Chapter 9: Subjective Probability and Bayesian Methodology

Stephen E. Chick

INSEAD; Technology Management Area;
Boulevard de Constance; 77300 Fontainebleau; France

Abstract

Subjective probability and Bayesian methods provide a unified approach to handle not only randomness from stochastic sample-paths, but also uncertainty about input parameters and response metamodels. The chapter surveys some basic concepts, principles and techniques useful for a subjective Bayesian approach to uncertainty analysis, data collection plans to reduce input uncertainty, response surface modeling, and expected value-of-information approaches to experimental designs for selection procedures. Some differences from the classical technique are identified.

If simulation is defined to be the analysis of stochastic processes through the generation of sample paths of the process, then Bayesian and subjective probability methods apply in several ways for the modeling, design and analysis of simulation experiments. By Bayesian methods, we refer here to parameter inference through repeated observations of data with Bayes' rule. Examples in simulation are input parameter inference using field data or the inference of metamodel parameters from simulation replications. The Bayesian approach entails postulating a 'prior probability' model that describes a modeler's initial uncertainty about parameters, a likelihood function that describes the distribution of data, given that a parameter holds a specific value, and Bayes' rule, which provides a coherent method of updating beliefs about uncertainty when data becomes available. By subjective probability, we refer to probability assessments for all unknown quantities, including parameters that can be inferred with Bayes' rule, as well as unknown quantities for which parameters cannot be inferred from repeated sampling of data (e.g. one-shot deals like the total potential market size for a particular new product from a simulated manufacturing facility). By frequentist, we mean methods based on sampling statistics from repeated observations, such as maximum likelihood (MLE) methods to fit input parameters, or ranking and selection procedures

Email address: `stephen.chick@insead.edu` (Stephen E. Chick).

that provide worst-case probability of correct selection guarantees based on repeated applications of the procedure. The chapter describes applications of Bayesian and subjective probability methods in simulation, and identifies some ways that the Bayesian approach differs from the frequentist approach that underlies much of simulation theory.

In the simulation community, Glynn (1986) first suggested Bayesian applications of *uncertainty analysis* for statistical input parameter uncertainty. In that paper, the traditional role of estimating $\alpha = h(E[Y])$ is extended to account for statistical input parameter uncertainty, so $\alpha(\theta) = h(E[Y \mid \theta])$ depends upon unknown parameters with distribution $p(\theta)$ that can be updated with data from the modeled system. Three questions he poses are (i) how to estimate the distribution of $\alpha(\Theta)$ induced by the random variable Θ, (ii) how to estimate the mean $E[\alpha(\Theta)]$, and (iii) estimation of credible sets, e.g. finding a, b so the probability $\Pr(\alpha(\Theta) \in [a, b])$ equals a prespecified value, like 0.95. Chick (1997) provided a review of the few works to that date that applied Bayesian ideas to simulation, then suggested a broader range of application areas than uncertainty analysis, including *ranking and selection*, *response surface modeling*, and *experimental design*.

The basic goal is to understand how uncertainty and decision variables affect system performance, so that better decisions can be made. The premise in this chapter is that representing *all* uncertainty with probability can aid decision-makers that face uncertainty. Stochastic uncertainty, the randomness in simulation models that occurs even if all parameters are known, is already widely modeled with probability. The subjective Bayesian approach also models input parameter and response surface uncertainty with probability distributions, a practice that has been less common in stochastic process simulation.

Probabilistic models for uncertainty are increasingly employed for at least three reasons. One, doing so allows the modeler to quantify how parameter uncertainty influences the performance of a simulated system. Parameters of models of real systems are rarely known with certainty. The Bayesian approach for uncertainty analysis overcomes some limitations of the classical approach for parameter and model selection (Chick, 2001; Barton and Schruben, 2001; Draper, 1995). Two, simulation experiments can be designed to run more efficiently (Chick and Inoue, 2001a; Santner et al., 2003). And three, Bayesian and subjective probability methods are not new but are increasingly implemented due to the development of improved computing power and Markov Chain Monte Carlo (MCMC) methods (Gilks et al., 1996).

This chapter describes the subjective Bayesian formulation for simulation. Section 1 presents the basics of subjective probability and Bayesian statistics in the context of quantifying uncertainty about one statistical input parameter. Section 2 summarizes the main ideas and techniques for addressing three

main challenges in implementing Bayesian inference: maximization, integration, and sampling variates from posterior distributions. Section 3 addresses input distribution selection when multiple candidate distributions exist. Section 4 presents a joint formulation for input and output modeling, and reviews applications for data collection to reduce input uncertainty in a way that reduces output uncertainty, and for response surface modeling and simulation experiments to reduce response surface uncertainty. Section 5 describes applications of Bayesian expected value of information methods for efficiently selecting the best of a finite set of simulated alternatives.

Simulation research with Bayesian methods has grown rapidly since the mid to late 1990s. A partial reference list is Chen and Schmeiser (1995); Chen (1996); Scott (1996); Nelson et al. (1997); Chen et al. (1999); Cheng (1999); Lee and Glynn (1999); Andradóttir and Bier (2000); Chick and Inoue (2001a,b); Chick (2001); Cheng and Currie (2003); Steckley and Henderson (2003); Chick et al. (2003); Zouaoui and Wilson (2003, 2004); Ng and Chick (2002), as well as applications to insurance, finance, waterway safety, civil engineering and other areas described in the Winter Simulation Conference Proceedings. Work on deterministic simulation with potentially important implications for stochastic simulation includes O'Hagan et al. (1999); Kennedy and O'Hagan (2001); Craig et al. (2001); Santner et al. (2003). Excellent references for subjective probability and Bayesian statistics in general, not just in simulation, include Lindley (1972); Berger (1985); Bernardo and Smith (1994), with special mention for de Finetti (1990), Savage (1972), and de Groot (1970).

1 Main Concepts

A stochastic simulation is modeled as a deterministic function of several inputs,

$$Y_r = g(\boldsymbol{\theta}_p, \boldsymbol{\theta}_e, \boldsymbol{\theta}_c; \boldsymbol{U}_r), \tag{1}$$

where Y_r is the output of the r-th replication. The vector of statistical input parameters $\boldsymbol{\theta}_p = (\theta_1, \theta_2, \ldots, \theta_{n_p})$ describes n_p sources of randomness whose values can be inferred from field data. For example, θ_1 may be a two-dimensional parameter for lognormally distributed service times, and θ_2 may be defect probabilities inferrable from factory data. Environmental parameters $\boldsymbol{\theta}_e$ are beyond the control of a decision maker, and no data is available for inference. Examples are the total potential market size for a new product, the general economic climate for a high-level model, or the actual climate for a production process influenced by temperature or humidity. The vector $\boldsymbol{\theta}_c$ represents all control parameters (decision variables) under direct control of the decision maker, such as production capacity, supply chain operating procedures, scheduling policies, and the number of servers at each node in a service system.

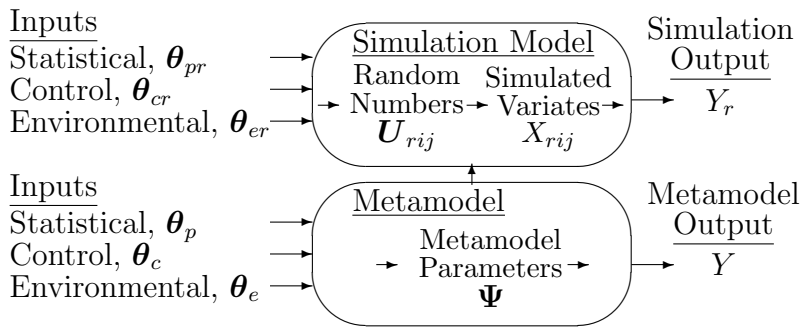

Fig. 1. Simulation Takes Multiple Types of Inputs and Metamodels Predict Outputs for Unsimulated Input Values

Random output Y_r for replication r can be generated even for the same inputs $(\boldsymbol{\theta}_{cr}, \boldsymbol{\theta}_{pr}, \boldsymbol{\theta}_{er})$ by sampling different portions of a random number stream $\boldsymbol{u}_r$, to obtain random variates x_{rij} (the j-th simulated variate using input parameter θ_i during replication r), as in the top row of Figure 1. We use upper case for random variables, lower case for realizations, and bold-face to emphasize that a quantity is a vector. We may suppress the r, as in X_{ij}, to describe data collected from the actual system being simulated. That data would be used to infer the parameters of the statistical distributions to describe the system.

One reason a simulation experiment may be run is to estimate the function $g()$ because its exact form is not known. **Metamodels** can be used to predict the output of a simulation model (or the simulated system) when a full simulation takes a long time to run. Section 4 describes some Bayesian methods to describe uncertainty about the parameters $\boldsymbol{\psi}$ of a metamodel.

A subjective probabilist represents all uncertain quantities with probability distributions. Uncertainty about statistical input parameters and environmental parameters are described as random quantities by the subjective Bayesian framework; we use a prior distribution $\pi(\boldsymbol{\theta}_p, \boldsymbol{\theta}_e)$ for the random quantity $(\boldsymbol{\Theta}_p, \boldsymbol{\Theta}_e)$. The specification of prior distributions and Bayesian inference with data are discussed in Section 1.1. Loss functions and the expected value of information follow in Section 1.2, with uncertainty analysis in Section 1.3.

1.1 Bayesian Modeling

An important simulation design issue is the selection of appropriate input distributions to characterize the stochastic behavior of the modeled system (see Chapter 6). Failure to select appropriate input distributions can result in misleading simulation output, and therefore poor system design decisions. This section reviews basic ideas and important theorems for inferring statistical parameters $\boldsymbol{\theta}_p$ from data with the Bayesian formalism. To simplify the discus-

sion, we focus on selecting a single statistical parameter θ for a given, fixed candidate model for input into a computer simulation. A candidate model could be, for example, a Bernoulli distribution. We therefore drop the extra subscripts from Equation (1) in this subsection. The subscripts are needed in later sections. Section 3 explores multiple candidate models for a given source of randomness.

For a Bayesian, the idea of exchangeability is preferred to the idea of independent and identically distributed (i.i.d.) random variables. Exchangeability is weaker than the i.i.d. assumption and plays a role in specifying probability models. Let $\boldsymbol{X}_N = (X_1, X_2, \ldots, X_N)$ be a generic vector of random variables on an outcome space Ω. A probability p on Ω is **exchangeable** if it is invariant with respect to permutations of the coordinates (e.g. $p(\boldsymbol{x}_n) = p(x_1, x_2, \ldots, x_n) = p(x_{h_1}, x_{h_2}, \ldots, x_{h_n})$ for permutations h on $\{1, 2, \ldots, n\}$ for arbitrary $n \leq N$.

Simulation is often concerned with conceptually infinite ($N \to \infty$) exchangeable sequences, e.g. no conceptual bound on the number of data observations or simulation replications. A key theorem (de Finetti, 1990; Bernardo and Smith, 1994) for infinite exchangeable sequences of Bernoulli random variables says that outcomes are conditionally independent, given the limiting fraction of heads, $\Theta = \lim_{N\to\infty} \sum_{i=1}^{N} X_i/N$, with some mixture distribution $\pi(\theta)$,

$$\lim_{N\to\infty} p(\boldsymbol{x}_n) = \int \left\{ \prod_{i=1}^{n} f(x_i \mid \theta) \right\} d\pi(\theta), \tag{2}$$

where $p(x_i \mid \theta) = f(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$ is viewed as a conditional probability when considered as a function of x_i and as a likelihood when written as a function of θ. A mixture written in the form of Equation (2) for an arbitrary parameter θ, distribution $\pi(\theta)$ and likelihood model f is called a **de Finetti-type representation**. The notation anticipates the convention of writing a **prior distribution** as $\pi(\cdot)$, representing the *a priori* belief that the parameter takes on a given value. Equation (2) is the basis for inference of statistical input parameter θ from data $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ via **Bayes' rule**,

$$p(\theta \mid \boldsymbol{x}_n) = \frac{\pi(\theta) p(\boldsymbol{x}_n \mid \theta)}{p(\boldsymbol{x}_n)} = \frac{\pi(\theta) \prod_{i=1}^{n} f(x_i \mid \theta)}{\int p(\boldsymbol{x}_n \mid \theta) d\pi(\theta)} \tag{3}$$

The first equality of Equation (3) is Bayes' rule and applies in general. The second equality follows from conditional independence. The **posterior probability** $p(\theta \mid \boldsymbol{x}_n)$ of θ given $\boldsymbol{x}_n$, summarizes uncertainty about θ via the likelihood model, the prior distribution and the data $\boldsymbol{x}_n$.

Bayesian methods require probability distributions to quantify initial uncertainty before data is observed. The **selection of a prior distribution** is controversial. Bruno de Finetti (1990) argues that a prior distribution is a

subjective expression of uncertainty, and that You (yes, You) may justifiably specify a different distribution than I, since we may have different beliefs about the likelihood of a given event. Savage (1972) suggests a process for eliciting a prior distribution from a modeler through the evaluation of 'fair bets' (as opposed to limiting frequencies). Kahneman et al. (1982) illustrate potential pitfalls with eliciting probability judgments and present techniques to counter them. While this may seem 'too subjective' and open to biases (Edwards, 1984), the ability to include prior information provides important flexibility and can be considered an advantage of the approach. Frequentist methods apply only with data, and problems remain (e.g. see Section 3).

To avoid the impression of subjectivity, several 'automated' mechanisms have nonetheless been proposed to support the selection of a prior distribution. When a lot of data is available, the likelihood function is the dominant term in Bayes' rule, rather than the prior distribution, so these methods may be helpful. The first approach is to obtain a prior distribution for a parameter through an *indifference judgment*. For example, for the unknown probability θ of a Bernoulli outcome, this would give a $\texttt{uniform}[0,1]$ distribution, the prior probability model used by Laplace (1812) to assess his prior probability that the sun would come up tomorrow. That approach is coordinate dependent (e.g. indifference over θ or $\log\theta$).

Jeffreys (1946) suggested $\pi(\theta) \propto |H(\theta)|^{1/2} d\theta$, where H is the expected information in one observation,

$$H(\theta) = E\left[-\frac{\partial^2 \log p(X \mid \theta)}{\partial \theta^2}\bigg|_{\theta} \right], \tag{4}$$

because it has the attractive property of being invariant with respect to coordinate changes in θ. It is 'uniform' with respect to the natural metric induced by the likelihood function (Kass, 1989). *Jeffreys' prior* for Bernoulli sampling is a $\texttt{beta}(1/2, 1/2)$ distribution. For some models, Jeffreys' prior is improper (does not integrate to one), but may be useful if the data results in a proper posterior after Bayes' rule is formally applied.

A third approach that is mathematically convenient is to assume a *conjugate prior* distribution, meaning that the posterior distribution has the same functional form as the prior distribution. For Bernoulli(θ) sampling, the $\texttt{beta}(\alpha, \beta)$ distribution with probability density function (pdf) $f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ is a conjugate prior. If data $\boldsymbol{x}_n$ is observed, with $s_n = \sum_{i=1}^n x_i$, then the posterior pdf is $f(\theta \mid \boldsymbol{x}_n) \propto \theta^{\alpha+s_n-1}(1-\theta)^{\beta+n-s_n-1}$, a $\texttt{beta}(\alpha+s_n, \beta+n-s_n)$ distribution. Conjugate prior distributions exist for all members of the regular exponential family (Bernardo and Smith, 1994), which includes the exponential, normal, gamma, lognormal, Wishart, Bernoulli, geometric, and Poisson distributions, and linear regression models with normally distributed error, among others. The $\texttt{uniform}[0,1]$ distribution is in the conjugate fam-

ily for Bernoulli sampling—it is a $\texttt{beta}(1,1)$ distribution. Prior distributions selected in this way are often selected to be as *noninformative* as possible, meaning that probability is spread 'evenly' over the space of parameters. Although 'evenly' is subjectively defined, heuristics are available for members of the regular exponential family, whose likelihood function can be written $p(x \mid \theta) = a(x)h_0(\theta)\exp\left[\sum_{j=1}^{d} c_j\phi_j(\theta)h_j(x)\right]$for some $a(\cdot), h_0(\cdot), c_j, \phi_j(\cdot), h_j(\cdot)$. The conjugate prior is $p(\theta) = [K(\boldsymbol{t})]^{-1}[h_0(\theta)]^{n_0}\exp\left[\sum_{j=1}^{d} c_j\phi_j(\theta)t_j\right]$, where $\boldsymbol{t} = (t_1, t_2, \ldots, t_d)$ is a hyperparameter. The posterior distribution given n conditionally independent data points then has parameters $n_0 + n$ and the sum of $\boldsymbol{t}$ and the sufficient statistics (Bernardo and Smith, 1994). The parameter n_0 is therefore interpreted by some to be the 'strength' of the prior, measured in terms of the number of samples. In that case, evenly spreading probability can be taken to mean selecting n_0 close to 0, while insuring that the prior is still proper.

Jaynes (1983) suggests a fourth approach that is common in image and signal processing: maximum entropy methods define 'diffuse' priors with respect to a background measure, subject to moment constraints on the parameters. Berger (1994) and Kass and Wasserman (1996) comment further on default prior distributions and sensitivity analysis with respect to them.

Probability modeling is inherently subjective—even so-called 'objective' methods require the subjective specification of a likelihood model. One standard Bayesian practice is to use a slightly informative conjugate distribution for the unknown mean, by choosing it to be proper but diffuse (Gilks et al., 1996). For example, the conjugate prior for an unknown mean of a normal distribution is also a normal distribution. A diffuse prior would be $\texttt{Normal}\left(0, \sigma_{big}^2\right)$ for some large σ_{big}^2. Conjugate prior distributions are mathematically convenient, but care is still required with their use, as with any statistical analysis, Bayesian or otherwise.

Classical asymptotic theorems (laws of large numbers, LLN; central limit theorems, CLT; e.g. Billingsley 1986) have Bayesian interpretations when considered to be conditional on the mean and standard deviation of an infinite exchangeable sequence. A Bayesian extension of the LLN allows for a sample average to converge to an 'unknown' mean (random variable) rather than to a 'true' mean.

Theorem 1 (Bayesian LLN) *Let X_i be an exchangeable sequence of random variables, and let $\bar{X}_n$ and $\bar{Y}_m$ be the averages of n and m of the X_i, respectively. If $Var[X_1] < \infty$, then the* probability *that*

$$\left|\bar{X}_n - \bar{Y}_m\right| > \epsilon$$

may be made arbitrarily small by taking n and m sufficiently large (de Finetti,

1990, p. 216).

Although the mode of a posterior distribution may not be the true mean, an asymptotic normality property holds for posterior distributions of parameters.

Theorem 2 (Posterior Normality) *For each n, let $p_n(\cdot)$ be the posterior pdf of the d-dimensional parameter θ_n given $\boldsymbol{x}_n = (x_1, \ldots, x_n)$, let $\tilde{\theta}_n$ be its mode, and define the $d \times d$ Bayesian observed information matrix Σ_n^{-1} by*

$$\Sigma_n^{-1} = -L_n''(\tilde{\theta}_n), \text{ where } L_n''(\vartheta) = \left.\frac{\partial^2 \log p_n(\theta \mid \boldsymbol{x}_n)}{\partial \theta^2}\right|_{\theta=\vartheta}. \tag{5}$$

Then $\phi_n = \Sigma_n^{-1/2}(\theta_n - \tilde{\theta}_n)$ converges in distribution to a standard (multivariate) normal random variable, if 3 technical conditions hold: (c1) 'Steepness': $\lim_{n\to\infty} \bar{\sigma}_n^2 \to 0$, where $\bar{\sigma}_n^2$ is the largest eigenvalue of Σ_n, (c2) 'Smoothness': Let $B_\delta(\tilde{\theta}_n) = \{\vartheta : \left|\vartheta - \tilde{\theta}_n\right| < \delta\}$. For any $\epsilon > 0$, there exists N and $\delta > 0$ such that, for any $n > N$ and $\vartheta \in B_\delta(\tilde{\theta}_n)$, the derivatives in Equation (5) exist and satisfy $I - A(\epsilon) \leq L_n''(\vartheta)\left\{L_n''(\tilde{\theta}_n)\right\}^{-1} \leq I + A(\epsilon)$, where I is a $d \times d$ identity matrix and $A(\epsilon)$ is a $d \times d$ symmetric positive semidefinite matrix whose largest eigenvalue tends to 0 as $\epsilon \to 0$, (c3) 'Concentration': For any δ, $\int_{B_\delta(\tilde{\theta}_n)} p_n(\theta) d\theta \to 1$ as $n \to \infty$. (Bernardo and Smith, 1994, Prop 5.14).

Theorem 2 asserts that uncertainty about the value of the unknown parameter value can be approximated asymptotically with a normal distribution. The Bayesian observed information Σ_n^{-1} is a measure of precision of the posterior distribution of θ, and behaves asymptotically like the frequentist observed information (which ignores the prior distribution) under rather general conditions, but the interpretation differs somewhat. The classical analog of Theorem 2 asserts that the MLE is asymptotically normally distributed about a 'true' parameter θ_0 (Law and Kelton, 2000), rather than describing uncertainty about θ. The mode $\tilde{\theta}_n$ is often called a MAP (maximum *a posteriori* probability) estimator. Conditions (c1-c2) basically insure that the the posterior mode is asymptotically shaped like a normal distribution, and (c3) insures that probability outside a neighborhood of $\tilde{\theta}_n$ is negligible. Bernardo and Smith (1994) also discuss alternate conditions.

The above results apply to conceptually infinite exchangeable sequences that can be used to infer statistical parameters, $\boldsymbol{\theta}_p$. Environmental parameters $\boldsymbol{\theta}_e$ do not have such sequences to help inference, but the subjective probability methods of de Finetti (1990); Savage (1972); Kahneman et al. (1982) still apply for assessing prior distributions. Exchangeability is relevant if a finite exchangeable sequence exists to help inference for $\boldsymbol{\theta}_e$.

Here are facts that link finite and infinite exchangeable sequences that are not used further in this paper, but are useful for further subjective probabil-

ity work. One, exchangeability is weaker than even conditional independence for finite sequences. For example, let $\Omega = \{0,1\}^N$ model $N = 3$ Bernoulli outcomes, and let $\theta_N = \sum_{i=1} X_i/N$. The subjective probability assessment $p((1,0,0)) = p((0,1,0)) = p((0,0,1)) = 1/3$ is an exchangeable Bernoulli model, but not independent, because $X_1 + X_2 + X_3 = 1$. Similarly, X_1 and X_2 are not conditionally independent, given $\theta_N = 1/3$ (This model is an atypical subjective assessment for coin flips, but matches well the 'hide a coin under a shell' game.) Two, suppose that each of the finite set of alternatives $\theta_N \in \{0/N, 1/N, \ldots, (N-1)/N, 1\}$ is judged equally likely for each N, then $\lim_{N\to\infty} p(\theta_N)$ converges in distribution to Laplace's (1812) continuous $\texttt{uniform}[0,1]$ prior distribution for $\theta = \lim_{N\to\infty} \theta_N$. Three, de Finetti (1990) derives Equation (2) as a consequence of having a conceptually infinite exchangeable sequence of Bernoulli outcomes, as opposed to directly assuming conditional independence. Four, judgments stronger than exchangeability, such as invariance to sums or to an ℓ_p-norm, may be required to justify de Finetti-type representations for other random variables (Barlow and Mendel, 1992; Chick and Mendel, 1998).

1.2 Loss and Value of Information

The fact that input uncertainty is described by probability distributions allows the modeler to (1) assess the **expected value of information** (EVI) of additional data collection, and (2) to perform an uncertainty analysis. The EVI is useful in experimental design. It measures the value of resolving uncertainty with respect to a **loss function** $\mathcal{L}(d,\omega)$ that describes the loss when a **decision** d is chosen and the state of nature is ω. Experiments can bring information about ω, so the expected improvement in the loss given by the experiment is a Bayesian experimental design criterion.

The value of information idea directly leads to the selection procedures in Section 5. A simplified version of that problem adapted from de Groot (1970, Sec. 11.8-9) illustrates the key concepts. Suppose we must decide whether or not the unknown mean W of a normal distribution is smaller (decision $d = 1$) or larger ($d = 2$) than w_0. Assume the variance σ^2 is known. Conditionally independent samples $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$, with $p(X_i) \sim \texttt{Normal}\,(w, \sigma^2)$ given $W = w$, can be used to infer the value of the mean. The decision maker designs a sampling experiment (chooses n) to balance the cost of sampling, cn, and the expected penalty if the wrong answer is chosen. Here the penalty for incorrect selection is the opportunity cost $\mathcal{L}\,(d,w)$, the difference between the actual value of w and w_0 when the wrong answer is selected, and 0 if the right answer is selected. Hence,

$$\mathcal{L}(1,w) = \begin{cases} 0 & \text{if } w \le w_0 \\ w - w_0 & \text{if } w > w_0, \end{cases} \quad \text{and } \mathcal{L}(2,w) = \begin{cases} w_0 - w & \text{if } w \le w_0 \\ 0 & \text{if } w > w_0. \end{cases}$$

Since the mean is not known exactly, there is a potential penalty for incorrectly specifying whether W is smaller or larger than w_0. We model uncertainty about W with a $\texttt{Normal}(\mu, 1/\tau)$ prior distribution, which is conjugate for normal sampling with an unknown mean and known variance (de Groot, 1970). Here τ is the *precision* in our uncertainty about W. Observing $\boldsymbol{X}_n = \boldsymbol{x}_n$ would reduce the uncertainty and result in the posterior distribution

$$p(w \mid \boldsymbol{x}_n) \sim \texttt{Normal}\left(z, \tau_n^{-1}\right),$$

where

$$\begin{aligned} z &= \text{posterior mean of } W = E[W \mid \boldsymbol{x}_n] = \frac{\tau\mu + n\bar{x}_n/\sigma^2}{\tau + n/\sigma^2} \\ \tau_n &= \text{posterior precision of } W = \tau + n/\sigma^2. \end{aligned}$$

The variance τ_n^{-1} equals the posterior variance approximation Σ_n in Equation (5) because Σ_n is based on a normal distribution approximation.

The posterior mean z influences the decision, but it depends upon n, which must be selected before $\boldsymbol{X}_n$ is observed. We therefore need the **predictive distribution** $p(z)$ of the posterior mean $Z = E[W \mid \boldsymbol{X}_n] = (\tau\mu + n\bar{X}_n/\sigma^2)/\tau_n$ to see how n samples influences the decision d. The conditional distribution of $\bar{X}_n$ given w is $\texttt{Normal}(w, \sigma^2/n)$. Mixing over the prior distribution of W implies that the predictive distribution for Z is $\texttt{Normal}(\mu, \tau_z^{-1})$, where

$$\tau_z = \tau(\tau + n/\sigma^2)/(n/\sigma^2). \tag{6}$$

The variance τ_z^{-1} of Z is 0 as $n \to 0$ (no new information). If $n \to \infty$ (perfect information about w), then $\mathrm{Var}[Z] \to \tau^{-1}$, the prior variance of W.

The experimental design that minimizes risk (the cost of sampling plus expected losses due to a potentially incorrect decision) is the n that minimizes a nested expectation, the inner expectation corresponding to the expected loss after $\boldsymbol{X}_n = \boldsymbol{x}_n$ is observed, the outer expectation averaging over $\boldsymbol{X}_n$.

$$\rho(n) = cn + E[E[\mathcal{L}(d(\boldsymbol{X}_n), W) \mid \boldsymbol{X}_n]] \tag{7}$$

A general technique for determining $E[\mathcal{L}(d(\boldsymbol{X}_n), W) \mid \boldsymbol{X}_n]$ is to obtain an auxiliary loss function $\mathcal{L}^*$ that has the same optimal decision, but simplifies

the loss function by making the loss of one of the decisions equal to 0. Adding a function of w does not change the optimal decision (de Groot, 1970). Set $\mathcal{L}^*(d,w) = \mathcal{L}(d,w) - \mathcal{L}(1,w)$, which is 0 if $d=1$ and is $w_0 - w$ if $d=2$. Then

$$E[\mathcal{L}^*(d(\boldsymbol{X}_n), W) \mid \boldsymbol{X}_n] = \begin{cases} 0 & \text{if } d(\boldsymbol{X}_n) = 1 \\ w_0 - Z & \text{if } d(\boldsymbol{X}_n) = 2. \end{cases} \tag{8}$$

The decision that minimizes the loss in Equation (8) is to assert $d(\boldsymbol{X}_n) = 2$ ('bigger') if the posterior mean exceeds the threshold, $Z > w_0$, and to assert $d(\boldsymbol{X}_n) = 1$ ('smaller') if $Z \leq w_0$.

The expectation over the outcomes of this experiment can be determined with well-known tables because the decision depends upon $\boldsymbol{X}_n$ only through Z, and Z has a normal distribution. Define $\phi(\cdot)$ and $\Phi(\cdot)$ to be the pdf and cdf of a standard normal random variable, respectively. The expected loss can be determined from the standard normal loss function $\Psi[s] = \int_s^\infty (t-s)\phi(t)dt = \phi(s) - s(1-\Phi(s))$ for expected lost sales in the newsvendor problem if demand is normally distributed (e.g. Nahmias, 2000, standardized loss p. 262).

$$\begin{aligned} E[E[\mathcal{L}^*(d(\boldsymbol{X}_n), W) \mid \boldsymbol{X}_n]] &= -\int_{w_0}^{\infty} (z - w_0) p(z \mid \boldsymbol{X}_n) dz \\ &= -\tau_z^{\frac{-1}{2}} \Psi[\tau_z^{\frac{1}{2}}(w_0 - \mu)] \end{aligned}$$

The expected loss of the original loss function is recovered by adding back $E[\mathcal{L}(1,W)]$, using the prior distribution of W for the expectation.

$$E[E[\mathcal{L}(d(\boldsymbol{X}_n), W) \mid \boldsymbol{X}_n]] = \tau^{\frac{-1}{2}} \Psi[\tau^{\frac{1}{2}}(w_0 - \mu)] - \tau_z^{\frac{-1}{2}} \Psi[\tau_z^{\frac{1}{2}}(w_0 - \mu)] \tag{9}$$

The EVI for m samples is the difference between Equation (9) when $n \to 0$ and when $n = m$ (τ_z depends on n). If $w_0 > \mu$, the EVI simplifies to $\tau_z^{\frac{-1}{2}} \Psi[\tau_z^{\frac{1}{2}}(w_0 - \mu)]$. Combine Equation (9) with Equations (6-7), note that $d\Psi/ds = \Phi(s) - 1$ and $d\tau_z/dn = -\tau^2\sigma^2/n^2$, and take the derivative with respect to n (relaxing the integer assumption) to obtain an optimality condition for the sample size.

$$\frac{\partial \rho}{\partial n} = \frac{1}{2}\tau_z^{-\frac{3}{2}} \phi[\tau_z^{\frac{1}{2}}(w_0 - \mu)] \cdot \frac{-\tau^2\sigma^2}{n^2} + c = 0$$

For diminishing costs $c \to 0$, the sample size is large. Since $\tau_z \to \tau$ as $n \to \infty$, the optimal sample size n is approximately

$$n^* = \left(\frac{\tau^{\frac{1}{2}} \sigma^2 \phi[\tau^{\frac{1}{2}}(w_0 - \mu)]}{2c} \right)^{\frac{1}{2}}. \tag{10}$$

This argument illustrates the basic ideas of loss functions, and the use of predictive distributions for future samples to infer the EVI of sampling. *The technique of adding functions of the unknowns can be useful to simplify the derivation of the optimal solution. Asymptotic approximations are a further tool to identify criteria-based sampling plans.* Extensions of this basic argument justify the value of information based selection procedures of Section 5.1 and Chick and Inoue (2001a,b, 2002).

An alternate mechanism to approximate the effect of information on parameter uncertainty is based on a thought experiment for the posterior probabilities of parameters. For members of the regular exponential family, the asymptotic variance approximation Σ_n in Equation (5) simplifies to the form $H^{-1}(\theta)/(n_0 + n)$, where H is the expected information from one observation (Equation (4)), when a canonical conjugate prior distribution is used (Bernardo and Smith, 1994). To approximate the effect of collecting m additional samples on the parameter uncertainty, one could presume that the posterior distribution changes from $\texttt{Normal}\left(\tilde{\theta}_n, \Sigma_n\right)$ to

$$\texttt{Normal}\left(\tilde{\theta}_n, \Sigma_n(n_0 + n)/(n_0 + n + m)\right). \tag{11}$$

This transformation reflects an appropriate scaling of the posterior precision, and the idea is used in a frequentist context for estimating how many replications are required to achieve a confidence interval of a given size (Law and Kelton, 2000). Chen (1996) uses this type of approximation for the Bayesian posterior distribution of the unknown means of several simulated systems in order to motivate a class of ranking and selection procedures called the $OCBA$. Ng and Chick (2001, 2004) use the idea to plan experiments to reduce input uncertainty in a way that reduces output uncertainty.

1.3 Uncertainty Analysis

The fact that uncertainty about inputs is described with probability distributions allows a modeler to engage in **uncertainty analysis**, in addition to sensitivity analysis. A sensitivity analysis tests how the mean simulation output depends upon one or more input parameters as that parameter is varied (estimating $E[g(\theta) \mid \mathcal{E}]$ as a function of θ, given all information $\mathcal{E}$). Uncertainty analysis entails propagating input parameter uncertainty about Θ through to uncertainty about outputs Y. Even if a simulation has no random number stream, a distribution on unknown inputs means that the output is random.

An unbiased estimator of the mean output $E[Y \mid \mathcal{E}]$ with both stochastic (from $\boldsymbol{u}$) and systemic (or parameter) uncertainty accounted for is obtained from the Bayesian model average (BMA) in Figure 2, which averages over random inputs sampling according to the distribution $p(\theta \mid \mathcal{E})$ (Draper, 1995; Chick,

2001). Zouaoui and Wilson (2003) explore the relative magnitude of stochastic and systemic uncertainty with variations on the BMA, and discuss how to update the estimate should new data become available (so the algorithm need not be rerun from scratch). Importance sampling (cf. Chapter 11) techniques can reweight estimates accordingly (with likelihood ratio determined as the ratio of the 'new' posterior divided by the 'old' distribution). Andradóttir and Glynn (2004) examine the estimation of $E[Y \mid \mathcal{E}]$ when there may be bias in the estimates of Y given θ, when quasi-random sequences are used in place of the pseudo-random sequences assumed by Figure 2, or when numerical techniques like Simpson's rule are employed to select values of θ. Another goal is to estimate the distribution of the conditional expectation $E[Y \mid \Theta, \mathcal{E}]$. When Y is a deterministic function of Θ, then naive Monte Carlo simulation can be used with traditional kernel estimation techniques to assess the distribution of $Y(\Theta)$. When the simulation is stochastic (depends on $\boldsymbol{u}$), then $E[Y \mid \theta, \mathcal{E}]$ is imperfectly estimated for any given θ. Given several technical conditions (e.g. univariate continuous-valued θ, monotonic mean response), Steckley and Henderson (2003) derive asymptotically optimal ways of selecting by cleverly selecting r and n in Figure 2 to produce a kernel density estimator based on the output. Their work builds upon Lee and Glynn (1999), which estimated the distribution function of $E[Y \mid \Theta, \mathcal{E}]$ for discrete θ.

2 Computational issues

Three basic computational issues for implementing a Bayesian analysis are maximization (e.g. find the MLE $\hat{\theta}$, or MAP $\tilde{\theta}$ estimators for a posterior distribution); integration, either to find a marginal distribution (e.g. find $p(\theta_1 \mid \boldsymbol{x}_n)$ from $p(\theta_1, \theta_2 \mid \boldsymbol{x}_n)$) or constant of proportionality for a posterior distribution (e.g. find $c^{-1} = \int f(\boldsymbol{x}_n \mid \theta) d\pi(\theta)$); and simulation (e.g. sample from $p(\theta \mid \boldsymbol{x}_n)$ in order to estimate $E[g(\theta) \mid \boldsymbol{x}_n]$). Techniques to address these issues are described in a variety of sources (e.g. Naylor and Smith, 1982; Evans and Swartz, 1995; Tanner, 1996; Gilks et al., 1996; The Mathworks, 2002).

---

for $r = 1, \ldots, r$ replications
  sample parameter θ_r from $p(\theta \mid \mathcal{E})$
  for $i = 1, 2, \ldots, n$
    generate simulation output y_{ri} given input θ_r
  end loop
end loop
Estimate $E[Y \mid \mathcal{E}]$ with $\bar{y} = \sum_{r=1}^{R} \frac{1}{R} \sum_{i=1}^{n} y_{ri}/n$.

Fig. 2. Bayesian Model Average (BMA)

For maximization, a number of methods are available including gradient-based methods (e.g. Newton-Raphson), gradient-free methods (e.g. Nelder-Mead), and simulation-based methods. The expectation-maximization (EM) algorithm is a technique for finding the MAP or MLE when there is missing data or nuisance parameters are to be integrated out (e.g. the MAP of $p(\theta_1 \mid \boldsymbol{x}_n)$ when it is 'messy' but $p(\theta_1, \theta_2 \mid \boldsymbol{x}_n)$ is easier to manipulate).

For integration, five general techniques apply (Evans and Swartz, 1995) when analytical results (e.g. conjugate priors) are not available: quadrature, asymptotic methods, Markov chain methods, importance sampling, adaptive importance sampling. Quadrature is useful when the number of dimensions is not too large. The Laplace method is an interesting asymptotic approximation for integrals $\int g(\theta) f(\theta \mid \boldsymbol{x}_n) d\pi(\theta)$. The Laplace method applies even if $f(\theta \mid \boldsymbol{x}_n)$ is only a likelihood when the constant of proportionality for the posterior is unknown, and can work well for integrating out nuisance parameters if regularity conditions hold. The method is based on asymptotic normality approximations like those used for Equation (5), and therefore require a large n. Another effective technique for approximating the density $p(\theta_1 \mid \boldsymbol{x}_n)$ (not just the MLE or MAP) when it is 'messy' but $p(\theta_1 \mid \theta_2, \boldsymbol{x}_n)$ and $p(\theta_2 \mid \theta_1, \boldsymbol{x}_n)$ are easy to manipulate is data augmentation, often called the IP algorithm (for imputation, posterior algorithm). The IP algorithm is a nice alternative to other kernel estimation methods, and is closely related to the Markov Chain Monto Carlo (MCMC) methods mentioned below. *Importance sampling (IS) remains one of the more powerful methods for efficient integration.* See Chapter 11.

For simulation of variates, classical methods for generating independent variates from posterior distributions may apply (see Chapter 4). *Posterior distributions are often known only up to a constant of proportionality (the numerator of Bayes' rule is easy to write, but the denominator may be hard to compute). It is therefore important to have a method to simulate variates for arbitrary functions proportional to posterior distributions. MCMC is the most important of those methods at present.* MCMC constructs a Markov chain whose stationary distribution is the desired posterior distribution (Chapter 4, this volume; Gilks et al. (1996)). The ARMS (adaptive rejection Metropolis sampler) combines adaptive rejection sampling, which is useful for logconcave posterior distributions, together with an MCMC-type Metropolis step to handle non-logconcave distributions (Gilks et al., 1995). States of a chain constructed with MCMC techniques can be sampled for input into the BMA of Figure 2. Samples and estimators based on MCMC need evaluation to assure reasonable convergence for estimators and faithfulness to the posterior distribution.

Figure 3 illustrates a qualitative feel for some of the approximation techniques for a gene linkage model (Tanner, 1996) that has a parameter $\theta \in [0, 1]$. A spreadsheet implementation of the EM algorithm identified the MAP $\tilde{\theta}$. The

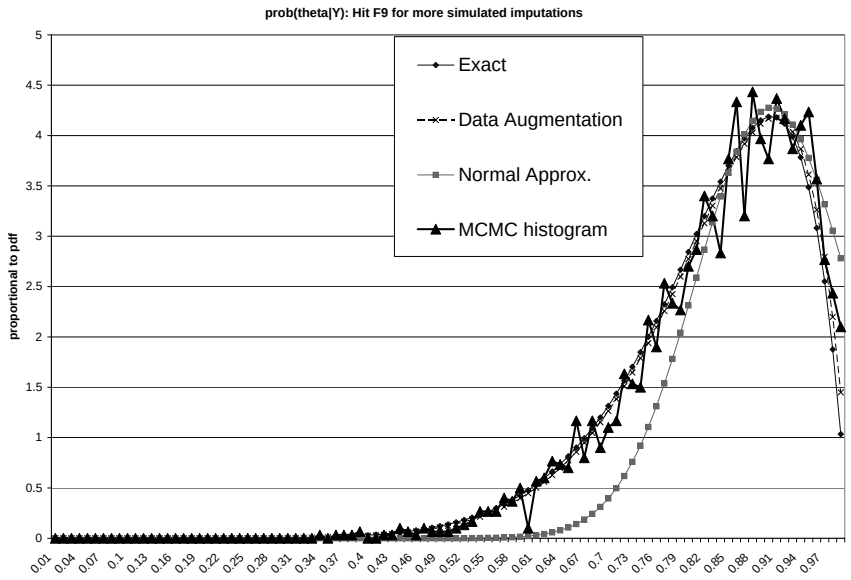

Fig. 3. Different approximations for the posterior pdf

asymptotic normal approximation of Theorem 2 provides a reasonable estimate of the mode and variance of the true posterior distribution, but does not model skewness well, particularly if $\tilde{\theta}$ is near the boundary or if few data points are available. Data augmentation results in a smoother kernel estimator than the empirical histogram estimator usually studied with MCMC methods. The MCMC estimator could be smoothed like the data augmentation to provide a much more accurate representation (Gilks et al., 1996) with an additional computational cost for smoothing.

For simple examples a spreadsheet is fine, but more powerful tools are needed in general to implement Bayesian inference. The `BUGS` and `WinBUGS` packages implement Gibbs sampling and some Metropolis sampling, and are available on the WWW (Spiegelhalter et al., 1996). BOA, for Bayesian output analysis (Smith, 2004), is a set of MCMC diagnostic tools for convergence and data analysis that functions with the `R` or `S-PLUS` statistical packages. Gauss and Matlab are also commonly used to program MCMC methods.

At present, it is possible to input randomized input parameters to some commercial discrete-event simulation packages to implement the BMA algorithm of Figure 2, but interfaces are not yet fully user friendly. A user-friendly tool to implement the BMA and other uncertainty analysis needs in commercial discrete-event simulation packages would be helpful. Uncertainty analysis for other Monte Carlo applications has been available as a spreadsheet tool for some time (e.g. Winston, 2000).

3 Input distribution and model selection

Selecting an input distribution to model a sequence of random quantities $X_1, X_2, \ldots$ is often more complicated than inferring a parameter of a single parametric distribution, as described in Section 1. There is often a finite number q of candidate distributions proposed to model a given source of randomness, with continuous parameters $\theta_m = (\vartheta_{m1}, \ldots, \vartheta_{md_m})$, where d_m is the dimension of θ_m, for $m = 1, \ldots, q$. For example, service times might be modeled by exponential, lognormal or gamma distributions ($q = 3$). Denote by $p(x \mid m, \theta_m)$ the probability density function (pdf) for X, given m and θ_m.

The classical approach in simulation (Law and Kelton, 2000) for input selection is to (a) find the MLE of each parameter, (b) perform a goodness-of-fit test, and (c) pick the 'best' fitting candidate distribution and input the MLE into the simulation. The Bayesian approach addresses some controversial aspects of the classical approach. Critiques of classical techniques in general include: goodness-of-fit and P-value criteria are difficult to interpret and inconclusive at best, and misleading at worst (Berger and Pericchi, 1996); use of a single distribution and parameter underestimates the uncertainty in the distribution's functional form and parameter (Draper, 1995); with few data points, few candidate distributions are rejected, and with many data points, all distributions are rejected (Raftery, 1995); there is no coherent method for selecting among nonrejected distributions; and classical techniques can reject the *a posteriori* most probable distribution (Lindley, 1957; Berger and Delampady, 1987). *In the simulation context, input uncertainty can make standard confidence intervals for the mean output almost meaningless if the classical approach is employed* (Chick, 2001; Barton and Schruben, 2001).

A Bayesian approach is no different than the approach in Section 1, except that a prior probability distribution needs to be placed on the model/parameter combination, $\pi(M = m, \theta_m)$, a mixed discrete-continuous model. If data $\boldsymbol{x}_n$ becomes available, the BMA then requires sampling from the joint posterior $p(m, \theta_m \mid \boldsymbol{x}_n)$. This can be accomplished by composition, sampling the input model then the parameter with $p(m \mid \boldsymbol{x}_n)p(\theta_m \mid m, \boldsymbol{x}_n)$. *While improper, noninformative prior distributions can formally be used when there is a single candidate model and enough data makes the posterior proper, this cannot be done when there are multiple candidate distributions.* The reason is that

$$p(M = m \mid \boldsymbol{x}_n) = \frac{\int p(\boldsymbol{x}_n \mid \theta_m, m)\pi(\theta_m \mid m)\pi(M = m)d\theta_m}{\sum_{i=1}^{q} \int p(\boldsymbol{x}_n \mid \theta_i, i)\pi(\theta_i \mid i)\pi(M = i)d\theta_i}$$

can be changed at will by changing an improper prior $\pi(\theta_m \mid M)$ to an equally valid improper prior $c\pi(\theta_m \mid M)$ for an arbitrary constant c. Proper (and therefore informative) prior distributions $\pi(\theta_m \mid M = m)$ are needed for each m in order to make Bayes' rule well-defined. Chick (2001) implements the

BMA with this model selection for discrete-event simulations, and suggests a method for assessing $\pi(\theta_m \mid M = m)$ based on moment methods. O'Hagan (1995); Berger and Pericchi (1996) proposed automated prior distribution selection techniques for the model selection problem that use variations on the theme of using part of the data with a noninformative prior.

Chick (2001) illustrated Bayesian input modeling in a stochastic simulation context when $q > 1$, and suggested a method-of-moments approach for assessing prior distributions for the unknown parameters of each candidate distribution. Zouaoui and Wilson (2004) noted a decoupling of stochastic uncertainty from two types of structural uncertainty (about candidate models and their parameters) under special conditions, and provided a variance reduction for the BMA and numerical analysis. Richardson and Green (1997) and Cheng and Currie (2003) present techniques for the nonregular case when a candidate itself is a mixture distribution (e.g., a mixture of 2 or 3 normal distributions).

Selecting models according to $p(M|\mathcal{E})$ is consistent in that if one of the entertained models is actually the true model, then the true model is selected if enough data is observed and some regularity conditions hold. When the true model is not among those being considered, Bayesian model selection chooses the model that is closest to the true model in terms of Kullback-Leibler divergence (Berk, 1966; Bernardo and Smith, 1994; Dmochowski, 1999).

4 Joint input-output models

Simulation is interested in both stochastic uncertainty, or randomness that occurs when all model parameters are known, and structural uncertainty, or uncertainty about model inputs when a real system is being simulated. This section describes an input-output model that quantifies the uncertainty in simulation outputs due to input uncertainty, data collection plans for reducing input uncertainty in a way that effectively reduces output uncertainty, mechanisms to select computer inputs to improve estimates of the system response, and mechanisms to help infer probability distributions for input parameters, given information about output parameters (the inverse problem).

Recall Figure 1. It is impossible to simulate all possible values of continuous input parameters in finite time on a finite number of computers. Metamodels describe the value of g at untested input values (see Chapter 18). This is useful when the simulation model requires extensive computation. Metamodel parameters $\boldsymbol{\psi} = (\phi_1, \phi_2, \ldots, \phi_{n_m})$ may include regression coefficients, or parameters of a Gaussian random function (GRF) model of the mean response (Cressie, 1993; Santner et al., 2003; van Beers and Kleijnen, 2003). Since the metamodel parameters are unknown, they are described as a random variable

$\boldsymbol{\Psi}$. The metamodel is

$$Y = g(\boldsymbol{\Theta}_p, \boldsymbol{\Theta}_e, \boldsymbol{\theta}_c; \boldsymbol{U}, \boldsymbol{\Psi}) \tag{12}$$

This formulation allows for Y to be predicted for unsimulated values of $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_e, \boldsymbol{\theta}_c)$ via the response model and $\boldsymbol{\Psi}$. Field data from a modeled system can be used (1) to infer the input parameters $\boldsymbol{\theta}_c$, along the lines of Sections 1 and 3, or (2) to understand the distribution of the outputs Y. Output from multiple simulation runs are used to infer $\boldsymbol{\Psi}$. This formulation generalizes several previous models that focused on $g(\boldsymbol{\theta}_p, \boldsymbol{u})$ (Cheng and Holland, 1997), or nonstochastic models that do not model randomness from $\boldsymbol{u}$ (Santner et al., 2003). The model of Chick (1997) differed in that simulation output was not considered to be exchangeable with a real system's output (a calibration issue).

4.1 Bayesian Metamodeling

Here we discuss inference for the normal linear model and Gaussian random function (GRF) metamodels. The normal linear model is

$$Y = \sum_{\ell=1}^{p} g_\ell(\boldsymbol{\theta})\beta_\ell + Z(\boldsymbol{\theta}; \boldsymbol{U}) = \boldsymbol{g}^T(\boldsymbol{\theta})\boldsymbol{\beta} + Z(\boldsymbol{\theta}; \boldsymbol{U}),$$

for *known* regression functions $g_1, \ldots, g_p$, unknown regression coefficients $\boldsymbol{\beta}$, and independent zero-mean random noise $Z(\cdot)$ with sampling variance σ^2. If σ^2 does not depend upon $\boldsymbol{\theta}$, then the conjugate prior is an inverse gamma distribution for σ^2 and a conditionally normal distribution for $\boldsymbol{\beta}$ given σ^2, if all factors are active (Bernardo and Smith, 1994). George and McCulloch (1996) and Cheng (1999) discuss techniques for estimating which factors are active with what probability. Ng and Chick (2002) describe an entropy-based experimental design criterion to identify which factors are active and reduce parameter uncertainty simultaneously.

If the g_i represent the individual dimensions of the unknown parameters $(\boldsymbol{\theta}_p, \boldsymbol{\theta}_e)$, the β_ℓ are gradients with respect to the inputs. If further the model has only statistical input parameters $\boldsymbol{\theta}_p$ for which data can be collected (but not parameters $\boldsymbol{\theta}_e$ for which no data is available), Ng and Chick (2001) and Zouaoui and Wilson (2003) indicated that output uncertainty can be decoupled asymptotically or under special conditions.

$$\begin{aligned} V_{tot} = \mathrm{Var}[\bar{Y} \mid \mathcal{E}] &\approx \text{stochastic} + \text{structural uncertainty} \\ &\approx \frac{\hat{\sigma}_0^2}{m} + \frac{\boldsymbol{\beta}^T \hat{H}_{\boldsymbol{\theta}_p}^{-1} \boldsymbol{\beta}}{n}, \end{aligned}$$

where $\hat{\sigma}_0^2$ is the estimate of the variance from m replications, the MLE $\hat{\boldsymbol{\theta}}_p$

and estimate $\hat{H}^{-1}_{\boldsymbol{\theta}_p}$ of the information in one observation are based on n data points, and technical conditions hold (e.g. those for Theorem 2). This adapted a frequentist result of Cheng and Holland (1997). Ng and Chick (2001) applied the result to uncertainty due to multiple input parameters, to provide sampling plans to collect further data (e.g. for arrival rates or for service time distributions) to reduce input parameter uncertainty in a way that optimally reduces output uncertainty in some sense. Ng and Chick (2004) extended that analysis by accounting for uncertainty in $\boldsymbol{\beta}$ as well, providing analytical results for members of the exponential family of distributions, and giving a numerical analysis.

Figure 4 summarizes qualitatively how ignoring input parameter uncertainty can significantly degrade confidence interval (CI) coverage. The experiment involved simulating the parameter estimation process for several simulation input parameters, inputting those parameters into a simulation, then generating a nominal 95% CI (see (Ng and Chick, 2004) for details). The values at $B = 0$ in all graphs represent the CI coverage and mean half width if no additional data is collected after that first round of simulation. The top row of graphs gives the empirical coverage and half width if the CI is based on the estimated variance V_{tot} involving both stochastic and structural uncertainty. The bottom row is based on computing the CI by inputting only the MLE of the input parameters into the simulation, and using only stochastic uncertainty $V_{stoch} = \hat{\sigma}_0^2/m$ to estimate the variance of the estimator. The values at $B > 0$ describe how the coverage would change if additional data for the several different input parameters were collected in an optimal way. Optimal here is defined by collecting data for the different parameters in a way that minimizes V_{tot} if the effect of additional samples is presumed to reduce V_{tot} as in Equation (11). There is a slight degradation in coverage, perhaps due to the fact that the simulation model was nonlinear and the approximation for V_{tot} is based on a local linear approximation. The bigger story from this example is that ignoring input parameter uncertainty can lead to almost meaningless CI statements if the mean of the simulated response is a nonlinear function of the input parameters. The effect is worse with fewer data points to estimate parameters. To date, much simulation research seeks analytical results for stochastic models, or mechanisms to reduce the variance of estimators due to stochastic noise. Those results need to be complemented with an *understanding of how performance depends on input uncertainty, and methods to reduce input uncertainty to effectively reduce output uncertainty.* Bayesian tools can help.

GRFs are well-known response models in deterministic simulations, particularly in geostatistics (Cressie, 1993; Santner et al., 2003), but are less well known in stochastic simulation. GRFs provide flexibility that the linear model does not, and are useful when g takes a long time to compute. The GRF for

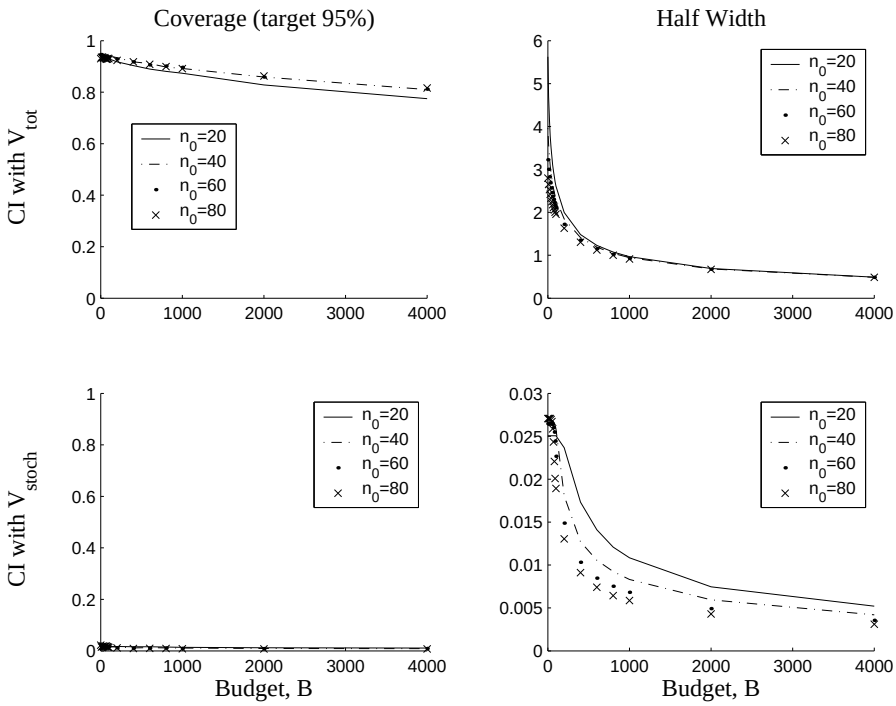

Fig. 4. Empirical coverage for a nominal 95% CI is poor if parameter uncertainty is ignored

an unknown nonstochastic g (no random number stream $\boldsymbol{u}$) is

$$Y(\boldsymbol{\theta}) = \sum_{\ell=1}^{p} g_\ell(\boldsymbol{\theta})\beta_\ell + Z(\boldsymbol{\theta}) = \boldsymbol{g}^T(\boldsymbol{\theta})\boldsymbol{\beta} + Z(\boldsymbol{\theta}) \tag{13}$$

for known regression functions $g_1, \ldots, g_p$ of $\mathbb{R}^d$, and unknown regression coefficients $\boldsymbol{\beta}$. The zero-mean random second-order process $Z(\boldsymbol{\theta})$ is such that for any distinct inputs $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$, the vector $(Y_1, \ldots, Y_m)$ has a multivariate normal distribution, conditional on $\boldsymbol{\beta}$. GRFs are determined by their mean $\boldsymbol{g}^T(\boldsymbol{\theta})\boldsymbol{\beta}$ and (auto)covariance function $C^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathrm{Cov}(Y(\boldsymbol{\theta}_1), Y(\boldsymbol{\theta}_2))$, defined independent of $\boldsymbol{\beta}$. It is common to assume strong stationarity ($(Y_1, \ldots, Y_m)$ and $(Y_1 + \boldsymbol{h}, \ldots, Y_m + \boldsymbol{h})$ have the same distribution), so $C^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = C(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$.

Inference for $g(\boldsymbol{\theta})$ at values of $\boldsymbol{\theta}_{r+1}$ not yet input to a simulation model is more flexible than linear regression via the correlation function $R(\boldsymbol{h}) = C(\boldsymbol{h})/C(0)$ for $\boldsymbol{h} \in \mathbb{R}^d$. See Santner et al. (2003) for examples. Kriging, a geostatistics term, is a best linear unbiased prediction (BLUP) for $g(\boldsymbol{\theta}_{r+1})$. An assessment of the uncertainty in $g(\boldsymbol{\theta}_{r+1})$ can be used as an experimental design technique to choose inputs to reduce response uncertainty (Santner et al., 2003). Example usage of GRFs includes input selection for efficient response surface estimation; percentile estimation; and model calibration (Sacks et al., 1989; O'Hagan et al., 1999; Kennedy and O'Hagan, 2001; Santner et al., 2003; Oakley, 2004). Stochastic observations can be modeled by assuming a measurement error, giving a so-called 'nugget effect' (Santner et al., 2003). Van

Beers and Kleijnen (2003) found that a special case of Equation (13) was a useful metamodel of a stochastic process simulation. *GRFs provide an effective mechanism for reducing the computational effort to get a good response estimate by selecting simulation inputs on areas where the mean response has the greatest uncertainty. More work is needed for GRFs in the stochastic simulation context.*

4.2 Inference of Input Parameters from Output Information

Uncertainty analysis examines the distribution of $g(\boldsymbol{\theta}_c, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_e; \boldsymbol{U})$ induced by the distribution of $(\boldsymbol{\Theta}_p, \boldsymbol{\Theta}_e)$ (whether g depends upon $\boldsymbol{U}$ or is deterministic). The question of calibration is whether the realization or distribution of $(\boldsymbol{\Theta}_p, \boldsymbol{\Theta}_e)$ can be inferred from data or probability assessments about Y.

For example, the arrival rate λ and service rate μ of an M/M/c queue are inputs to a simulation model. If they are unknown, their distribution can be inferred from field data. The distribution of λ, μ induces a distribution on system outputs, such as the average queue content $\bar{Q}$, and the 'clock speed' as measured by the autocorrelation of the queue occupancy process. The inverse problem is whether inputs λ and μ can be inferred from observation on the outputs such as the queue occupancy at discrete times $Q(\tau_1), \ldots, Q(\tau_n)$. McGrath et al. (1987); McGrath and Singpurwalla (1987); Armero and Bayarri (1997) comment on the potential to infer input parameters for queues from outputs, but mostly evaluate the expected value of information for inferring λ, μ from interarrival and service time data, along with potentially one observation $Q(\tau_1)$. The problem is hard because the likelihood function involves transient state transition probabilities, which are complicated for many stochastic processes. Larson (1990) describes frequentist inference with incomplete arrival time data, but using transactional data on arrivals and departures.

For epidemic models, O'Neill (2002) provides a review of tools and techniques. Chick et al. (2003) illustrate the inference of epidemic model (input) parameters from (output) data on disease incidence and germ contamination by estimating the likelihood model for outputs, given inputs, using approximations of the stationary distribution to help assess the likelihood function. Kennedy and O'Hagan (2001) provide a GRF model to infer $(\boldsymbol{\theta}_p, \boldsymbol{\theta}_e)$ given observations of $Y(\boldsymbol{\theta}_c)$ to help calibrate input parameters given output observations, assuming a nonstochastic response function g and that output data can be observed with random error. Also see Craig et al. (2001). Part of the problem of inferring input parameters from outputs arises from allowing the computer output to be an imperfect model of reality (e.g. with bias modeled as a Gaussian random field). The joint input-output model of Chick (1997) did not allow for the inference of input parameters from system outputs because it considered

input models to be adequate for representing reality, but did not consider the outputs of the model to be exchangeable with the corresponding observations from the real system. The reason is that the lack of detail in computer models may not reflect the reality of the actual system. The joint input-output model augmented with an explicit bias metamodel helps quantify model validation concerns. The bias term accounts for the potential that model output might not be perfectly exchangeable with computer model output, conditional upon the values of the input parameters matching the values of the corresponding parameters in the real system.

Kraan and Bedford (2003) assess subjective prior distributions for inputs to a nonstochastic g that best induces a desired distribution for the outputs, in the sense of Kullback-Leibler divergence, even without data for Bayesian inference.

5 Ranking and selection

This section compares and contrasts frequentist and Bayesian approaches to ranking and selection. The objective of ranking and selection is to select the best of a finite number k of systems, where best is defined in terms of the expected performance of a system (e.g. Chapter 17). In the notation of Section 4, $\boldsymbol{\theta}_c$ assumes one of a discrete set of values indexed by $i \in \{1, 2, \ldots, k\}$, and the goal is to identify the system i that maximizes $w_i = E[g(\boldsymbol{\theta}_{ci}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_e; \boldsymbol{U})]$. The means are inferred from observing simulation replications

$$Y_{ir} = w_i(\boldsymbol{\theta}_c, \boldsymbol{\theta}_p, \boldsymbol{\theta}_e) + \sigma_i(\boldsymbol{\theta}_c, \boldsymbol{\theta}_p, \boldsymbol{\theta}_e) z(\boldsymbol{\theta}_c, \boldsymbol{\theta}_p, \boldsymbol{\theta}_e; \boldsymbol{U}_{ri}) \tag{14}$$

for $i = 1, 2, \ldots, k; r = 1, 2, \ldots$, where the function g is rewritten in terms of the mean w_i, standard deviation σ_i, and a zero-mean unit-variance noise z. The usual formulation of the problem does not explicitly specify a dependence upon $(\boldsymbol{\Theta}_p, \boldsymbol{\Theta}_e)$, so we leave those parameters implicit in this section, and the response from different systems is not related with a metamodel $\boldsymbol{\psi}$ (although continuous optimization problems may use $\boldsymbol{\psi}$, see Chapter 18).

Both two-stage and sequential procedures are possible. In a two-stage procedure, a first stage consisting of r_0 replications for each system are run, initial estimates of $(\mathbf{w}, \boldsymbol{\sigma})$ are obtained in order to determine how many more samples to collect for each system, and then a second stage collects the samples and selects a system based on all output. Sequential procedures repeatedly allocate samples in a series of consecutive stages.

Different approaches to ranking and selection use different criteria for allocating samples and for measuring the quality of a selection. Frequentist ap-

proaches, like indifference zone (IZ) methods, provide correct selection guarantees over repeated applications of a selection procedure (see Chapter 17). Bayesian approaches, like the value of information procedures (VIP, Chick and Inoue, 2001a), and optimal computing budget allocation (OCBA, Chen, 1996), use posterior distributions to quantify the evidence for correct selection. These approaches are described and distinguished below.

Typical assumptions common to all three approaches are that the means and variances of the output of each system may differ, and that the output is independent and normally distributed, *conditional* on w_i and σ_i^2, for $i = 1, \ldots, k$.

$$\{Y_{ij} : j = 1, 2, \ldots\} \overset{i.i.d.}{\sim} \texttt{Normal}\left(w_i, \sigma_i^2\right)$$

Although the normality assumption is not always applicable, it is often possible to batch a large number of outputs so that independence and normality is approximately satisfied (we return to correlated output in Section 5.3). It will be easier at times to refer to the precision $\lambda_i = 1/\sigma_i^2$ instead of the variance. Set $\mathbf{w} = (w_1, \ldots, w_k)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$. Let $w_{[1]} \leq w_{[2]} \leq \ldots \leq w_{[k]}$ be the ordered means. In practice, the ordering $[\cdot]$ is unknown, and the best system, system $[k]$, is to be identified with simulation. A problem instance ('configuration') is denoted by $\boldsymbol{\chi} = (\mathbf{w}, \boldsymbol{\sigma}^2)$. Let r_i be the number of simulation replications for system i run so far. Let $\bar{y}_i = \sum_{j=1}^{r_i} y_{ij}/r_i$ be the sample mean and $\hat{\sigma}_i^2 = \sum_{j=1}^{r_i}(y_{ij} - \bar{y}_i)^2/(r_i - 1)$ be the sample variance. Let $\bar{y}_{(1)} \leq \bar{y}_{(2)} \leq \ldots \leq \bar{y}_{(k)}$ be the ordering of the sample means based on all replications seen so far. Equality occurs with probability 0 in contexts of interest here. The quantities r_i, $\bar{y}_i, \hat{\sigma}_i^2$ and (i) may change as more replications are observed.

Because output is random, correct selection cannot be guaranteed with probability 1 with a finite number of replications. A correct selection occurs when the selected system, system $\mathfrak{D}$, is the best system, $[k]$. Selection is based on a procedure's estimates $\hat{w}_i$ of w_i, for $i = 1, \ldots, k$, after all replications are observed. Usually overall sample means are the estimates, $\hat{w}_i = \bar{y}_i$, and the system with the best sample mean is selected as best, $\mathfrak{D} = (k)$, although some procedures may vary due to screening (Goldsman et al., 2002) or weighted samples (Dudewicz and Dalal, 1975).

The commonality between the three approaches stops there, and the differences now begin. The IZ (Chapter 17) approach seeks to guarantee a bound for the evidence for correct selection, with respect to repeated applications of the procedure to a given problem instance, for all problem instances within a specified class. For normally distributed output, the most widely-used criterion is a lower bound P^* for the probability of correct selection (Bechhofer et al., 1995), subject to the indifference zone constraint that the best system be at least a pre-specified amount $\delta^* > 0$ better than the others. Formally, the probability of correct selection ($\mathrm{PCS}_{\mathrm{iz}}$) is the probability that the system

selected as best (system $\mathfrak{D}$) is the system with the highest mean (system $[k]$), conditional on the problem instance. The probability is with respect to the simulation output Y_{ij} from the procedure (the realizations y_{ij} determine $\mathfrak{D}$).

$$\mathrm{PCS}_{\mathrm{iz}}(\boldsymbol{\chi}) \stackrel{\mathrm{def}}{=} \Pr\left(w_{\mathfrak{D}} = w_{[k]} \,|\, \mathbf{w}, \boldsymbol{\sigma}^2\right) \tag{15}$$

The validity of an IZ procedure comes from proving statements like

$$\mathrm{PCS}_{\mathrm{iz}}(\boldsymbol{\chi}) \geq P^*, \text{for all } \boldsymbol{\chi} = (\mathbf{w}, \boldsymbol{\sigma}^2) \text{ such that } w_{[k]} \geq w_{[k-1]} + \delta^*. \tag{16}$$

See Bechhofer et al. (1995) for other examples of the IZ approach with other distributions and indifference-zone structures. Early IZ procedures were statistically conservative in the sense of requiring many replications. More recent work with screening or sequential IZ procedures reduces the number of replications (Nelson et al., 2001; Goldsman et al., 2002).

Bayesian procedures model the evidence for correct selection with the posterior distribution of the unknown means and variances, given the data seen so far from a single application of the procedure. The basic Bayesian formulation is developed before the differences between the VIP and OCBA approaches are described. Given all output $\mathcal{E}$ seen so far for a single application of a procedure, the posterior probability of correct selection is

$$\mathrm{PCS}_{\mathrm{Bayes}} = \Pr\left(W_{\mathfrak{D}} = W_{[k]} \,|\, \mathcal{E}\right) \tag{17}$$

$$= 1 - E[\mathcal{L}_{0-1}(\mathfrak{D}, \mathbf{W}) \,|\, \mathcal{E}], \tag{18}$$

where the 0-1 loss function $\mathcal{L}_{0-1}(\mathfrak{D}, \mathbf{w})$ equals 1 if $w_{\mathfrak{D}} = w_{[k]}$ and 0 otherwise. The expectation is taken over the decision, $\mathfrak{D}$, and the posterior distribution of the unknown means and variances. Assuming a noninformative prior distribution for the unknown mean and variance, the posterior marginal distribution for the unknown means W_i given $r_i > 3$ samples is $\mathtt{St}\,(\bar{y}_i, r_i/\hat{\sigma}_i^2, \upsilon_i)$, a shifted Student t distribution with mean $\bar{y}_i$, degrees of freedom $\upsilon_i = r_i - 1$, and variance $(\hat{\sigma}_i^2/r_i)\upsilon_i/(\upsilon_i - 2)$ (de Groot, 1970; Chick and Inoue, 2001a).

A comparison of Equation (15) and Equation (17) emphasizes the difference between the PCS based on frequentist methods and Bayesian methods. Frequentist methods provide worst-case bounds for $\mathrm{PCS}_{\mathrm{iz}}(\boldsymbol{\chi})$ subject to constraints on $\boldsymbol{\chi}$, and $\mathrm{PCS}_{\mathrm{iz}}(\boldsymbol{\chi})$ is estimated by counting the fraction of correct selections from repeated applications of the procedure. Bayesian methods provide a measure of evidence for correct selection given the data seen in a single application of a procedure. Equation (18) emphasizes the link to the Bayesian decision-theoretic methods in Section 1.2. That link can be extended by generalizing the opportunity cost to the selection context here, $\mathcal{L}_{oc}(\mathfrak{D}, \mathbf{w}) = w_{[k]} - w_{\mathfrak{D}}$. The loss is 0 when the best system is selected, and is the difference between the best and the selected system otherwise. This is

an alternate measure of evidence for correct selection that makes more sense than PCS when simulation output represents financial value. The posterior expectation of the opportunity cost of a potentially incorrect selection is

$$\mathrm{EOC}_{\mathrm{Bayes}} = E\left[\mathcal{L}_{oc}(\mathfrak{D}, \mathbf{W}) \,|\, \mathcal{E}\right] = E\left[W_{[k]} - W_{\mathfrak{D}} \,|\, \mathcal{E}\right]. \tag{19}$$

The frequentist $\mathrm{EOC}_{\mathrm{iZ}}(\boldsymbol{\chi}) = E\left[w_{[k]} - w_{\mathfrak{D}} \,|\, \mathbf{w}, \boldsymbol{\sigma}^2\right]$ differs by having the expectation taken only over randomized $\mathfrak{D}$ for a given $\boldsymbol{\chi}$. This formalism is sufficient to describe the basic ideas behind the VIP and OCBA selection procedures.

5.1 Value of Information Procedures (VIPs)

Value of Information Procedures (VIPs) allocate additional samples in order to improve the expected value of information (EVI) of those samples with respect to a loss function. Chick and Inoue (2001a) provide four procedures. Both two-stage and sequential procedures are given for both the 0-1 loss function and opportunity cost. The EVI of additional samples are measured using the predictive distribution of additional output, along with the expected loss from Equation (18) or Equation (19), with ideas paralleling those in Section 1.2.

After the first stage of sampling of $r_i = r_0$ replications per system in a VIP, the posterior distribution of the unknown mean and variance of each system is used as a prior distribution for the second stage. If noninformative prior distributions are used for the unknown mean and variance, the unknown means have a t distribution as described after Equation (18). The goal is to determine the second-stage allocation $\mathbf{r}' = (r'_1, r'_2, \ldots, r'_k)^T$ that minimizes the *expected loss* to a decision maker after all replications have been run. Let $\boldsymbol{x}_{r'_i} = (x_{i,r_0+1}, \ldots, x_{i,r_0+r'_i})$ denote the second-stage output for system i, let $\boldsymbol{x}_{\mathbf{r}'} = (\boldsymbol{x}_{r'_1}, \ldots, \boldsymbol{x}_{r'_k})$ denote all second-stage output, and let $\mathfrak{D}(\boldsymbol{x}_{\mathbf{r}'})$ be the system with the highest overall sample mean after both stages. Given $\boldsymbol{x}_{\mathbf{r}'}$ and a loss function $\mathcal{L}$, the expected loss is $E[\mathcal{L}\left(\mathfrak{D}(\boldsymbol{x}_{\mathbf{r}'}), \mathbf{W}\right) \mid \boldsymbol{x}_{\mathbf{r}'}]$. Since $\mathbf{r}'$ is chosen before the second stage, we take the expectation with respect to the predictive distribuion of $\boldsymbol{X}_{\mathbf{r}'}$. Let $\mathbf{c} = (c_1, \ldots, c_k)$ be the cost per replication of each system. The total cost of replications plus the expected loss for selecting the system with the best overall sample mean is (cf. Equation (7))

$$\rho^*(\mathbf{r}') \stackrel{\mathrm{def}}{=} \mathbf{c}\mathbf{r}' + E\left[E[\mathcal{L}\left(\mathfrak{D}(\boldsymbol{X}_{\mathbf{r}'}), \mathbf{W}\right) \mid \boldsymbol{X}_{\mathbf{r}'}]\right]. \tag{20}$$

If there is a computing budget constraint (e.g. for CPU time), samples are allocated to solve the following optimization problem.

$$
\begin{aligned}
\min_{\mathbf{r}'} \quad & \rho^*(\mathbf{r}') \\
\text{s.t.} \quad & \mathbf{c}\mathbf{r}' = B \\
& r'_i \geq 0 \quad \text{for } i = 1, \ldots, k
\end{aligned}
\tag{21}
$$

Gupta and Miescke (1994) solved a special case of Problem (21). If $k = 2$ and $c_1 = c_2$, the optimal second-stage allocation minimizes the absolute difference of the posterior precision for the mean of each system, regardless of whether the 0-1 loss or opportunity cost is used. For the opportunity cost, $k \geq 2$, $c_1 = \ldots = c_k = 1$, and known precision, Gupta and Miescke (1996) provide an optimal allocation if $B = 1$.

Those special cases are not sufficient to address the complexity of problems found in simulation, and approximations are required to obtain readily computable allocations. Chick and Inoue (2001a) derived asymptotically optimal allocations that minimize a bound on the expected loss in Equation (20), a formulation that allows for unequal, unknown variances, different sampling costs, and a balance between sampling costs and the EVI of the samples. *The bound is obtained by examining the $k-1$ pairwise comparisons between the system with the highest first-stage sample mean and each other system.* The asymptotic approximation is like that for Equation (10). If $k = 2$, the bound is tight for the opportunity cost but is loose for the 0-1 loss due to an extra asymptotic approximation. *Sequential procedures* require one more approximation, as the number of replications of each system may be different after a given stage. This means that the EVI requires assessing the difference $W_i - W_j$ of t distributed random variables with different degrees of freedom (the Behrens-Fisher problem). The Welch (1938) approximation can be used to approximate the EVI and expected loss in Equation (21) (Chick and Inoue, 2001a). The resulting EOC and PCS VIPs minimize the following measures of incorrect selection.

$$
\begin{aligned}
\text{EOC}_{\text{Bonf}} &= \sum_{i \neq \mathfrak{D}} E\left[\max\{0, W_i - W_{\mathfrak{D}}\} \,|\, \mathcal{E}\right] \\
1 - \text{PCS}_{\text{Bonf}} &= \sum_{i \neq \mathfrak{D}} \Pr\left(\{W_i \geq W_{\mathfrak{D}}\} \,|\, \mathcal{E}\right)
\end{aligned}
$$

5.2 OCBA Procedures

The third approach is the optimal computing budget allocation (OCBA) (Chen, 1996). The OCBA is based on several approximations, including the thought experiment in Equation (11) that approximates how additional replications would affect uncertainty about each W_i. Samples are allocated sequentially in a greedy manner to maximize an approximation to $\text{PCS}_{\text{Bayes}}$ at each

stage. The approximations made by the original OCBA essentially assume that (a) the system with the best sample mean based on replications seen so far is to be selected, (b) a normal distribution can approximate uncertainty about each W_i, and (c) the effect of an additional small number of replications r' beyond the r_i done so far for system i, but none for the other systems, changes the uncertainty about the means to

$$\begin{aligned} p(\tilde{W}_i) &\sim \texttt{Normal}\left(\bar{y}_i, \hat{\sigma}_i^2/(r_i + r')\right) \\ p(\tilde{W}_j) &\sim \texttt{Normal}\left(\bar{y}_j, \hat{\sigma}_j^2/r_j\right) \qquad \text{for } j \neq i. \end{aligned}$$

This induces an *estimated approximate probability of correct selection* that approximates the probability of correct selection with respect to $p(\tilde{\mathbf{W}})$ and Slepian's inequality (Chapter 17).

$$\text{EAPCS}_i = \prod_{j:j\neq(k)} \Pr\left(\left\{\tilde{\mathbf{W}} : \tilde{W}_j < \tilde{W}_{(k)}\right\} \mid \mathcal{E}\right). \tag{22}$$

A small number r' of replications are allocated at each stage to the q systems with the largest $\text{EAPCS}_i - \text{PCS}_{\text{Slep}}$, where the posterior evidence for correct selection is approximated with Slepian's inequality,

$$\text{PCS}_{\text{Slep}} = \prod_{j:j\neq(k)} \Pr\left(\left\{\mathbf{W} : W_j < W_{(k)}\right\} \mid \mathcal{E}\right),$$

and using the Welch approximation for the differences $W_j - W_{(k)}$. Chen et al. (2004, references therein) explored several variations on this theme, including varying r' and q; and the use of either normal or t distributions for W_i.

5.3 Comments

Surprisingly few papers to date compare representatives from each of the IZ, VIP and OCBA approaches. Chick and Inoue (2001a) found that VIPs compared favorably with the Combined Procedure of Nelson et al. (2001), when PCS_{iz} was measured against the average number of replications per system. Inoue et al. (1999) compared the VIP, OCBA, and modified versions of the IZ procedure of Rinott (1978). Both two-stage and sequential VIPs performed well empirically over a broad class of problems. The VIP based on opportunity cost fared best with respect to several performance criteria. The VIP based on the 0-1 loss performed slightly less well than the opportunity cost procedure, even with respect to PCS_{iz}, because of an extra approximation in its derivation. The fully sequential OCBA was also empirically very effective. The OCBA performed less well when run as a two-stage procedure, rather than sequentially, or when the values of r' and q were large (Inoue et al., 1999).

Branke, Chick and Schmidt (2005) evaluate a variety of procedures, and introduce new 'stopping rules' for the VIP and OCBA procedures. The sequential stopping rule $\mathcal{S}$ samples τ replications per stage of sampling (after the first) until a fixed sampling budget is exhausted. That rule was used by Inoue et al. (1999). The new EOC stopping rule samples τ replications per stage (after the first) until a Bonferroni-like bound for the EOC is reached ($\text{EOC}_{\text{Bonf}} \leq \beta^*$ for some user-specified β^*). Other stopping rules, such as for PCS_{Slep} are implemented similarly. Figure 5 gives a typical result that compares VIPs and OCBA procedures endowed with these stopping rules ($\tau = 1$), and a version of a leading sequential IZ procedure, $\mathcal{KN}$++ (Goldsman et al., 2002), adapted for independent output. As the number of replications or target posterior EOC values are changed for VIP and OCBA procedures, a different average number of replications and evidence for correct selection are observed. The curves are dependent upon the specific problem instance, which in this case is a single problem instance with evenly spaced means and a common variance, but some observations can be made.

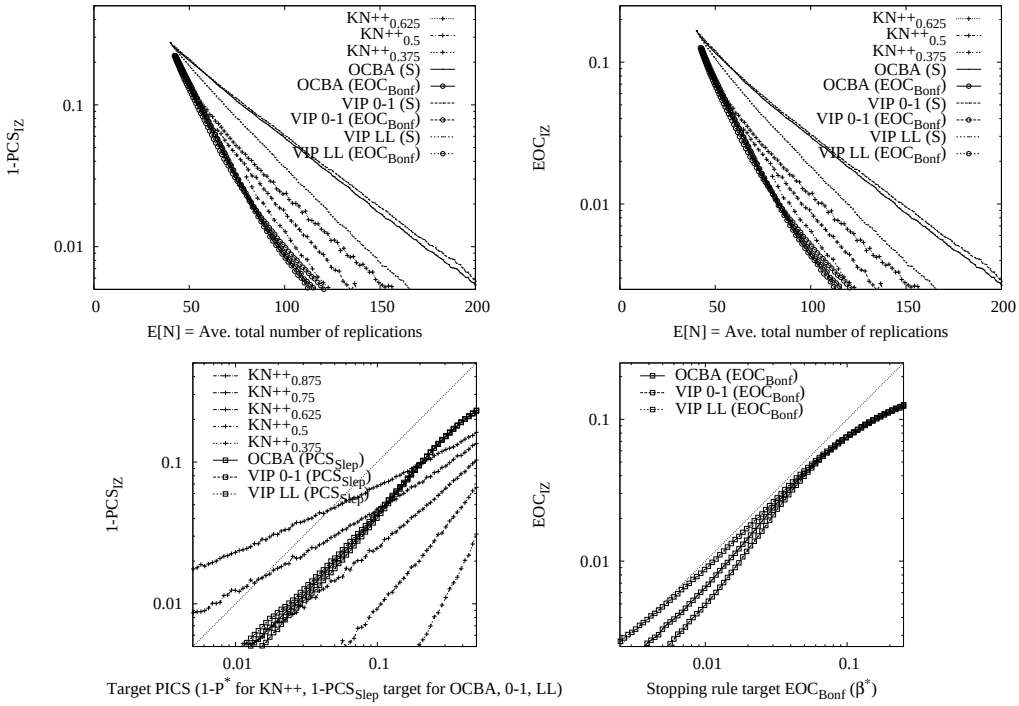

Fig. 5. Estimated evidence for correct selection as a function of the average number of replications per stage (top row), and relationship of parameter target versus empirical result (bottom row), for several several selection procedures ($k = 10$ systems, $w_1 - w_i = (i-1)/2$; $\sigma_i^2 = 1$ for $i = 1, 2, \ldots, k$, with $r_0 = 5$, estimates based on 10^5 macroreplications).

In the top graphs of Figure 5, lower curves mean that more evidence per correct selection is obtained per average number of replications. For $\mathcal{KN}$++, δ^* was fixed to certain levels equal to, less than, and greater than the true difference in means between the best and second best (0.5 in this case), and the PCS goal P^*

was varied to obtain similar curves. The $\mathcal{S}$ stopping rule allows a practitioner to completely control the number of replications. Procedure $\mathcal{KN}$++ improves efficiency here by allowing a procedure to stop early if the evidence for correct selection is high early, or to continue sampling if more information is needed. The Bayesian VIP and OCBA procedures similarly can adapt the amount of sampling as a function of the posterior evidence for correct selection so far (here measured with $\mathrm{EOC}_{\mathrm{Bonf}}$), but improved upon $\mathcal{KN}$++ here due to more flexibility as to which systems get sampled. The bottom graphs in Figure 5 show the relationship between the targeted and empirically observed evidence for correct selection. The diagonal represents perfect calibration. The OCBA and VIP are slightly conservative (below the line), meaning that the observed $\mathrm{PCS}_{\mathrm{iz}}$ and $\mathrm{EOC}_{\mathrm{iz}}$ for this problem instance is somewhat 'better' than the corresponding Bayesian target levels. The graphs for $\mathcal{KN}$++ are straight, but may significantly underdeliver or overdeliver relative to the desired $\mathrm{PCS}_{\mathrm{iz}}$ depending upon whether δ^* is selected to be larger, similar to, or smaller than the difference in means between the top two systems. A small δ^* results in excessive sampling for a given target P^*. For $\mathcal{KN}$++, the desired PCS and empirical $\mathrm{PCS}_{\mathrm{iz}}$ tracked each other better for slippage configurations (all nonbest systems have the same mean) when δ^* was exactly the difference between the best and second best. Branke, Chick and Schmidt (2005) provide a more thorough analysis and empirical study.

Common random numbers (CRN) can be used to sharpen contrasts across systems in two stage procedures. To date, the IZ approach provides more choices for procedures that accomodate CRN than the Bayesian approach. Kim and Nelson (2001) describe recent progress for IZ procedures with CRN. Chick and Inoue (2001b) present two-stage VIP procedures that handle CRN for both the opportunity cost and 0-1 loss functions. The procedures allow for simulating only a subset of systems during a second stage, but are not sequential. Because of correlation across systems, information is gained about the systems not simulated during the second stage using missing data techniques. Fu et al. (2004) examine CRN for the OCBA with sequential sampling.

CRN is typically implemented by using the same random number generators $\boldsymbol{u}_r$ to synchronize the stochastic uncertainty between simulated systems, so that the outputs Y_{ri} and Y_{rj} are positively correlated for replication r of systems i and j. When there is also input uncertainty, so that $(\boldsymbol{\Theta}_p, \boldsymbol{\Theta}_e)$ are described with probability rather than assumed to be a fixed input value, a second method for inducing CRN exists. Common values of $(\boldsymbol{\theta}_{pr}, \boldsymbol{\theta}_{er})$ are used for the r-th simulation of each system. This can be useful for sharpening contrasts for systems even when input uncertainty is modeled.

Cross-fertizilation between the approaches is also possible. Chick and Wu (2005) applied the opportunity cost idea from the VIP approach to the IZ framework to develop a two-stage procedure with a frequentist expected op-

portunity cost bound. The net effect is to replace the two parameters P^*, δ^* of the IZ approach with a single parameter Δ, the maximum acceptable expected opportunity cost, so that the frequentist expected opportunity cost $E\left[w_{[k]} - w_{\mathfrak{D}} \,|\, \boldsymbol{\chi} = (\mathbf{w}, \boldsymbol{\sigma}^2)\right] \leq \Delta$ for all problem instances $\boldsymbol{\chi}$ (not just those in an indifference zone).

6 Discussion and future directions

Bayesian methods apply to simulation experiments in a variety of ways, including uncertainty analysis, ranking and selection, input distribution modeling, response surface modeling, and experimental design. One main theme is to represent all uncertainty with probability distributions, to update probability using Bayes' rule, and to use the expected value of information as a technique to make sampling decisions (e.g. the opportunity cost and 0-1 loss functions for selection procedures, or the Kullback-Leibler divergence for parameter estimation for linear response models). The other main theme is to use simulation to efficiently estimate quantities of interest for a Bayesian analysis. Asymptotic approximations are often helpful when exact optimal solutions are difficult to obtain. Research opportunities include:

- Input modeling and uncertainty analysis: kernel estimation for conditional means, with variability due to input uncertainty; improved modeling of prior distributions for statistical input parameters to obtain better models of uncertainty for simulation outputs (e.g. the conjugate prior distributions for parameters of an M/M/1 queue result in the absence of moments for quantities like the stationary average queue length, even conditioning on stability), including results for the calibration/inverse problem.
- Response modeling: extending the Gaussian random function work in the world of stochastic simulation; sampling plans for input parameter inference to optimally reduce output uncertainty, including nonasymptotic results, to help understand what data is most important to collect to infer the value of inputs for simulations.
- Ranking and selection: VIP procedures based on nonasymptotic EVI allocations for samples; analytical or empirical work to evaluate the IZ, VIP and OCBA approaches.
- Experimental designs: Estimating quantiles or other non-expected value goals; non-Gaussian output for ranking and selection and GRFs.
- Computational methods: improvements in MCMC and other sampling methods for posterior distributions.

Acknowledgments. Portions of this chapter were published by Chick (2004) in the *Proceedings of the Winter Simulation Conference.*

References

[1] Andradóttir, S., Bier, V. M., 2000. Applying Bayesian ideas in simulation. Simulation Theory and Practice 8, 253–280.

[2] Andradóttir, S., Glynn, P. W., 2004. Computing Bayesian means using simulation. Georgia Tech, ISYE Technical Report.

[3] Armero, C., Bayarri, M., 1997. A Bayesian analysis of a queueing system with unlimited service. J. Statistical Planning and Inference 58, 241–261.

[4] Barlow, R. E., Mendel, M. B., 1992. De Finetti-type representations for life distributions. J. American Statistical Association 87, 1116–1122.

[5] Barton, R. R., Schruben, L. W., 2001. Simulating real systems. In submission.

[6] Bechhofer, R. E., Santner, T. J., Goldsman, D. M., 1995. Design and Analysis for Statistical Selection, Screening, and Multiple Comparisons. John Wiley & Sons, Inc., New York.

[7] Berger, J. O., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd Edition. Springer-Verlag, New York.

[8] Berger, J. O., 1994. An overview of robust Bayesian analysis. TEST 3, 5–124.

[9] Berger, J. O., Delampady, M., 1987. Testing precise hypotheses (with discussion). Statistical Science 2, 317–352.

[10] Berger, J. O., Pericchi, L. R., 1996. The intrinsic Bayes factor for model selection and prediction. J. American Statistical Association 91, 109–122.

[11] Berk, R., 1966. Limiting behaviour of posterior distributions when the model is incorrect. Annals of Mathematical Statistics 37 (1), 51–58.

[12] Bernardo, J. M., Smith, A. F. M., 1994. Bayesian Theory. Wiley, Chichester, UK.

[13] Billingsley, P., 1986. Probability and Measure, 2nd Edition. John Wiley & Sons, Inc., New York.

[14] Branke, J., Chick, S.E., Schmidt, C., 2005. Selecting a selection procedure. INSEAD, Technology Management Area working paper.

[15] Chen, C.-H., 1996. A lower bound for the correct subset-selection probability and its application to discrete event simulations. IEEE Transactions on Automatic Control 41 (8), 1227–1231.

[16] Chen, C.-H., Yücesan, E., Dai, L., Chen, H., 2004. Efficient computation of optimal budget allocation for discrete event simulation experiment. IIE Transactions , to appear.

[17] Chen, H., Schmeiser, B., 1995. Monte Carlo estimation for guaranteed-coverage non-normal tolerance intervals. J. Statistical Computation and Simulation 51 (2-4), 223–238.

[18] Chen, H.-C., Chen, C.-H., Lin, J., Yücesan, E., 1999. An asymptotic allocation for simultaneous simulation experiments. In: Farrington, P. A., Nembhard, H. B., Sturrock, D., Evans, G. (Eds.), Proc. 1999 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 359–366.

[19] Cheng, R. C. H., 1999. Regression metamodelling in simulation using Bayesian methods. In: Farrington, P. A., Nembhard, H. B., Sturrock, D., Evans, G. (Eds.), Proc. 1999 Winter Simulation Conference. IEEE, Inc., Piscataway,

NJ, pp. 330–335.
[20] Cheng, R. C. H., Currie, C. S., 2003. Prior and candidate models in the Bayesian analysis of finite mixtures. In: Chick, S., Sánchez, P., Ferrin, D., Morrice, D. (Eds.), Proc. 2003 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 392–398.
[21] Cheng, R. C. H., Holland, W., 1997. Sensitivity of computer simulation experiments to errors in input data. J. Statistical Computing and Simulation 57, 219–241.
[22] Chick, S. E., 1997. Bayesian analysis for simulation input and output. In: Andradóttir, S., Healy, K., Withers, D., Nelson, B. (Eds.), Proc. 1997 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 253–260.
[23] Chick, S. E., 2001. Input distribution selection for simulation experiments: Accounting for input uncertainty. Operations Research 49 (5), 744–758.
[24] Chick, S. E., 2004. Bayesian methods for discrete event simulation. In: Ingalls, R., Rossetti, M., Smith, J., Peters, B. (Eds.), Proc. 2004 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 89–100.
[25] Chick, S. E., Inoue, K., 2001a. New two-stage and sequential procedures for selecting the best simulated system. Operations Research 49 (5), 732–743.
[26] Chick, S. E., Inoue, K., 2001b. New procedures for identifying the best simulated system using common random numbers. Management Science 47 (8), 1133–1149.
[27] Chick, S. E., Inoue, K., 2002. Corrigendum: New selection procedures. Operations Research 50 (3), 566.
[28] Chick, S. E., Mendel, M. B., 1998. New characterizations of the no-aging property and the ℓ_1-isotropic models. J. Appl. Probability 35 (4), 903–910.
[29] Chick, S. E., Soorapanth, S., Koopman, J. S., 2003. Inferring infection transmission parameters that influence water treatment decisions. Management Science 49 (7), 920–935.
[30] Chick, S. E., Wu, Y., 2005. Selection procedures with frequentist expected opportunity cost bounds. Operations Research 53, to appear.
[31] Craig, P., Goldstein, M., Rougier, J., Seheult, A. H., 2001. Bayesian forecasting for complex systems using computer simulators. J. American Statistical Association 96, 717–729.
[32] Cressie, N. A., 1993. Statistics for Spatial Data. J. Wiley & Sons, New York.
[33] de Finetti, B., 1990. Theory of Probability, v. 2. J. Wiley & Sons, New York.
[34] de Groot, M. H., 1970. Optimal Statistical Decisions. McGraw-Hill, New York.
[35] Dmochowski, J., 1999. Intrinsic priors via Kullback-Leibler geometry. In: Bayesian Statistics 5. Oxford University Press, pp. 543–549.
[36] Draper, D., 1995. Assessment and propogation of model uncertainty (with discussion). J. Royal Statistical Society, Series B 57 (1), 45–97.
[37] Dudewicz, E. J., Dalal, S. R., 1975. Allocation of observations in ranking and selection with unequal variances. Sanhkya B37 (1), 28–78.
[38] Edwards, A. W. F., 1984. Likelihood. Cambridge University Press.
[39] Evans, M., Swartz, T., 1995. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. Statistical Science

10 (3), 254–272.

[40] Fu, M. C., Hu, J.-Q., Chen, C.-H., Xiong, X., 2004. Optimal computing budget allocation under correlated sampling. In: Ingalls, R., Rossetti, M., Smith, J., Peters, B. (Eds.), Proc. 2004 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 595–603.

[41] George, E. I., McCulloch, R. E., 1996. Stochastic search variable selection. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman and Hall, London, pp. 203–214.

[42] Gilks, W. R., Best, N. G., Tan, K. K. C., 1995. Adaptive rejection metropolis sampling. Applied Statistics 44, 455–472.

[43] Gilks, W. R., Richardson, S., Spiegelhalter, D. J., 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall, London.

[44] Glynn, P., 1986. Problems in Bayesian analysis of stochastic simulation. In: Wilson, J. R., Henriksen, J. O., Roberts, S. D. (Eds.), Proc. 1986 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 376–383.

[45] Goldsman, D., Kim, S.-H., Marshall, W. S., Nelson, B. L., 2002. Ranking and selection for steady-state simulation: Procedures and perspectives. INFORMS Journal on Computing 14 (1), 2–19.

[46] Gupta, S. S., Miescke, K. J., 1994. Bayesian look ahead one stage sampling allocations for selecting the largest normal mean. Statistical Papers 35, 169–177.

[47] Gupta, S. S., Miescke, K. J., 1996. Bayesian look ahead one-stage sampling allocations for selecting the best population. J. Statistical Planning and Inference 54, 229–244.

[48] Inoue, K., Chick, S. E., Chen, C.-H., 1999. An empirical evaluation of several methods to select the best system. ACM Transactions on Modeling and Computer Simulation 9 (4), 381–407.

[49] Jaynes, E. T., 1983. E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics. D. Reidel Publishing Co., Dordrecht, Holland.

[50] Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. London. Ser. A. 186, 453–461.

[51] Kahneman, D., Slovic, P., Tversky, A., 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge.

[52] Kass, R. E., 1989. The geometry of asymptotic inference. Statistical Science 4 (3), 188–234.

[53] Kass, R. E., Wasserman, L., 1996. The selection of prior distributions by formal rules. J. American Statistical Association 91 (435), 1343–1370.

[54] Kennedy, M., O'Hagan, A., 2001. Bayesian calibration of computer models (with discussion). J. Royal Statistical Society B 63, 425–464.

[55] Kim, S.-H., Nelson, B. L., 2001. A fully sequential procedure for indifference-zone selection in simulation. ACM TOMACS 11, 251–273.

[56] Kraan, B., Bedford, T., 2003. Probabilistic inversion of expert judgements in the quantification of model uncertainty. Research Paper No. 2003/12, Strathclyde Business School, Glasgow.

[57] Laplace, P. S., 1812. Theorie Analytique des Probabilites. Courcier, Paris.

[58] Larson, R. C., 1990. The queue inference engine: Deducing queue statistics from transactional data. Management Science 36, 586–601.

[59] Law, A. M., Kelton, W. D., 2000. Simulation Modeling & Analysis, 3rd Edition. McGraw-Hill, Inc, New York.

[60] Lee, S.-H., Glynn, P., 1999. Computing the distribution function of a conditional expectation via Monte Carlo: Discrete conditioning spaces. In: Farrington, P. A., Nembhard, H. B., Sturrock, D., Evans, G. (Eds.), Proc. 1999 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 1654–1663.

[61] Lindley, D. D., 1957. A statistical paradox. Biometrika 44, 187–192.

[62] Lindley, D. V., 1972. Bayesian Statistics, A Review. Society of Industrial and Applied Mathematics, Philadelphia.

[63] McGrath, M. F., Gross, D., Singpurwalla, N. D., 1987. A subjective Bayesian approach to the theory of queues I–modeling. Queueing Systems 1, 317–333.

[64] McGrath, M. F., Singpurwalla, N. D., 1987. A subjective Bayesian approach to the theory of queues II–inference and information in M/M/1 queues. Queueing Systems 1, 335–353.

[65] Nahmias, S., 2000. Production and Operations Management, 4th Edition. McGraw Hill.

[66] Naylor, J., Smith, A., 1982. Applications of a method for the efficient computation of posterior distributions. Appl. Statist. 31, 314–225.

[67] Nelson, B. L., Schmeiser, B. W., Taaffe, M. R., Wang, J., 1997. Approximation-assisted point estimation. Operations Research Letters 20 (3), 109–118.

[68] Nelson, B. L., Swann, J., Goldsman, D., Song, W., 2001. Simple procedures for selecting the best simulated system when the number of alternatives is large. Operations Research 49 (6), 950–963.

[69] Ng, S.-H., Chick, S. E., 2001. Reducing input distribution uncertainty for simulations. In: Peters, B., Smith, J., Rohrer, M., Madeiros, D. (Eds.), Proc. 2001 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, p. 364–371.

[70] Ng, S.-H., Chick, S. E., 2004. Reducing parameter uncertainty for stochastic systems. Tech. rep., INSEAD, Technology Management Area, Fontainebleau, France.

[71] Ng, S.-H., Chick, S. E., 2004. Design of follow-up experiments for improving model discrimination and parameter estimation. Naval Research Logistics 51 (8), 1129-1148.

[72] Oakley, J., 2004. Estimating percentiles of uncertain computer code outputs. Appl. Statist. 53, 83–93.

[73] O'Hagan, A., 1995. Fractional Bayes factors for model comparison (with discussion). J. Royal Statistical Society, Series B 56, 99–118.

[74] O'Hagan, A., Kennedy, M., Oakley, J., 1999. Uncertainty analysis and other inference tools for complex computer codes (with discussion). In: Bayesian Statistics 6. Oxford University Press, pp. 503–524.

[75] O'Neill, P. D., 2002. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Math. Biosci. 180, 103–114.

[76] Raftery, A. E., 1995. Bayesian model selection in social research (with dis-

cussion by A. Gelman, D.B. Rubin, and R.M. Hauser, and a rejoinder). In: Marsden, P. V. (Ed.), Sociological Methodology 1995. Blackwells, Cambridge, Mass., pp. 111–196.

[77] Richardson, S., Green, P., 1997. On Bayesian analysis of mixtures with an unknown number of components. J. Royal Statistical Society, B 59 (4), 731–792.

[78] Rinott, Y., 1978. On two-stage selection procedures and related probability-inequalities. Communications in Statistics A7 (8), 799–811.

[79] Sacks, J., Welch, W., Mitchell, T., Wynn, H., 1989. Design and analysis of computer experiments. Statistical Science 4, 409–435.

[80] Santner, T. J., Williams, B. J., Notz, W. I., 2003. The Design and Analysis of Computer Experiments. Springer, New York.

[81] Savage, L. J., 1972. The Foundations of Statistics. Dover Publications, Inc., New York.

[82] Scott, E. M., 1996. Uncertainty and sensitivity studies of models of environmental systems. In: Charnes, J. M., Morrice, D. J., Brunner, D. T., Swain, J. J. (Eds.), Proc. 1996 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 255–259.

[83] Smith, B. J., 2004. Bayesian output analysis program (BOA) v. 1.1 user's manual. Tech. rep., Univ. Iowa, `http://www.public-health.uiowa.edu/boa/`.

[84] Spiegelhalter, D. J., Best, N. G., Gilks, W. R., Inskip, H., 1996. Hepatitis B: a case study in MCMC methods. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman and Hall, London, pp. 21–44.

[85] Steckley, S. G., Henderson, S. G., 2003. A kernel approach to estimating the density of a conditional expectation. In: Chick, S., Sánchez, P., Ferrin, D., Morrice, D. (Eds.), Proc. 2003 Winter Simulation Conference. IEEE, Inc., Piscataway, NJ, pp. 383–391.

[86] Tanner, M. A., 1996. Tools for Statistical Inference: Methods for exploration of posterior distributions and likelihood functions, 3rd Edition. Springer.

[87] The Mathworks, 2002. Getting Started With MATLAB. Natick, MA, USA.

[88] van Beers, W., Kleijnen, J. P. C., 2003. Kriging for interpolation in random simulation. J. Operational Research Society 54, 255–262.

[89] Welch, B. L. 1938. The significance of the difference between two means when the population variances are unequal. Biometrika 25, 350–362.

[90] Winston, W., 2000. Financial Models Using Simulation and Optimization, 2nd Edition. Palisade Corp.

[91] Zouaoui, F., Wilson, J. R., 2003. Accounting for parameter uncertainty in simulation input modeling. IIE Transactions 35, 781–792.

[92] Zouaoui, F., Wilson, J. R., 2004. Accounting for input-model and input-parameter uncertainties in simulation. IIE Transactions 36 (11), 1135-1151.