more complex models, encouraging parallel processing MCMC implementations.

REFERENCES

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

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1996), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3–66.

Best, N. G., Cowles, M. K., and Vines, K. (1995), "CODA: Convergence Diagnostics and Output Analysis Software for Gibbs Sampling Output, Version 0.30," Medical Research Council, Biostatistics Unit, Cambridge, U.K.

Brooks, S. P., and Roberts, G. O. (1998), "Assessing Convergence of Markov Chain Monte Carlo Algorithms," *Statistics and Computing*, 8, 319–335.

Carlin, B. P., and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), London: Chapman and Hall.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis–Hastings Algorithm," *American Statistician*, 49, 327–335.

Cowles, M. K., and Carlin, B. P. (1994), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 883–904.

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

Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995a), "Efficient Parameterization for Normal Linear Mixed Effects Models," *Biometrika*, 82, 479–488.

——— (1995b), "Efficient Parameterization for Generalized Linear Mixed Models," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 47–74.

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

Gelfand, A. E., Smith, A. F. M., and Lee, T-M. (1992), "Bayesian Analysis of Constrained Parameter and Truncated Data Problems," *Journal of the American Statistical Association*, 87, 523–532.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

Gelman, A., Meng, X-L., and Stern, H. S. (1995), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733–807.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, 6, 721–741.

Geweke, J. F. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrika*, 57, 1317–1340.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. E., McNeil, A. J., Sharples, L. D., and Kirby, A. J. (1993), "Modeling Complexity: Applications of Gibbs Sampling in Medicine," *Journal of the Royal*

Statistical Society, Ser. B, 55, 39–52.
Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society*, Ser. C, 41, 337–348.
Ripley, B. D. (1987), *Stochastic Simulation*, New York: Wiley.
Ritter, C., and Tanner, M. A. (1992), "The Gibbs Stopper and the Griddy Gibbs Sampler," *Journal of the American Statistical Association*, 87, 861–868.
Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.
Rubin, D. B. (1988), "Using the SIR Algorithm to Simulate Posterior Distributions," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, London: Oxford University Press, pp. 395–402.
Smith, A. F. M., and Gelfand, A. E. (1992), "Bayesian Statistics Without Tears," *American Statistician*, 46, 84–88.
Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 3–23.
Spiegelhalter, D., Best, N., and Carlin, B. P. (1998), "Bayesian Deviance, the Effective Number of Parameters and the Comparison of Arbitrarily Complex Models," technical report, MRC Biostatistics Unit, Cambridge, U.K.
Spiegelhalter, D. J., Thomas, A., Best, N., and Gilks, W. R. (1995), "BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50," Medical Research Council, Biostatistics Unit, Cambridge, U.K.
Tanner, M. A. (1993), *Tools for Statistical Inference* (2nd ed.), New York: Springer-Verlag.
Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.
Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Ann. Statist.*, 22, 1701–1762.
Wakefield, J., Gelfand, A. E., and Smith, A. F. M. (1992), "Efficient Computation of Random Variates via the Ratio-of-Uniforms Method," *Statist. and Comput.*, 1, 129–133.
West, M. (1992), "Modeling With Mixtures," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 503–524.

The Variable Selection Problem

Edward I. GEORGE

The problem of variable selection is one of the most pervasive model selection problems in statistical applications. Often referred to as the problem of subset selection, it arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is uncertainty about which subset to use. This vignette reviews some of the key developments that have led to the wide variety of approaches for this problem.

1. INTRODUCTION

Suppose that $\mathbf{Y}$, a variable of interest, and $\mathbf{X}_1, \ldots, \mathbf{X}_p$, a set of potential explanatory variables or predictors, are vectors of n observations. The problem of variable selection, or subset selection as it is often called, arises when one wants to model the relationship between $\mathbf{Y}$ and a subset of $\mathbf{X}_1, \ldots, \mathbf{X}_p$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when p is large and $\mathbf{X}_1, \ldots, \mathbf{X}_p$ is thought to contain many redundant or irrelevant variables.

The variable selection problem is most familiar in the linear regression context, where attention is restricted to normal linear models. Letting γ index the subsets of $\mathbf{X}_1, \ldots, \mathbf{X}_p$ and letting q_γ be the size of the γth subset, the problem is to select and fit a model of the form

$$\mathbf{Y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \varepsilon, \tag{1}$$

where $\mathbf{X}_\gamma$ is an $n \times q_\gamma$ matrix whose columns correspond to the γth subset, $\boldsymbol{\beta}_\gamma$ is a $q_\gamma \times 1$ vector of regression coefficients, and $\varepsilon \sim N_n(0, \sigma^2 I)$. More generally, the variable selection problem is a special case of the model selection problem where each model under consideration corresponds to a distinct subset of $\mathbf{X}_1, \ldots, \mathbf{X}_p$. Typically, a single model class is simply applied to all possible subsets. For example, a wide variety of relationships can be considered with generalized linear models where $g(E(\mathbf{Y})) = \alpha + \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma$ for some link function g (see the vignettes by Christensen and McCulloch). Moving further away from the normal linear model, one might instead consider relating $\mathbf{Y}$ and subsets of $\mathbf{X}_1, \ldots, \mathbf{X}_p$ with nonparametric models such as CART or MARS.

The fundamental developments in variable selection seem to have occurred either directly in the context of the linear model (1) or in the context of general model selection frameworks. Historically, the focus began with the linear model in the 1960s, when the first wave of important developments occurred and computing was expensive. The focus on the linear model still continues, in part because its analytic tractability greatly facilitates insight, but also because many problems of interest can be posed as linear variable selection problems. For example, for the problem of nonparametric function estimation, $\mathbf{Y}$ represents the values of the unknown function, and $\mathbf{X}_1, \ldots, \mathbf{X}_p$ represent a linear basis, such as a wavelet basis or a spline basis. However, as advances in computing technology have allowed for the implementation of richer classes of models, treatments of the variable selection problem by general model selection approaches are becoming more prevalent.

One of the fascinating aspects of the variable selection problem has been the wide variety of methods that have

Edward I. George holds the Ed and Molly Smith Chair and is Professor of Statistics, Department of MSIS, University of Texas, Austin, TX 78712 (E-mail: *egeorge@mail.utexas.edu*). This work was supported by National Science Foundation grant DMS-98.03756 and Texas ARP grants 003658.452 and 003658.690.