# Monte Carlo Integration in Bayesian Statistical Analysis

*By*

*Robert L. Wolpert* [1]

*Institute of Statistics and Decision Sciences*

DP# 90-A-06

A review of Monte Carlo methods for approximating the high-dimensional integrals that arise in Bayesian statistical analysis. Emphasis is on the features of many Bayesian applications which make Monte Carlo methods especially appropriate, and on Monte Carlo variance-reduction techniques especially well suited to Bayesian applications. A generalized logistic regression example is used to illustrate the ideas, and high-precision formulas are given for implementing Bayesian Monte Carlo integration.

1. Robert Wolpert is Associate Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC, 27706 (*rlw@isds.duke.edu*). This research was supported in part by NSF Grants No. SES-8605867 and SES-8921227.

2. *AMS 1980 subject classifications.* Primary 62F15; secondary 65C06.

3. *Key words and phrases.* Bayesian analysis, numerical integration, simulation, Monte Carlo method, variance reduction.

## 1. BACKGROUND: BAYESIAN ANALYSIS

In Bayesian analysis uncertainty about a quantity $\theta$ is represented in the form of a probability measure $\pi(d\theta)$ with which one can calculate the probability $\mathbf{P}^\pi[\theta \in A]$ that $X$ lies in any measurable set $A$, or (more generally) the expectation $\mathbf{E}^\pi g(\theta)$ of any function of interest that depends on $\theta$. Upon observing the value of some random quantity $X$ whose probability distribution depends on $\theta$, Bayes' theorem gives the rule for calculating the probability distribution $\pi^\star(d\theta)$ for uncertainty about $\theta$ *after* observing $X{=}x$ as the ratio

$$\pi^\star(d\theta) = \frac{f(x|\theta)\,\pi(d\theta)}{\int_\Theta f(x|\theta')\,\pi(d\theta')} \tag{1}$$

where $\pi(d\theta)$ is the probability distribution representing uncertainty about $\theta$ *before* observing $X{=}x$, and where $f(x|\theta)$ is the probability density function for $X$ (with respect to an arbitrary dominating measure $\mu(dx)$), evaluated at the observed value $X{=}x$. The expectation of some function of interest $g(\theta)$ (possibly a vector) is given by a similar ratio:

$$\begin{aligned}\overline{g} \equiv& \mathbf{E}[g(\theta)|X{=}x]\\ =& \int_\Theta g(\theta)\,\pi^\star(d\theta)\\ =& \frac{\int_\Theta g(\theta) f(x|\theta)\,\pi(d\theta)}{\int_\Theta f(x|\theta)\,\pi(d\theta)}.\end{aligned} \tag{2}$$

Although we do not require that $\pi(d\theta)$ be properly normalized or even that $\pi(\Theta) < \infty$, we will assume throughout that $\int_\Theta f(x|\theta)\,\pi(d\theta) < \infty$ and $\int_\Theta |g(\theta)|\, f(x|\theta)\,\pi(d\theta) < \infty$ for every $x$, so both numerator and denominator will be well-defined in (1), (2), and similar formulas to follow.

The posterior covariance matrix $\Sigma \equiv \mathbf{E}[(g(\theta)-\overline{g})(g(\theta)-\overline{g})'|X{=}x]$ gives one indication of how well $g(\theta)$ is determined by the prior distribution $\pi(d\theta)$ and the observation $X{=}x$, and so measures the accuracy or precision of the estimate $\overline{g}$ of $g(\theta)$. It too is given by a ratio of integrals

$$\begin{aligned}\Sigma \equiv& \mathbf{E}[(g(\theta)-\overline{g})(g(\theta)-\overline{g})'|X{=}x]\\ =& \frac{\int (g(\theta)-\overline{g})(g(\theta)-\overline{g})' f(x|\theta)\,\pi(d\theta)}{\int f(x|\theta)\,\pi(d\theta)}.\end{aligned} \tag{3}$$

A more complete representation of the uncertainty attendant $g(\theta)$ following the observation of $X{=}x$ would be the (joint) *posterior density function* for $g(\theta)$, if it exists. Although $\xi = g(\theta)$ may not have a density function (its distribution will not be absolutely continuous if $\pi(d\theta)$ is supported on a lower-dimensional manifold, for example, or if $g(\theta)$ is constant on a set of positive measure) it can be approximated arbitrarily well by a distribution with a density of the form

$$\begin{aligned}\pi^\star_\epsilon(\xi) \equiv& \int_\Theta K_\epsilon(\xi - g(\theta))\,\pi^\star(d\theta)\\ =&\frac{\int_\Theta K_\epsilon(\xi - g(\theta)) f(x|\theta)\,\pi(d\theta)}{\int_\Theta f(x|\theta)\,\pi(d\theta)}\end{aligned} \tag{4}$$

for any approximate identity $K_\epsilon(x) = \epsilon^{-q}K(x/\epsilon) \geq 0$ satisfying $\int_{R^q} K_\epsilon(x)\,dx = 1$ with suitably small $\epsilon > 0$. For plotting marginal posterior densities and for many other purposes it is enough to know $\pi^\star_\epsilon(\xi)$ at a few hundred points $\{\xi_j\}$, *i.e.*, to find the posterior expectation of a several-hundred-dimensional function $\{K_\epsilon(\xi_j - g(\theta))\}_{j\in J}$.

If the method of selecting a probability distribution $\pi(d\theta)$ to represent knowledge about $\theta$ before the experiment does not depend on the distribution of $X$, then Bayesian statistical analysis based on $\pi^\star$ is consistent with the Likelihood Principle, *i.e.*, depends on the observation $X{=}x$ only through the "likelihood function"

$$L(\theta) = L(\theta|X{=}x) = f(x|\theta).$$

Birnbaum and others (reviewed and extended by Berger and Wolpert, 1988) have shown that any violation of the Likelihood Principle also violates either the Weak Conditionality Principle (which asserts that, if one randomly selects between two experiments, then only the experiment performed is relevant) or the Sufficiency Principle (which asserts that all evidence about $\theta$ from observing $X$ is also contained in any sufficient statistic $T(X)$). Since these two principles are widely held, it is hard to justify the use of a statistical procedure inconsistent with the Likelihood Principle.

Although the likelihood function $L(\theta) = f(x|\theta)$ is defined as a probability density function *for* $X$ *given* $\theta$, with respect to some arbitrary reference measure $\mu(dx)$, the Bayesian statistician (in light of Equation (1)) regards it as the probability density function *for* $\theta$ *given* $X{=}x$, with respect to a reference measure proportional to $\pi(d\theta)$. The likelihood function is central for non-Bayesian statistical methods as well: the maximum likelihood estimate (MLE) $\hat g$ of $g(\theta)$ is just the function $g(\theta)$ evaluated at a point $\hat\theta$ where $L(\theta)$ attains its maximum

$$\begin{aligned}\hat g =& \widehat{g(\theta)}\\ =& g(\hat\theta).\end{aligned}$$

For a (nearly) uniform prior distribution $\pi(\theta)$ the MLE $\hat g$ is (nearly) the *mode* of the distribution of $g(\theta)$ under the posterior $\pi^\star(d\theta)$, while $\overline{g}$ in (2) is the *mean.*

The difference between a mode and mean doesn't seem so dramatic, and in many cases it is not; the real issues arise in trying to represent the degree of certainty with which $g(\theta)$ is known, following the observation of $X{=}x$. The Bayesian uses the same measure $\pi^\star$ given in (1) to evaluate the *posterior probabilities* that $g(\theta)$ lies in specified sets (especially *posterior HPD regions*) or to evaluate the *posterior covariance* $\Sigma$ of $\overline{g}$ as in (3) or the marginal *posterior density* for some components of $g(\theta)$ as in (4), and thus

stays faithful to the likelihood principle, while the frequentist constructs *p-values*, *confidence sets* and *standard errors* by considering what the likelihood function *would have been* if other, near-by values of $X = \tilde{x} \approx x$ had been observed instead of $X{=}x$, and in so doing leaves the Likelihood Principle behind.

Unfortunately the integrals necessary for calculating the ratios in Equations (2–4) are seldom amenable to analytical methods. Frequently the parameter $\theta$ takes values in a high-dimensional space, making the integral resistant to quadrature methods; the simplest quarterly or monthly time-series applications lead to problems in four or twelve dimensions, for example, and problems in five to ten or twenty dimensions and more are common in many fields of application. Tensor-product quadrature methods are unthinkable in such problems.

One recourse is to appeal to large-sample asymptotic normality and approximate the mean $\overline{g}$ in (2) by the mode $\hat{g}$, and the covariance $\Sigma$ in (3) by the inverse of the information matrix. Another is to approximate the likelihood function $L(\theta) = f(x|\theta)$ and the prior distribution $\pi(\theta)$ by members of some conjugate pair of density families for which the integrals can be evaluated in closed form for a suitable class of functions $g(\theta)$ (*e.g.*, polynomials). With the emergence of fast desk-top computers and appropriate numerical algorithms a third choice has emerged: to approximate the integrals in (2–4) through Monte Carlo integration.

## 2. BACKGROUND: MONTE CARLO INTEGRATION

The theory of Monte Carlo integration is simple. If $\theta_i \in \Theta$ is a sequence of synthetic random variates, each drawn from some probability distribution $\Pi(d\theta)$ that dominates $\pi(d\theta)$, then let $w(\theta)$ be the Radon-Nikodym derivative

$$w(\theta) = f(x|\theta)\frac{\pi(d\theta)}{\Pi(d\theta)}, \tag{5}$$

and consider the weighted averages

$$\overline{g}_n \equiv \frac{\sum_{i=1}^n g(\theta_i)w(\theta_i)}{\sum_{i=1}^n w(\theta_i)}. \tag{6}$$

It is easy to calculate the expectations of the numerator and denominator in Equation (6), both of which are well-defined and finite by the earlier assumptions that both the likelihood function $f(x|\theta)$ and the product $g(\theta)f(x|\theta)$ are integrable with respect to the prior $\pi(d\theta)$, *i.e.*, $\int_\Theta f(x|\theta)\,\pi(d\theta) < \infty$ and $\int_\Theta |g(\theta)|\,f(x|\theta)\,\pi(d\theta) < \infty$. The numerator has expectation

$$\begin{aligned}\mathbf{E}\sum_{i=1}^n g(\theta_i)w(\theta_i) =& \sum_{i=1}^n \int_\Theta g(\theta_i)f(x|\theta_i)\frac{\pi(d\theta_i)}{\Pi(d\theta_i)}\Pi(d\theta_i)\\ =& n \times \int_\Theta g(\theta)f(x|\theta)\,\pi(d\theta),\end{aligned}$$

whereas the denominator has expectation

$$\begin{aligned}\mathbf{E}\sum_{i=1}^n w(\theta_i) =& \sum_{i=1}^n \int_\Theta f(x|\theta_i)\frac{\pi(d\theta_i)}{\Pi(d\theta_i)}\Pi(d\theta_i)\\ =& n \times \int_\Theta f(x|\theta)\,\pi(d\theta).\end{aligned}$$

Thus $\overline{g}_n$ is the ratio of unbiased estimates of the numerator and denominator of (2); it doesn't follow that $\overline{g}_n$ is an unbiased estimate of $\overline{g}$, of course, but it suggests that the approach is promising. Notice that the sample size and joint distribution of $\{\theta_i\}_{i\le n}$ weren't specified; we are free to choose $n$ and $\Pi(d\theta)$ in any convenient way, and the $\{\theta_i\}$ need not be stochastically independent. If they are at least *mixing* (*e.g.*, if each $\theta_i$ is independent of all but finitely many $\theta_j$) then two applications of the strong ergodic theorem (one to the numerator and one to the denominator) give an immediate proof of the consistency of Monte Carlo estimation, *i.e.*, the almost-sure convergence of $\overline{g}_n$ to $\overline{g}$.

The Monte Carlo estimate of $\Sigma = \mathbf{E}[(g(\theta)-\overline{g})(g(\theta)-\overline{g})'|X{=}x]$ is given by the matrix expression

$$\Sigma_n \equiv \frac{\sum_{i=1}^n (g(\theta_i)-\overline{g}_n)(g(\theta_i)-\overline{g}_n)'w(\theta_i)}{\sum_{i=1}^n w(\theta_i)} \tag{7}$$

and that of the marginal density $\{\pi^\star_\epsilon(\xi_j)\}$ by the weighted kernel density estimate

$$\pi^\star_{\epsilon n}(\xi_j) \equiv \frac{\sum_{i=1}^n K_\epsilon(\xi_j - g(\theta_i))w(\theta_i)}{\sum_{i=1}^n w(\theta_i)}. \tag{8}$$

If $\xi = g(\theta)$ has an absolutely continuous posterior distribution then, for fixed $\epsilon$, the sequence of estimates $\overline{g}_n$ will not be consistent estimators of the density function $\pi^\star_0(\xi_j)$; in fact, the sequence will converge pointwise almost surely to $\pi^\star_\epsilon(\xi)$, the convolution of the true density $\pi^\star_0(\xi)$ with the kernel $K_\epsilon(\xi)$. If a sequence $\epsilon_n \to 0$ converges to zero sufficiently slowly that $(\epsilon_n)^q n \to \infty$ (*e.g.*, $\epsilon_n = n^{-(q+1)^{-1}}$) then one can verify pointwise convergence of $\pi^\star_{\epsilon_n n}(\xi_j)$ to $\pi^\star_0(\xi_j)$ at all points of continuity $\xi_j$.

The quantities $\overline{g}_n$, $\overline{\Sigma}_n$, and $\pi^\star_{\epsilon n}(\xi_j)$ are random variables; we should choose $n$ and the joint sampling distribution of the variates $\{\theta_i\}$ in such a way as to minimize their computation time and some measure of their likely errors in estimating $\overline{g}$, $\overline{\Sigma}$, and $\pi^\star_0(\xi_j)$.

The mean square error $\sqrt{\mathbf{E}|\overline{g}_n - \overline{g}|^2}$ in Monte Carlo importance sampling falls off as $\sigma/\sqrt{n}$ for some constant $\sigma > 0$ if $w, gw \in \mathcal{L}^2(\Theta, \Pi(d\theta))$. For independent, identically-distributed (*iid*) sampling it is easy to show that the constant is approximately

$$\sigma \approx \frac{\sqrt{\int \big(g(\theta) - \overline{g}\big)^2 w(\theta)^2 \Pi(d\theta)}}{\int w(\theta)\Pi(d\theta)},$$

but *variance reduction techniques* can often lead to much smaller constants. In $d$ dimensions, the errors of tensor-product quadrature rules of local order $m$ fall off as $n^{-m/d}$; the commonly used fourth-order Runge-Kutta scheme, for example, has errors that fall as $n^{-4/d}$, faster than those of *iid* Monte Carlo methods in dimensions $d \le 7$. For this reason Monte Carlo methods were widely discounted in the 1950's, until the discovery by Hammersley and Morton of "antithetic accelertion" (or, at least, its popularization; see Tukey (1957)), the most important of several methods of reducing the constant $\sigma$ above. For small $d$ Monte Carlo integration may not be efficient *asymptotically*, in the limit as $n \to \infty$, but it can be quite efficient for achieving moderate precision with modest $n$ if $\sigma$ is made sufficiently small. Popular acceleration methods for reducing $\sigma$ enough to make Monte Carlo practical for Bayesian integration include

1. **importance sampling**, the use of a sampling distribution $\Pi(d\theta)$ lending relatively little weight to the "unimportant" regions to which $f(x|\theta)\,\pi(d\theta)$ gives little weight and instead concentrating on those areas to which $f(x|\theta)\,\pi(d\theta)$ gives great weight (or, roughly, choosing $\Pi(d\theta)$ to insure that $w(\theta)$ is bounded and nearly constant). See Curtiss, *et al.* (1951) for an account of the early development of importance sampling by Fermi, von Neumann, and Ulam. Specific cases or generalizations include *Russian roulette, splitting, stratified sampling,* and *conditional Monte Carlo.*
2. **control variables**, the indirect use of Monte Carlo methodology to estimate the difference $\int \big(g(\theta) - h(\theta)\big)w(\theta)\Pi(d\theta) \big/ \int w(\theta)\Pi(d\theta) = \overline{g} - \overline{h}$ for some function $h(\theta) \approx g(\theta)$ with known expectation $\overline{h}$ (see Fieller and Hartley, 1954). A generalization of this method is the *regression* method.
3. **antithetic variates**, the use of non–*iid* variates (especially negatively-correlated pairs). Generalizations of this include *random quadrature* methods. See Hammersley and Morton (1956).

These and other methods (*e.g.*, orthogonal polynomials) are discussed in Hammersley and Handscomb (chapter 5, 1964), Rubinstein (1981, 1986), and Wilson (1984).

A renewed interest in Monte Carlo methods accompanied the appearance in the early 1980's of widely-available minicomputers and low-cost microcomputers. Kloek and van Dijk (1978) introduced *adaptive* importance sampling, the dynamic adjustment of the sampling distribution $\Pi(d\theta)$ to improve incrementally the algorithm's efficiency, while Naylor and Smith (1982), Smith *et al.* (1985), and others at the University of Nottingham developed adaptive orthogonalization techniques improving the efficiency of both quadrature and Monte Carlo techniques. Geman and Geman (1984) in a study of Bayesian methods for image reconstruction developed and applied the theory of Gibbs sampling schemes, later recognized by Gelfand and Smith (1988) to be a broadly useful tool in high-dimensional Bayesian analysis. Tanner and Wong (1987) introduced a related technique (substitution sampling). Geweke (1988) proved a central limit theorem for Monte Carlo sampling schemes in Bayesian analysis. Many other recent contributions deserve mention in this active field.

One of the reasons these methods are so successful in Bayesian statistics is that statisticians seldom must integrate pathological functions with great local variation; indeed, likelihood functions and their products with prior densities can often be approximated strikingly well (sometimes after a nonlinear transformation removing positivity or monotonicity constraints) by simple elliptically symmetric functions such as multivariate Student $t$ or the Gaussian forms

$$f(x|\theta)\pi(\theta) \approx ce^{-\frac{1}{2}(\theta-\hat{\theta})'\Lambda(\theta-\hat{\theta})} \tag{9}$$

for some vector $\hat{\theta}$ and Hermitian form $\Lambda$. Important problems with non-unimodal integrands do arise (*e.g.*, in estimating the central tendency in problems with broad-tailed distributions, such as Cauchy or Student $t$), and some problems exhibit sufficient skewness to require novel, asymetric sampling distributions, but Bayesian statistical methods are now more practical than ever before precisely because the integrands encountered in Equations (2–4) above are commonly well enough behaved that Monte Carlo methods work well, even in high dimensional problems. In the example below the parameters in a seven-dimensional variation on a logistic regression model are fit to about two decimals of precision in several minutes' computation time on a desktop workstation by Monte Carlo methods, while a five-parameter submodel could not be fit to $\pm 10\%$ precision in a week-end of computation using fourth-order Runge-Kutta. Asymptotically the quadrature method must win out, of course, since eventually $c_1 n^{-4/7} < c_2 n^{-1/2}$, but the asymptotics are not terribly relevant to the practical problem of finding an approximate solution to the problem at hand. The solution can be found using Monte Carlo methods, and (apparently) not using quadrature methods.

It is worth noting that nowhere in the discussion of the Monte Carlo method and methods of accelerating its convergence has it been necessary to consider the dimension of the parameter space $\Theta$. Quadrature methods calling for the evaluation of an integrand at points of a lattice require an amount of computation which increases exponentially in the dimension, while Monte Carlo does not.

Of course this point can be emphasized too strongly. In many high-dimensional

problems the matrix $\Lambda$ in (9) above is badly conditioned, so the measure $f(x|\theta)\,\pi(d\theta)$ is supported in a neighborhood of some lower dimensional (and possibly curved) space; a failure to detect this situation or imprecision in identifying the space can lead to inefficient sampling schemes and even to gross undetected estimation errors. If the curvature is significant then nonlinear reparametrizations may be required. Graphical methods for exploring likelihood contours can be helpful in revealing this and similar problems and in suggesting nonlinear reparametrizations to correct them.

## 3. SPECIAL FEATURES OF BAYESIAN MONTE-CARLO ANALYSIS

Implementing Bayesian statistical analysis calls for the numerical estimation of several integrals such as those in Equations (2–4) which share certain features (illustrated in the example below) that make some Monte Carlo techniques seem especially appropriate. Each of the integrands in the numerators and denominators of equations (2–4) is a product of three terms: a multidimensional function of interest like $g(\theta)$, $(g(\theta)-\overline{g})$, $K_\epsilon(\xi_j-g(\theta))$, or simply 1; a nonnegative likelihood function $f(x|\theta)$; and a nonnegative prior density function, $\pi(\theta)$. In many applications:

1. The integrand is high-dimensional. The function $g(\theta)$ often has hundreds of components (*e.g.*, in kernel density estimation) each of which is easy to compute (*e.g.*, components might include $[\theta]_j$, $[\theta]_j[\theta]_k$, $1_{A_j}(\theta)$, or $K_\epsilon(\xi_j-g(\theta))$). Using Monte Carlo techniques, all the terms in Equations (4–6) can be calculated simultaneously using the same sequence of random deviates $\{\theta_i\}$.
2. The likelihood function $L(\theta) = f(x|\theta)$ in each of the required integrals is often slow and expensive to compute. When no sufficient statistics are available (*e.g.*, when exponential families are inappropriate), calculating $f(x|\theta)$ may require a loop through the entire dataset for each distinct value of $\theta$. The computational burden of generating random deviates $\theta_i \sim \Pi(d\theta)$ is usually negligible when compared to that of calculating $f(x|\theta_i)$, so there is little reward for using especially efficient random number generators; conversely, variance-reduction techniques are quite important to limit the number of points where $f(x|\theta_i)$ must be evaluated.
3. The prior density function $\pi(\theta)$ in each of the required integrals is also comparatively expensive to compute if subjective prior distributions are used (summarized in tables, or using density functions drawn with a "mouse.") Some prior densities intended to be "noninformative" are especially expensive computationally, *e.g.*, the *reference priors* of Berger and Bernardo (1989) or even Jeffreys' priors (1960) in multi-dimensional models which are not exponential families.
4. For many problems the product of the likelihood function and prior density is unimodal and "bell-shaped," *i.e.*, the negative logarithm $\ell(\theta) \equiv -\log[f(x|\theta)\pi(\theta)]$ is well approximated by a quadratic form in a neighborhood of its minimum.
5. High precision isn't important; $\pm 5\%$ or $\pm 1\%$ is usually quite adequate. Other uncertainties and approximations arising in the modeling process usually make it inappropriate to seek machine accuracy (6-16 decimals) in statistical calculations.
6. Integrals are moderately high-dimensional. In applications to structured, hierarchical Bayesian models and in routine economic time series $\Theta$ often range from 4–24 dimensions or more, while latent variable models and nonparametric survival or density estimation lead to problems with hundreds of variables.

7. As insights and experience are gained the function of interest $g(\theta)$ often changes; as new data become available the likelihood function $f(x|\theta)$ changes; even the prior density $\pi(\theta)$ may change. "Optimal" methods tailored to a particular integrand aren't helpful unless the tailoring can be done and revised almost instantly.
8. Speed is important in some applications and not in others. For interactive elicitation of prior densities in hierarchical models it would be useful to calculate predictive distributions and prior marginal densities within seconds. This is not yet possible for any but the simplest of models.

## 4. A BIOASSAY EXAMPLE

In Wolpert and Warren-Hicks (1990) details are given of a hierarchical Bayesian analysis combining laboratory data and field observations to study the effects on fish survival of three features often associated with so-called "acid rain" (low pH and high concentrations of calcium and monomeric aluminum). Multicollinearity in some observational datasets makes it difficult or impossible to use all three quantities to good advantage in a model selected through the use of field observations alone; hierarchical Bayesian models provide a coherent logical structure for combining field data and bioassay data, despite uncertainties about the differences between field and laboratory settings, and offer a way to circumvent the multicollinearity.

### Threat and Tolerance

Denote by $\mathbf{X}_i$ the vector of explanatory variables $\mathbf{X}_i = (\mathrm{pH}_i, \log[\mathrm{Al}]_i, \log[\mathrm{Ca}]_i)$ associated with the water chemistry of some lake (indexed by $i$). With each such vector is associated a "threat" $\zeta_i$, but the correspondence $\mathbf{X}_i \mapsto \zeta_i = \zeta(\mathbf{X}_i)$ is uncertain. Initially take the association to be linear

$$\zeta_i = \mathbf{X}_i\beta = X_{i1} + X_{i2}\beta_2 + X_{i3}\beta_3$$

for an uncertain 3-vector $\beta$, normalized by the constraint $\beta_1 = 1$. With each lake in the field dataset associate an uncertain "tolerance" $\tau_i$ representing unrecorded and uncontrolled environmental factors (food supply, water temperature, *etc.*) which may affect the lake's ability to support a brook trout population, with the understanding that the lake *will* support brook trout if $\tau_i \ge \zeta_i$ and it *will not* support trout if $\tau_i < \zeta_i$. If tolerances are taken to be random, drawn independently of $\mathbf{X}_i$ from a specified location-scale family with standardized CDF $\Psi(x)$ (say, the logistic) and uncertain location and scale parameters $\alpha_F$ and $\sigma_F$, then the probability that a lake with the explanatory variables $\mathbf{X}_i$ would be viable for brook trout would be

$$\mathbf{P}[\tau_i \ge \zeta_i] = 1 - \Psi\left(\frac{\zeta_i - \alpha_F}{\sigma_F}\right) = \left(1 + e^{(\mathbf{X}_i\beta - \alpha_F)/\sigma_F}\right)^{-1}.$$

In the laboratory bioassay experiments fish are presumed to be endowed with unobservable logistically distributed tolerances $\tau_j$ and are presumed to experience a constant hazard of $h_j = 0$ if $\tau_j \ge \mathbf{X}_j\beta$, and $h_j = (\mathbf{X}_j\beta - \tau_j)c$ if $\tau_j < \mathbf{X}_j\beta$; it follows that the probability $p_j$ of surviving for at least the duration $t_j$ of the bioassay is

$$p_j = Z_j + \frac{Z_j(1-Z_j)}{1 + c\sigma_L t_j}\ {}_2F_1(1, 2; 1 + c\sigma_L t_j; 1 - Z_j)$$

where $Z_j = \left(1 + e^{(\mathbf{X}_j\beta - \alpha_L)/\sigma_L}\right)^{-1}$ is the probability that $\mathbf{P}[\tau_j \geq \mathbf{X}_j\beta]$ and where ${}_2F_1(a,b;c;z)$ is Gauss' hypergeometric function. The second term in this expression is the probability of a right-censored death time. The likelihood function for the seven-dimensional parameter $\theta = (\beta_2, \beta_3, \alpha_F, \sigma_F, \alpha_L, \sigma_L, c)$ on the basis of 177 field observations ($P_i = 1$ for presence, $P_i = 0$ for absence of fish) and 164 bioassay observations ($S_j$ surviving and $D_j$ dead fish) is

$$\begin{aligned} L(\theta) = &\prod_{i=1}^{177} \left(1 + e^{(\mathbf{X}_i\beta - \alpha_F)/\sigma_F}\right)^{-P_i} \left(1 + e^{-(\mathbf{X}_i\beta - \alpha_F)/\sigma_F}\right)^{P_i - 1} \\ &\times \prod_{j=1}^{164} \left(Z_j + \frac{Z_j(1-Z_j)}{1 + c\sigma_L t_j}\ {}_2F_1(1,2;1 + c\sigma_L t_j; 1 - Z_j)\right)^{S_j} \\ &\times \prod_{j=1}^{164} \left(1 - Z_j - \frac{Z_j(1-Z_j)}{1 + c\sigma_L t_j}\ {}_2F_1(1,2;1 + c\sigma_L t_j; 1 - Z_j)\right)^{D_j} \end{aligned}$$

The functions of interest include the marginal densities for each of the four parameters $(\beta_2, \beta_3, \alpha_F, \sigma_F)$ governing field observations, the posterior mean and covariance matrix for these parameters, and the predictive distributions for the probability of fish presence $\left(1 + e^{(\mathbf{X}\beta - \alpha_F)/\sigma_F}\right)^{-1}$ for several specified chemistries $X$ (all of which are reported in Wolpert and Warren-Hicks (1990)). Features of this model include:

1. High-dimensional integrand (several hundred dimensions for the density estimation).
2. Slow computation for the likelihood function; transcendental and even special functions must be calculated for each observation at every point $\theta$.
3. Slow computation of the prior density, if the Jeffreys prior is used; again transcendental functions must be calculated inside a loop, and now a $7 \times 7$ determinant must be evaluated as well.
4. A unimodal and "bell-like" likelihood function (after logarithmic transformations for $\sigma_F$ and $\sigma_L$), with nearly elliptical contours for each pair of parameters.
5. Satisfactory and attainable precision of $\pm 1\%$ with several minutes' computation on a desktop Unix workstation.
6. Modest dimensionality (7) of the parameter space $\Theta$. We also consider models with quadratic terms in the uncertain dependence of threat $\zeta$ upon the three explanatory variables, increasing the dimension to 13; the dimension increases quickly as higher-order terms or more explanatory variables are added.
7. Changing *functions of interest* (predictive distributions were added), *likelihood function* (observations were added, and probit models were considered), and *prior density* (both uniform and Jeffreys priors were studied).
8. Badly conditioned linear approximations in the original coordinate system; the elliptical likelihood contours are highly eccentric, and the condition numbers for the information matrices at the posterior modes are high.

Wolpert and Warren-Hicks (1990) found approximations to the predictive survival distributions and posterior parameter distributions for this model using Monte Carlo methods with antithetic variates drawn from multivariate Student $t$ sampling distributions.

A two-step adaptive importance sampling scheme was used with initial location vectors and dispersion matrices suggested by an analysis of the likelihood function's behaviour near its maximum, later improved using Monte Carlo estimates of the mean and covariance. The degrees-of-freedom parameters were chosen to match the tail behaviour of the likelihood function.

## 5. APPENDIX: IMPLEMENTATION DETAILS

Elementary statistics textbooks sometimes give "shortcuts" for calculating the sample mean and variance from running totals $\mathrm{S}_X \equiv \sum X_i$ and $\mathrm{S}_{XX} \equiv \sum X_i^2$; unfortunately these widely-implemented formulas represent $S_n^2$ as the small difference of two large numbers, and so entail a great loss of precision. To find $S_3^2 = 2/3$ correctly to $d$ decimals using this technique for the data set $\{2999, 3000, 3001\}$, for example, requires that all intermediate calculations be carried out accurately to about $7 + d$ decimals; even for this simple problem *zero* decimals are available in Fortran single-precision! The expected number of bits of precision lost is about $\log_2\left(1 + (\mu/\sigma)^2\right)$ in general, or 23 bits for the example.

One remedy for this unnecessary loss of precision is to use the defining relation for $S_n^2$: first compute $\overline{X}_n$, then sum the squared deviations $(X_i - \overline{X}_n)^2$. This is unattractive because it requires that all $n$ observations be stored while the usual but imprecise formulas require only the three summary statistics $N$, $\mathrm{S}_X$, and $\mathrm{S}_{XX}$. An alternative is to initialize $\overline{X}_0 \equiv 0$ and $S_0^2 \equiv 0$, and then for $n \geq 1$ use the recursive formulas:

$$\begin{aligned}
\Delta_n \equiv& [X_n - \overline{X}_{n-1}] \\
\overline{X}_n \equiv& \overline{X}_{n-1} + \frac{1}{n}\Delta_n \\
S_n^2 \equiv& [S_{n-1}^2 + \frac{1}{n}\Delta_n^2](1 - \frac{1}{n})
\end{aligned}$$

which lead to full-precision values for $\overline{X}_n$ and $S_n^2$.

The same precision problem arises in calculating the weighted mean $\overline{g}_n$ in Equation (6) and especially in estimating the precision (7); the use of antithetic variates to reduce $\sigma$ can exacerbate the problem, since $[1 + (\mu/\sigma)^2]$ is then so large. The algorithm presented below gives high-precision recursive estimates $\overline{g}_n$ of the mean vector $\overline{g}$, $\ddagger\Sigma_n$ of the covariance matrix $\ddagger\Sigma$, and $\mathrm{MSE}_n$ of the estimation error matrix $\mathbf{E}\,(\overline{g}_n - \overline{g})(\overline{g}_n - \overline{g})'$, for any antithetic sampling scheme.

### Recursive Estimation Formulas

Let $\pi(d\theta)$ be a prior measure and $L(\theta) = f(x|\theta)$ a likelihood function on some measure-space $\big(\Theta,\ \mathcal{F},\ d\theta\big)$ and let $g(\theta)$ be an $\mathbb{R}^q$ valued measurable function on $\Theta$. The problem at hand is to find a sequence of estimates

$$\begin{aligned}
\overline{g}_n \approx\ \overline{g}\ \equiv& \frac{\int g(\theta)L(\theta)\,\pi(d\theta)}{\int L(\theta)\,\pi(d\theta)} \\
\ddagger\Sigma_n \approx\ \ddagger\Sigma \equiv& \frac{\int \big(g(\theta) - \overline{g}\big)\big(g(\theta) - \overline{g}\big)' L(\theta)\,\pi(d\theta)}{\int L(\theta)\,\pi(d\theta)} \\
\mathrm{MSE}_n \approx\ \mathbf{E}& \left[\big(\overline{g}_n - \overline{g}\big)\big(\overline{g}_n - \overline{g}\big)' \mid X{=}x\right].
\end{aligned}$$

For each $n$ choose an integer $m_n$ and draw an antithetic series $\theta_1^{(n)},...,\theta_{m_n}^{(n)}$ consisting of $m_n$ not-necessarily-independent draws $\theta_j^{(n)}$, each with marginal distribution $\Pi(d\theta)$. The number $m_n$ of draws need not be constant as $n$ varies. Within each antithetic series estimates will be calculated of $\overline{g}$ and $\Sigma$, each an average weighted by $\mathrm{w}_j^{(n)} = w(\theta_j^{(n)})$ at the observed points $\theta_j^{(n)}$.

*Within each antithetic series*

For each $n$ set $\mathrm{w}_0^{(n)} \equiv 0$, $\overline{g}_0^{(n)} \equiv 0$, $\mathrm{msq}_0^{(n)} \equiv 0$. For $1 \le j \le m_n$, set

$$\begin{aligned}
\omega_j^{(n)} &\equiv w(\theta_j^{(n)}) && = L(\theta_j^{(n)})\,\pi(d\theta_j^{(n)})/\Pi(d\theta_j^{(n)}) \\
\mathrm{w}_j^{(n)} &\equiv \mathrm{w}_{j-1}^{(n)} + \omega_j^{(n)} && = \sum_{i\le j} \omega_i^{(n)} \\
h_j^{(n)} &\equiv \omega_j^{(n)}/\mathrm{w}_j^{(n)} && \\
\\
g_j^{(n)} &\equiv g(\theta_j^{(n)}) && \\
\delta_j^{(n)} &\equiv g_j^{(n)} - \overline{g}_{j-1}^{(n)} && \\
\overline{g}_j^{(n)} &\equiv \overline{g}_{j-1}^{(n)} + h_j^{(n)}\delta_j^{(n)} && = \sum_{i\le j} \omega_i^{(n)} g_i^{(n)} \Big/ \sum_{i\le j} \omega_i^{(n)} \\
\mathrm{msq}_j^{(n)} &\equiv (1-h_j^{(n)})\big[\mathrm{msq}_{j-1}^{(n)} + h_j^{(n)}(\delta_j^{(n)})(\delta_j^{(n)})'\big] && \\
& && = \sum_{i\le j} \omega_i^{(n)} (g_i^{(n)} - \overline{g}_j^{(n)})(g_i^{(n)} - \overline{g}_j^{(n)})' \Big/ \sum_{i\le j} \omega_i^{(n)}
\end{aligned}$$

The four summary statistics from the $n^{th}$ antithetic series are:

| | | |
|---|---|---|
| $m_n$ | | The number of function evaluations in the series; |
| $\mathrm{w}_n$ | $\equiv \mathrm{w}_{m_n}^{(n)}$ | The total weight for the series; |
| $g_n$ | $\equiv \overline{g}_{m_n}^{(n)}$ | The $w(\theta_j^{(n)})$-weighted average of the vectors $g(\theta_j^{(n)})$; |
| $\mathrm{msq}_n$ | $\equiv \mathrm{msq}_{m_n}^{(n)}$ | The mean-square variation, *i.e.*, the $w(\theta_j^{(n)})$-weighted average of the matrices $(g_i^{(n)} - g_n)(g_i^{(n)} - g_n)'$. |

There is no information in these "within" statistics about the variability arising from the importance sampling because of the allowed dependence among the $\theta_j^{(n)}$ for a given $n$. This variability will be reflected in a "between" mean-square summarizing the variability of the vector quantities $g_n$ across repeated independent replicates.

If samples were drawn directly from the posterior density, the function $w(\theta)$ would be constant and $g_n$ and $\mathrm{msq}_n$ would be the unweighted average and sample variance of the $\{g_i\}_{i\le m_n}$, respectively. In the common case $m_n = 2$ the formulas above reduce to

$$\begin{aligned}
m_n &= 2 \\
\mathrm{w}_n &= \omega_1^{(n)} + \omega_2^{(n)} \\
g_n &= \left(\tfrac{\omega_1^{(n)}}{\omega_1^{(n)}+\omega_2^{(n)}}\right) g_1^{(n)} + \left(\tfrac{\omega_2^{(n)}}{\omega_1^{(n)}+\omega_2^{(n)}}\right) g_2^{(n)} \\
\mathrm{msq}_n &= \left(\tfrac{\omega_1^{(n)}}{\omega_1^{(n)}+\omega_2^{(n)}}\right)\left(\tfrac{\omega_2^{(n)}}{\omega_1^{(n)}+\omega_2^{(n)}}\right)\big(g_1^{(n)} - \overline{g}_2^{(n)}\big)\big(g_1^{(n)} - \overline{g}_2^{(n)}\big)'
\end{aligned}$$

or simply $w_n = 2$, $\overline{g} = (g_1^{(n)} + g_2^{(n)})/2$ and $\mathrm{msq}_n = (g_1^{(n)} - \overline{g}_2^{(n)})(g_1^{(n)} - \overline{g}_2^{(n)})'/4$ for direct sampling from the posterior.

*Combining Antithetic series*

For estimating mean-square estimation errors it is useful to have both w-weighted and $\mathrm{w}^2$-weighted averages; the latter we distinguish with asterisk superscripts. Initialize $N_n \equiv 0$, $W_0 \equiv 0$, $W_0^* \equiv 0$, $\overline{g}_0 \equiv 0$, $\overline{g}_0^* \equiv 0$, $\Sigma_0 \equiv 0$, $\mathrm{MSB}_0 \equiv 0$, and $\mathrm{MSB}_0^* \equiv 0$; with each succeeding $n$ set

$$\begin{aligned}
N_n &\equiv N_{n-1} + m_n &&= \sum_{j\le n} m_j \\
\mathrm{w}_n &\equiv \mathrm{w}_{m_n}^{(n)} &&= \sum_{i\le m_n} \omega_i^{(n)} \\
g_n &\equiv \overline{g}_{m_n}^{(n)} &&= \sum_{i\le m_n} \omega_i^{(n)} g_i^{(n)} \big/ \mathrm{w}_n \\
\mathrm{msq}_n &\equiv \mathrm{msq}_{m_n}^{(n)} &&= \sum_{i\le m_n} \omega_i^{(n)} (g_i^{(n)} - g_n)(g_i^{(n)} - g_n)' \big/ \mathrm{w}_n
\end{aligned}$$

$$\begin{aligned}
W_n &\equiv W_{n-1} + \mathrm{w}_n &&= \sum_{j\le n} \mathrm{w}_j & \qquad W_n^* &\equiv W_{n-1}^* + (\mathrm{w}_n)^2 &&= \sum_{j\le n} (\mathrm{w}_j)^2 \\
H_n &\equiv \mathrm{w}_n / W_n & & & H_n^* &\equiv (\mathrm{w}_n)^2 / W_n^* & & \\
\Delta_n &\equiv [g_n - \overline{g}_{n-1}] & & & \Delta_n^* &\equiv [g_n - \overline{g}_{n-1}^*] & & \\
\overline{g}_n &\equiv \overline{g}_{n-1} + H_n \Delta_n &&= \sum_{j\le n} \mathrm{w}_j g_j / W_n & \qquad \overline{g}_n^* &\equiv \overline{g}_{n-1}^* + H_n^* \Delta_n^* &&= \sum_{j\le n} (\mathrm{w}_j)^2 g_j / W_n^*
\end{aligned}$$

$$\begin{aligned}
\mathrm{MSW}_n &\equiv (1-H_n)\mathrm{MSW}_{n-1} + H_n\,\mathrm{msq}_n \\
&= \frac{\sum_{j\le n}\mathrm{w}_j\,\mathrm{msq}_j}{\sum_{j\le n}\mathrm{w}_j} \\
&= \frac{\sum_{j\le n}\sum_{i\le m_j}\omega_i^{(j)}(g_i^{(j)}-g_j)(g_i^{(j)}-g_j)'}{\sum_{j\le n}\sum_{i\le m_j}\omega_i^{(j)}} \\
\mathrm{MSB}_n &\equiv (1-H_n)\big[\mathrm{MSB}_{n-1} + H_n(\Delta_n)(\Delta_n)'\big] \\
&= \frac{\sum_{j\le n}\mathrm{w}_j(g_j-\overline{g}_n)(g_j-\overline{g}_n)'}{\sum_{j\le n}\mathrm{w}_j} \\
\mathrm{MSB}_n^* &\equiv (1-H_n^*)\big[\mathrm{MSB}_{n-1}^* + H_n^*(\Delta_n^*)(\Delta_n^*)'\big] \\
&= \frac{\sum_{j\le n}(\mathrm{w}_j)^2(g_j-\overline{g}_n^*)(g_j-\overline{g}_n^*)'}{\sum_{j\le n}(\mathrm{w}_j)^2} \\
\Sigma_n &\equiv \mathrm{MSW}_n + \mathrm{MSB}_n \\
&= H_n\,\mathrm{msq}_n + (1-H_n)\big[\Sigma_{n-1} + H_n(\Delta_n)(\Delta_n)'\big] \\
&= \frac{\sum_{j\le n}\sum_{i\le m_j}\omega_i^{(j)}(g_i^{(j)}-\overline{g}_n)(g_i^{(j)}-\overline{g}_n)'}{\sum_{j\le n}\sum_{i\le m_j}\omega_i^{(j)}}
\end{aligned}$$

*The Estimates*

With these in hand we can now estimate $\overline{g}$, the posterior expectation of $g(\theta)$, by

$$\begin{aligned}
\overline{g}_n &= \textstyle\sum\sum\omega_i^{(j)}g_i^{(j)} \Big/ \sum\sum\omega_i^{(j)} \\
&\approx \int w(\theta)g(\theta)\Pi(d\theta) \Big/ \int w(\theta)\Pi(d\theta) \\
&= \int g(\theta)L(\theta)\,\pi(d\theta) \Big/ \int w(\theta)\Pi(d\theta) \\
&= \overline{g}
\end{aligned}$$

and the posterior variance by

$$\begin{aligned}
\not\Sigma_n &\equiv \mathrm{MSW}_n + \mathrm{MSB}_n \\
&= \textstyle\sum\sum \omega_i^{(j)} (g_i^{(j)} - \overline{g}_n)(g_i^{(j)} - \overline{g}_n)' / \sum\sum \omega_i^{(j)} \\
&\approx \int w(\theta)(g(\theta) - \overline{g})(g(\theta) - \overline{g})' \Pi(d\theta) \Big/ \int w(\theta) \Pi(d\theta) \\
&= \int (g(\theta) - \overline{g})(g(\theta) - \overline{g})' L(\theta)\, \pi(d\theta) \Big/ \int L(\theta)\, \pi(d\theta) \\
&= \not\Sigma
\end{aligned}$$

The mean-square error of estimation is more interesting. A second-order Taylor-series expansion of the function $f(x,y,z) = xy/z^2$ about the means $\mu_X{=}0$, $\mu_Y{=}0$, and $\mu_Z{\neq}0$ of three real-valued random variables $X$, $Y$, and $Z$ (not necessarily independent) reveals that $\mathbf{E}(XY/Z^2) \approx (\mathbf{E}\, XY)/(\mathbf{E}\, Z)^2$. Thus the mean-square estimation error is

$$\begin{aligned}
\mathbf{E}(\overline{g}_n - \overline{g})(\overline{g}_n - \overline{g})' &= \mathbf{E} \left( \frac{\sum_{j\le n} \mathrm{w}_j (g_j - \overline{g})}{\sum_{j\le n} \mathrm{w}_j} \right) \left( \frac{\sum_{j\le n} \mathrm{w}_j (g_j - \overline{g})}{\sum_{j\le n} \mathrm{w}_j} \right)' \\
&\approx \frac{\mathbf{E}\big( \sum_{j\le n} \mathrm{w}_j (g_j - \overline{g}) \big) \big( \sum_{j\le n} \mathrm{w}_j (g_j - \overline{g}) \big)'}{\big( \mathbf{E} \sum_{j\le n} \mathrm{w}_j \big)^2} \\
&= \frac{\mathbf{E} \sum_{j\le n} (\mathrm{w}_j)^2 (g_j - \overline{g})(g_j - \overline{g})'}{\big( \mathbf{E} \sum_{j\le n} \mathrm{w}_j \big)^2} \\
&\approx \frac{\sum_{j\le n} (\mathrm{w}_j)^2 (g_j - \overline{g}_n)(g_j - \overline{g}_n)'}{\big( \sum_{j\le n} \mathrm{w}_j \big)^2} \\
&= \frac{1}{(W_n)^2} \sum_{j\le n} (\mathrm{w}_j)^2 \big( g_j - \overline{g}_n^* + \overline{g}_n^* - \overline{g}_n \big) \big( g_j - \overline{g}_n^* + \overline{g}_n^* - \overline{g}_n \big)' \\
&= \frac{1}{(W_n)^2} \sum_{j\le n} (\mathrm{w}_j)^2 \big[ \big( g_j - \overline{g}_n^* \big) \big( g_j - \overline{g}_n^* \big)' + \big( \overline{g}_n^* - \overline{g}_n \big) \big( \overline{g}_n^* - \overline{g}_n \big)' \big] \\
= \mathrm{MSE}_n \equiv \frac{W_n^*}{(W_n)^2} &\left[ \mathrm{MSB}_n^* + (\overline{g}_n^* - \overline{g}_n)(\overline{g}_n^* - \overline{g}_n)' \right]
\end{aligned}$$

This error estimate has two components: $\mathrm{MSB}_n^*$, a measure of how much the $g_j$ differ, and $(\overline{g}_n^* - \overline{g}_n)(\overline{g}_n^* - \overline{g}_n)'$, a measure of how much the $w_j$ differ. The first component can be made small by capturing as much variability as possible *within* each antithetic series (in $\mathrm{MSW}_n$) and leaving as little as possible in $\mathrm{MSB}_n \approx \mathrm{MSB}_n^*$, while the second component can be made small by sampling from an importance function similar to the posterior distribution so the weights $w_j$ will be nearly constant and hence the $(w_j)$-weighted and $(w_j)^2$-weighted means $\overline{g}_n$ and $\overline{g}_n^*$ will be nearly equal.

With *iid* sampling from the posterior density function the mean-square error in esti-

mating the $i^{th}$ component of $\overline{g}$ would have been

$$\begin{aligned}\frac{1}{N_n}\Sigma^{ii} &\approx \frac{1}{N_n}\Sigma_n^{ii}\\ &= \mathrm{RE}_n^i \times \mathrm{MSE}_n^{ii}\end{aligned}$$

where, following Hammersley and Handscomb, the "relative efficiency" for the $i^{th}$ component is defined to be

$$\begin{aligned}\mathrm{RE}_n^i &\equiv \left(\frac{1}{N_n}\Sigma_n^{ii}\right)/\mathrm{MSE}_n^{ii}\\ &= \frac{(W_n)^2\left[\mathrm{MSW}_n^{ii}+\mathrm{MSB}_n^{ii}\right]}{N_n W_n^*\left[\mathrm{MSB}_n^{*ii}+|(\overline{g}_n^*-\overline{g}_n)_i|^2\right]}.\end{aligned}$$

The RE indicates the efficiency of a given importance-sampling scheme, relative to the benchmark of *iid* sampling directly from the posterior distribution. The given procedure with $n$ function evaluations attains the same precision as would *iid* sampling with RE$\times n$ function evaluations. For some problems relative efficiencies of well over 100% are possible with well-chosen antithetic schemes.

## 6. REFERENCES


Berger, J.O. and Wolpert, R.L. (1988). *The Likelihood Principle (2nd Edition).* Institute of Mathematical Statistics Press, Hayward.

Berger, J.O. and Bernardo, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association* **84**, 200–207.

Curtiss, J.H. *et al.* (1951). Monte Carlo method. *National Bureau of Standards Applied Mathematics Series,* **12**.

Fieller, E.C. and Hartley, H.O. (1954). Sampling with control variables. *Biometrika* **41**, 494–501.

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.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geweke, J. (1988). Antithetic acceleration of Monte Carlo integration in Bayesian inference. *Journal of Economics* **38**, 73–89.

Hammersley, J.M. and Morton, K.W. (1956). A new Monte Carlo technique: antithetic variates. *Proceedings of the Cambridge Philosophical Society* **52**, 449–475.

Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods.* Chapman and Hall, London.

Jeffreys, H. (1960). *Theory of Probability (3rd edition).* Oxford University Press, London.

Kloek, T. and van Dijk, H.K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica* **46**, 1–19.

Naylor, J.C. and Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics* **31**, 214–225.

Rubinstein, R.Y. (1981). *Simulation and the Monte Carlo method.* John Wiley and Sons, New York.

Rubinstein, R.Y. (1986). *Monte Carlo optimization, simulation, and sensitivity of queueing networks.* John Wiley and Sons, New York.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Communications in Statistics* **A14**, 1079–1102.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Tukey, J.W. (1957). Antithesis or regression? *Proceedings of the Cambridge Philosophical Society* **53**, 923–924.

Wilson, J.R. (1984). Variance reduction techniques for digital simulation. *American Journal of Mathematics and Management Science* **4**, 277–313.

Wolpert, R.L. and Warren-Hicks, W.J. (1990). Bayesian hierarchical logistic models for combining field and laboratory data. *In preparation.*