Abstract

ZOUAOUI, FAKER. Accounting For Input Uncertainty in Discrete-Event Simulation. (Under the direction of James R. Wilson.)

The primary objectives of this research are formulation and evaluation of a Bayesian approach for selecting input models in discrete-event stochastic simulation. This approach takes into account the model, parameter, and stochastic uncertainties that are inherent in most simulation experiments in order to yield valid predictive inferences about the output quantities of interest. We use prior information to specify the prior plausibility of each candidate input model that adequately fits the data, and to construct prior distributions on the parameters of each model. We combine prior information with the likelihood function of the data to compute the posterior model probabilities and the posterior parameter distributions using Bayes' rule. This leads to a Bayesian Simulation Replication Algorithm in which: (a) we estimate the parameter uncertainty by sampling from the posterior distribution of each model's parameters on selected simulation runs; (b) we estimate the stochastic uncertainty by multiple independent replications of those selected runs; and (c) we estimate model uncertainty by weighting the results of (a) and (b) using the corresponding posterior model probabilities. We also construct a confidence interval on the posterior mean response from the output of the algorithm, and we develop a replication allocation procedure that optimally allocates simulation runs to input models so as to minimize the variance of the mean estimator subject to a budget constraint on computer time. To assess the performance of the algorithm, we propose some evaluation criteria that are reasonable within both the Bayesian and frequentist paradigms. An experimental performance evaluation demonstrates the advantages of the Bayesian approach versus conventional frequentist techniques.

ACCOUNTING FOR INPUT UNCERTAINTY IN DISCRETE-EVENT SIMULATION

by

Faker Zouaoui

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

OPERATIONS RESEARCH

Raleigh

2001

APPROVED BY:

_____	_____
_____	_____
	Chair of Advisory Committee

To
my wife Çiğdem, my daughter Alya,
and
my parents Ammar and Radhia

Biography

Faker Zouaoui was born on October 6, 1972, in Sfax, Tunisia. He graduated from the English Pioneer School of Ariana, Tunisia, in 1991. He received a government scholarship to pursue his studies in Bilkent University, Ankara, Turkey, where he received a Bachelor of Science in Industrial Engineering in 1995, and a Master of Science in Industrial Engineering in 1997. He also worked as a teaching assistant in the Industrial Engineering Department of Bilkent University for two years.

In August 1997, he joined North Carolina State University to pursue his doctoral studies. He served as a teaching and research assistant in the Statistics and Industrial Engineering Departments. After completing his Ph.D. degree in Operations Research, he plans to pursue a career in industry as an operations research consultant. His research interests include simulation modeling, Bayesian statistics, and stochastic processes.

Acknowledgments

I thank Professor James R. Wilson for his guidance and support in the execution of this research and the composition of this dissertation. I also thank Professors Stephen D. Roberts, Sujit Ghosh, and Bibhuti B. Bhattacharyya for serving as advisors on my committee and for their constructive input. This research was supported by the National Science Foundation under grant number DMI 9900164.

Finally, I would like to express my gratitude to many other people who supported me in this effort. My parents Ammar and Radhia, my sister Neila, and my parents-in-law Fehmi and Fikriye gave me encouragement when it was most needed. My brother Achraf and my friend Souheyl listened to hours of complaining through long distance calls. My daughter Alya endured my bad moods and absences with a smile. But most importantly Çiğdem, my wife, did all of the above and did it unceasingly.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statement and Research Objectives

Discrete-event simulation is a widely used tool for the analysis of the performance of complex stochastic systems with applications in manufacturing, service, and production systems. One of the main problems in the design of simulation experiments is to determine or specify valid input models that characterize the stochastic behavior of the modeled system (Wagner and Wilson, 1996). Since in practice, apart from rare situations, a model specification is never "correct," simulation practitioners must carefully address the issue of modeling input processes. Otherwise, the simulation output data will be misleading and possibly damaging or costly when used for decision making.

In the presence of data, the usual approach to model selection in the simulation community rests on the use of classical statistical techniques. Strategies for doing this are commonly guided by a series of significance tests, often based on the approximate asymptotic distribution of a test statistic (e.g., chi-squared goodness-of-fit tests). Goodness-of-fit measures based on these tests such as P-values are difficult to interpret and might be highly misleading (Berger and Delampady, 1987). In small samples, these tests have a very low power to detect a lack of fit between the empirical distribution and each candidate distribution, resulting in an inability to reject any of the candidate distributions. In large samples, minor discrepancies often appear to be statistically significant, resulting in rejection of all candidate distributions (Raftery, 1995). Moreover, analysis or comparison of multiple models, especially nonnested

ones, is very difficult in a classical framework (Miller, 1990).

Most fundamentally, any approach that selects a single model, fixes its parameters, and then yields an output inference conditionally on that model fails to account fully for the *model* and *parameter* uncertainties that are inherent in the basic estimation problem. This may lead to understatement of the inferential uncertainty assessments about the output quantities of interest, sometimes to a dramatic extent (Kass and Raftery, 1995). Some studies in the simulation literature have analyzed separately the sensitivity of output quantities to input parameter uncertainty (Cheng and Holland, 1997), and input model uncertainty (Gross and Juttijudatta, 1997). All these studies indicated that such an effect can be dramatic. None of these studies, however, considered the joint effect of both parameter and model uncertainty.

All these difficulties can be avoided, if one adopts a Bayesian approach which incorporates prior information into the model selection process in a formal and rigorous manner. In principle if prior information of adequate quality is available, then we can compute the posterior probabilities using Bayes' rule for all competing models. A composite inference can then be made that takes account of model and parameter uncertainty in a formally justifiable way. As a practical matter this idea was rejected in the statistical community for many decades because it is computationally quite expensive, if not impossible, in some problems. However, with the recent development of sampling-based approaches for calculating marginal and posterior probabilities (Gelfand, 1990), many statisticians have adopted the Bayesian approach to accounting for uncertainty in model selection. The approach known nowadays as Bayesian Model Averaging (BMA) (Raftery, 1996) has been used successfully in practical applications drawn from a broad diversity of fields such as econometrics, artificial intelligence, and medicine. See for example Draper (1995), Volinsky et al. (1997), and Hoeting et al. (1999).

In this dissertation, we have developed a Bayesian formulation of discrete-event stochastic simulation experiments using the BMA approach that takes into account the model, parameter, and stochastic uncertainties in order to yield valid predictive inferences about the output quantities of interest. This formulation leads to a "Bayesian Simulation Replication Algorithm" for designing simulation experiments that will be particularly useful in practice when input uncertainties are large. We have also derived point and credible (confidence) interval estimates for the posterior mean of the output response. These statistics will be used to evaluate the perfor-

mance of the Bayesian formulation compared to the classical and bootstrap methods using some Monte Carlo experiments.

1.2 Organization of the Dissertation

The remainder of this dissertation is organized into four chapters. In Chapter 2 we review briefly the current literature on the classical methods for selecting input models in discrete-event simulation, and we argue that these methods can lead to invalid (that is, unreliable) conclusions. To lay the foundation for our approach to solve this problem, we survey Bayesian model selection procedures and discuss in detail the BMA approach as a coherent mechanism to quantify the effects of model and parameter uncertainties on the distribution of simulation-generated performance measures of interest. We also review the results of previous attempts in the simulation literature to use Bayesian techniques, mainly in output analysis. In Chapter 3 we develop a Bayesian input modeling methodology to account for parameter and stochastic uncertainty. We also discuss point and interval estimation using the classical, bootstrap, and Bayesian methods; and we conduct some Monte Carlo experiments to compare the performance of all these approaches. In Chapter 4 we extend our Bayesian framework presented in Chapter 3 to account for model uncertainty. This is the first attempt in the simulation literature to quantify the effect of model and parameter uncertainty as well as the usual stochastic uncertainty on the output quantity of interest. We also develop a replication allocation procedure that optimally allocates simulation runs to input models so as to minimize the variance of the estimator of the posterior mean response subject to a budget constraint on the total amount of simulated experimentation (computer time) that is available. A Monte Carlo experiment will be used to illustrate these concepts, and evaluate the performance of the Bayesian approach versus the classical frequentist approach. Finally, in Chapter 5 we summarize the main contributions of this work, and we recommend some directions for further research on Bayesian methods for simulation input modeling.

Chapter 2

Literature Review

The literature review is organized into four sections. In Section 2.1 we describe the conventional approach to simulation input modeling and its major limitations. We also present briefly the recent literature on this topic. In Section 2.2 we survey the Bayesian model selection procedures in the statistical literature. This lays the foundation for our research to propose an alternative way for selecting valid input models in stochastic simulations when the uncertainties in the input processes are large. In Section 2.3 we discuss the Bayesian Model Averaging (BMA) approach as a coherent mechanism to quantify the effects of model and parameter uncertainties on the distribution of the performance measures of interest. Finally, in Section 2.4 we review the results of previous attempts in the simulation literature to use Bayesian techniques.

2.1 Simulation Input Modeling

Discrete-event simulation models typically have random components that mimic the stochastic behavior of the system under consideration. Interarrival and service times are examples of random components in a single-server queueing system. *Input modeling* is the process of selecting probability distributions or *input models* on these random components that match as close as possible the probabilistic mechanism associated with the system. Then, given that the random components follow particular probability distributions, the discrete-event simulation proceeds by generating random values from these distributions.

To select input models for a simulation experiment, a typical simulation analyst assumes that each input model is represented as a sequence of i.i.d random variables having a common distribution as one of the standard distributions (e.g., gamma, uniform, normal, beta, ...) included in almost all simulation languages. He later generates some summary statistics (e.g., min, max, quartiles, ...) and some graphical plots (e.g., histograms) to hypothesize some candidate distributions from a list of well-known distributions. Then, he fits these distributions to the data using maximum likelihood or moment matching. He finally verifies the fits using some visual inspection or goodness-of-fit tests and pick up the best fitting distribution. Among these tests are informal graphical techniques (Frequency comparisons and probability plots) as well as statistical goodness-of-fit tests (Kolmogorov-Smirnov, Chi-Squared, and Anderson-Darling). For a comprehensive discussion of these procedures, see Law and Kelton (2000). There are a number of software packages to support such simple input modeling approach, including ExpertFit (Law and McComas, 2000), Stat::fit and the Arena Input Analyzer (Kelton et al., 1998). Unfortunately, simple models often fail for one of the following reasons (Nelson and Yamnitsky, 1998):

- The shapes of the standard distributions are not flexible enough to represent some characteristics of the observed data.
- The input process may not be independent, either in time sequence or with respect to other input processes in the simulation.
- The input process changes over time.
- No data are available from which to select a family or assess the fit.

In the last two decades, several articles appeared in the simulation literature suggesting techniques that can be useful when simple models fail. We describe briefly these techniques, and highlight further complications in the input modeling process that received little or no attention from the simulation community.

2.1.1 Flexible Input Models

In many applications, simulation analysts often encounter data sets with many characteristics that are not captured by any of the standard distributions available in

their simulation packages. Several flexible distributions are proposed in the simulation literature with more flexible distributional shapes. Schmeiser and Deutsch (1977) proposed a four parameter family of probability distributions suitable for simulation. Swain et al. (1988) later developed a software package called FITTR1 to fit Johnson's translation system (Johnson, 1949). However, these distributions are inadequate for data sets having anomalies as simple as being bimodal with tails. Later Avramidis and Wilson (1994) developed the IDPF procedure to extend the parameterization of a standard distribution family by minimizing the sum of square errors between the inverse distribution function of this family and its inverse distribution function with a polynomial filter. They developed the IDPF software to improve the fit on the four parameter Johnson family. However, their method does not not always yield an acceptable fit to the data, and it does not always greatly improve the fit obtained from the Johnson family.

Finally, Wagner and Wilson (1996) developed a flexible, interactive, graphical methodology for modeling a broad range of input processes. They exploited the properties of Beziér curves to develop a flexible univariate distributional family, called the Bézier distribution, which is capable of taking an unlimited number of shapes. Bézier curves are often used to approximate a smooth univariate function on a bounded interval to pass in the vicinity of selected control points $\{\boldsymbol{p}_i = (x_i, z_i)^T : i = 0, \ldots, n\}$. A Bézier distribution with $n+1$ control points is defined as

$$\begin{aligned} \boldsymbol{P}(t) &= \{x(t), F_X(x(t))\}^T \\ &= \sum_{i=0}^{n} B_{n,i}(t)\boldsymbol{p}_i, \quad \forall t \in [0,1], \end{aligned}$$

where $\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ are arranged to ensure the basic requirements of a cumulative distribution function, and $B_{n,i}(t)$ is the Bernstein polynomial

$$B_{n,i}(t) = \begin{cases} \frac{n!}{i!(n-1)!} t^i (1-t)^{n-i}, & \text{for } t \in [0,1] \text{ and } i = 0, 1, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

The Bézier distribution can be constructed with or without data in a software package called PRIME that integrates graphical and statistical procedures to help select and visualize an appropriate representation of the input process.

2.1.2 Time-dependent Input Processes

Most of the input models that were built in the commercial simulation languages rely on the i.i.d assumption. However, many simulation models are driven by input processes which are time dependent. For example, the nonhomogeneous Poisson process (NHPP) is a generalization of the homogeneous Poisson process that allows for time dependent arrival rate. Lee et al. (1991) fitted an NHPP with an exponential rate function having polynomial and trigonometric components to capture long term and cyclic behaviors. Khul et al. (1997) extended these ideas to allow for multiple periodicities.

Cario and Nelson (1996) developed AutoRegressive To anything (ARTA) processes to construct a stationary time series with a given marginal distribution function and first p autocorrelations. They implemented their approach in ARTAFACTS and ARTAGEN software packages to generate a stationary ARTA process as input to a simulation model.

2.1.3 Dependent Input Models

In many simulation studies, successive observations on an input process may be autocorrelated. The TES methodology and software (Jagerman and Melamed, 1992) can be useful to model such processes. Another complication occurs when several input models are observed simultaneously in the simulation (such as correlated processing times at different workstations for an arriving job). Deveroye (1986) presented methods for generating random vectors from multivariate distributions such as multivariate normal distribution. Johnson (1987) developed a multivariate extension to the Johnson family. Cario and Nelson (1997) described a method to obtain random vectors with arbitrary marginal distributions and correlation matrix.

2.1.4 Subjective Input Models

There are several software packages that can be used to select probability distributions in the absence of data. The INSIGHT simulation environment (Roberts, 1983) displays standard distributions given user-specified parameters. DeBrota et al. (1989) described the VISIFIT software for matching a Johnson bounded distribution to subjective information. VIBES (AbouRizk et al., 1991) is another software designed for

the subjective estimation of generalized beta distributions. Finally, PRIME (Wagner and Wilson, 1996) offers an excellent visual interactive capability to model quite flexible input processes from expert information.

2.1.5 Limitations

In a given simulation study, once we identify reasonable probability distributions on the input processes and estimate their parameters, we proceed by generating random variates from these distributions using different random number streams either within one simulation run or over multiple runs. Inference is then made conditional on these models and their parameters for the output quantities of interest to the simulation analyst. We usually focus on estimating the mean response of the output with some measures of stochastic variability due to the internal generation of random numbers within the simulation model.

Simple models may fail because they cannot provide an adequate fit to many characteristics of the observed or subjective data, but complex models with all the recent advances taken into consideration may also fail to give reasonable predictive inference on the output of interest because of the following complications:

- Anomalies can occur in the use of classical goodness-of-fit tests which are based upon asymptotic approximations. In small samples, these tests can have very low power to detect lack of fit between the empirical distribution and each candidate distribution. In large samples, practically insignificant discrepancies between the empirical and candidate distributions often appear to be statistically significant, leading to the rejection of apparently satisfactory distributions. A dramatic example of this was discussed by Raftery (1986).

- In many situations more than one model provides a good fit to the data, but each one leads to substantially different inference on the quantity of interest. For example, Draper (1995) describes an application in oil industry where 13 forecasting models are appropriate for making predictions of oil prices, but each model gives completely different forecasts.

- Simulation analysts generally seek to estimate the unknown parameters based only on the observed data for the adopted input model. Subjective information

on the input processes, which is often available in practice, is not included formally in the model formulation.

- Conditioning on a single model ignores model uncertainty and fixing the model parameters ignores parameter uncertainty. Kass and Raftery (1995) provide examples where model and parameter uncertainty can be a big part of the overall uncertainty about the quantities of interest.

We propose in the next section the Bayesian model selection as an alternative approach for selecting valid input distributions. The Bayesian formalism, although computationally more intensive, provides solutions to most of the above difficulties.

2.2 Bayesian Model Selection

The issue of model selection can be addressed in two components: model adequacy and model choice (Gelfand, 1996). The first component checks if any proposed model is adequate in the presence of data or any other sources of information on the input process. The second component is a harder task which involves the selection of the best model within a collection of adequate models.

2.2.1 Model Adequacy

The most basic technique for checking the fit of a model to data, without requiring any more substantive information than is in the existing data and model, is to compare the data to the posterior predictive distribution (Gelman et al., 1995). Analogous to the classical P-value, we describe in this section the Bayes P-value as a measure for the statistical significance of the lack-of-fit.

If desired, further diagnostic tools can still be used such as cross-validation predictive distributions (Gelfand, 1996) and hypothesis-testing procedures (Bernardo and Smith, 1994). However, we do not recommend discrete-event simulation practitioners to use them routinely because they are computationally more intensive and may require further modeling assumptions.

Notation. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be the observed data and θ be the vector of parameters. To avoid confusion with the observed data $\boldsymbol{x}$, we define $\tilde{\boldsymbol{x}}$ as the replicated data that could have been observed. This represents the data we would have seen

tomorrow if the experiment that produced $\boldsymbol{x}$ today were replicated with the same model and the same value of θ that produced the observed data. We will work with the distribution of $\tilde{\boldsymbol{x}}$ given the current state of knowledge, that is, with the posterior predictive distribution,

$$p(\tilde{\boldsymbol{x}}|\boldsymbol{x}) \;=\; \int p(\tilde{\boldsymbol{x}}|\boldsymbol{x},\theta)p(\theta|\boldsymbol{x})d\theta. \tag{2.1}$$

Test quantities. We measure the discrepancy between model and data by defining test quantities, the aspects of the data we wish to check. A test quantity, or discrepancy measure, $T(\boldsymbol{x},\theta)$, is a scalar summary of parameters and data that is used as a standard when comparing observed data to predictive observations. Test quantities play the role in Bayesian model checking that test statistics play in classical testing. We use the notation $T(\boldsymbol{x})$ for a test statistic, which is a test quantity that depends only on the data.

Ideally, the test quantities will be chosen to reflect aspects of the model that are relevant to the scientific purposes to which the inference will be applied. They are often chosen to measure a feature of the data not directly addressed by the probability model such as ranks of the sample or correlation of model residuals. However, omnibus measures are useful for routine checks of fit. One general goodness-of-fit measure is the χ^2 discrepancy quantity, written here in terms of univariate data:

$$\chi^2 \text{ discrepancy: } T(\boldsymbol{x},\theta) \;=\; \sum_i \frac{(x_i - E(X_i|\theta))^2}{\text{var}(X_i|\theta)}, \tag{2.2}$$

where the summation is over the sample of observations. When θ is known or estimated value plugged-in, this quantity resembles the classical χ^2 goodness-of-fit measure.

***P*-values.** Lack of fit of the data with the posterior predictive distribution can be measured by the tail-area probability, or P-value, of the test quantity, and usually computed using posterior simulations of $(\theta,\tilde{\boldsymbol{x}})$. We define the P-value, for the familiar classical test and then in the Bayesian context.

The classical P-value for the test statistic $T(\boldsymbol{x})$ is

$$\text{classical } P\text{-value} \;=\; \Pr(T(\tilde{\boldsymbol{x}}) \geq T(\boldsymbol{x})|\theta), \tag{2.3}$$

where the probability is taken over the distribution of $\tilde{\boldsymbol{x}}$ with θ fixed. The classical P-value is in general a function of θ. A plug-in point estimate for θ such as the maximum likelihood estimate is often used to compute the P-value.

In the Bayesian approach, test quantities can be functions of the unknown parameters as well as data because the test quantity is evaluated over draws from the posterior distribution of the unknown parameters. The P-value is defined as the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity:

$$\begin{aligned} \text{Bayes } P\text{-value} &= \Pr\left(T(\tilde{\boldsymbol{x}},\theta) \geq T(\boldsymbol{x},\theta)|\boldsymbol{x}\right) \\ &= \int\int I_{\{T(\tilde{\boldsymbol{x}},\theta)\geq T(\boldsymbol{x},\theta)\}}\, p(\theta|\boldsymbol{x})\, p(\tilde{\boldsymbol{x}}|\theta)\, d\theta\, d\tilde{\boldsymbol{x}}, \end{aligned} \tag{2.4}$$

where I is the indicator function. In this formula, we have used the conditional independence property of the predictive distribution that $p(\tilde{\boldsymbol{x}}|\theta,\boldsymbol{x}) = p(\tilde{\boldsymbol{x}}|\theta)$.

In practice, we usually compute the posterior predictive distribution using simulation. If we already have L simulation draws from the posterior density of θ, we just draw $\tilde{\boldsymbol{x}}$ from the predictive distribution for each simulated θ; we now have L draws from the joint posterior density $p(\tilde{\boldsymbol{x}},\theta|\boldsymbol{x})$. The posterior predictive check is the comparison between the realized test quantities $T(\boldsymbol{x},\theta^l)$, and the predictive test quantities $T(\tilde{\boldsymbol{x}}^l,\theta^l)$. The estimated P-value is just the proportion of these L draws for which the test quantity equals or exceeds its realized value; that is, for which $T(\tilde{\boldsymbol{x}}^l,\theta^l) \geq T(\boldsymbol{x},\theta^l)$, $l = 1,\ldots,L$.

Interpretation of P-values. A model is suspected if the tail-area probability for some meaningful test quantity is close to 0 or 1. Major failures of the model, typically corresponding to extreme tail-area probabilities (less than 0.01 or more than 0.99), can be addressed by expanding the model in an appropriate way. Lesser failures might also suggest model improvements or might be ignored in the short term if the failure appears not to affect the main inferences. We will often evaluate a model with respect to several test quantities.

It is important not to interpret P-values as numerical evidence. For example, a P-value of 0.00001 is virtually no stronger, in practice, than 0.001; in either case, the aspect of the data measured by the test quantity is inconsistent with the model. A slight improvement in the model could bring either P-value to a reasonable range. The P-value measures statistical significance not practical significance.

2.2.2 Model Choice

Suppose we have a finite set $\mathcal{M} = \{M_1, \ldots, M_K\}$ of adequate candidate models that describes the observed data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Here we assume that the set $\mathcal{M}$ is indexed discretely on all the plausible models which can be nested or nonnested. In some applications, which occur rarely in simulation models, $\mathcal{M}$ can be indexed continuously. See Draper (1995) for a discussion on continuous model expansion. Let

$$p(M_k), \quad k = 1, \ldots, K \quad \text{and} \quad \textstyle\sum_k p(M_k) = 1,$$

denote the prior probability that M_k is the true model.

Assume that for each M_k the model specification for the data is

$$p(\boldsymbol{x}, \theta_k | M_k) \;\; = \;\; p(\boldsymbol{x} | M_k, \theta_k)\, p(\theta_k | M_k),$$

where θ_k represents the set of parameters specified by M_k, $p(\boldsymbol{x}|M_k, \theta_k)$ is the likelihood, and $p(\theta_k|M_k)$ is a proper prior distribution on θ_k which reflects the modeler's lack of knowledge on its true value.

Now if we want to select the single most probable model of the several entertained in $\mathcal{M}$, we would select M_{k^*}, such that

$$\max_k p(M_k|\boldsymbol{x}) \;\; = \;\; p(M_{k^*}|\boldsymbol{x}), \tag{2.5}$$

where

$$p(M_k|\boldsymbol{x}) \;\; = \;\; \frac{p(M_k)\, p(\boldsymbol{x}|M_k)}{\sum_{j=1}^{K} p(M_j)\, p(\boldsymbol{x}|M_j)}, \tag{2.6}$$

is the posterior probability of model M_k, and

$$p(\boldsymbol{x}|M_k) \;\; = \;\; \int p(\boldsymbol{x}|M_k, \theta_k)\, p(\theta_k|M_k)\, d\theta_k, \tag{2.7}$$

is the marginal likelihood or prior predictive density of model M_k.

Unless there is a sufficient penalty for using more than one model, selecting M_{k^*} for predicting a future observation does not aknowledge model uncertainty, and can lead to poor miscalibrated predictions. This is easily seen since the unconditional predictive density is

$$p(x_{n+1}|\boldsymbol{x}) \;\; = \;\; \sum_{k=1}^{K} p(M_k|\boldsymbol{x})\, p(x_{n+1}|\boldsymbol{x}, M_k), \tag{2.8}$$

where

$$p(x_{n+1}|\boldsymbol{x}, M_k) \;=\; \int p(x_{n+1}|\boldsymbol{x}, M_k, \theta_k)\, p(\theta_k|\boldsymbol{x}, M_k)\, d\theta_k, \tag{2.9}$$

is the predictive density under model M_k, and $p(\theta_k|\boldsymbol{x}, M_k)$ is the posterior distribution for θ_k given model M_k and data $\boldsymbol{x}$. This can be calculated as

$$p(\theta_k|\boldsymbol{x}, M_k) \;=\; \frac{p(\boldsymbol{x}|M_k, \theta_k)\, p(\theta_k|M_k)}{p(\boldsymbol{x}|M_k)}. \tag{2.10}$$

Equation (2.8) is an average of the posterior predictive distributions under each of the models considered, weighted by their posterior model probabilities. This approach, known in the statistical literature as Bayesian Model Averaging (BMA), acknowledge both aspects of uncertainty: model and parameter uncertainty. Madigan and Raftery (1994) note that averaging over all the models in this fashion provides better average predictive ability as measured by the logarithmic scoring rule (Good, 1952), than using any single model M_k. Considerable empirical evidence now exists to support this theoretical claim (see Draper (1995) and Raftery et al. (1996)). In the next section, we will give a detailed description of the BMA approach and discuss its implementation details.

2.3 Bayesian Model Averaging

2.3.1 Historical Perspective

In the statistical literature, early work related to model averaging includes Roberts (1965) who suggested a distribution which combines the opinion of two experts. This distribution, essentially a weighted average of posterior distributions of two models, is similar to BMA. Leamer (1978) expanded on this idea and presented the basic paradigm for BMA. He also pointed out the fundamental idea that BMA accounts for the uncertainty involved in selecting the model. The drawbacks of ignoring model uncertainty were recognized by many authors (see Dijkstra (1988) for a review), but little progress was made until new theoretical developments and computational power enabled researchers to overcome the difficulties related to implementing BMA (Hoeting et al., 1999). George (1999) reviews Bayesian model selection and discusses BMA in the context of decision theory. Draper (1995), Kass and Raftery (1995) all review BMA and the costs of ignoring model uncertainty.

Although, the BMA approach is not well-known in the simulation community, Cooke (1994) and Scott (1996) investigated the effect of structural uncertainty in simulating environmental systems. They created test scenarios in which they presented the same amount of information to each expert. The simulation results based on each expert opinion were compared against each other and against experimental data. The results showed variation amongst the predictions, and differences between the model predictions and experimental data. Chick (1997, 1999) outlined the basic methodology for implementing the BMA approach in discrete-event simulation. He proposed a Simulation Replication Algorithm for selecting input distributions parameters for simulation replications. We argue in the next section that Chick's algorithm suffers from some theoretical and practical deficiencies which motivated our research in later chapters.

In this section we discuss the basic idea and general implementation issues of Bayesian model averaging as described in the statistical literature. A major part of our research will focus on how to use the BMA approach in a discrete-event simulation experiment to account for model uncertainty, parameter uncertainty, as well as the usual stochastic uncertainty (see Chapter 4).

2.3.2 Basic Formulation

Let y denote the unknown quantity of interest such as a future observation (see Section 2.2.2) or a utility of a course of action. The desire is usually to express uncertainty about y in the light of the data $\boldsymbol{x}$. If $\mathcal{M} = \{M_1, \ldots, M_k\}$ denotes the set of all models considered, then the posterior distribution of y given the data $\boldsymbol{x}$, is

$$p(y|\boldsymbol{x}) = \sum_{k=1}^{K} p(M_k|\boldsymbol{x})\, p(y|\boldsymbol{x}, M_k) \tag{2.11}$$

$$= \sum_{k=1}^{K} p(M_k|\boldsymbol{x}) \int p(y|\boldsymbol{x}, M_k, \theta_k)\, p(\theta_k|\boldsymbol{x}, M_k)\, d\theta_k. \tag{2.12}$$

While in theory BMA represents an attractive solution to the problem of accounting for model and parameter uncertainty, it is not yet part of the standard statistical tool kit. This is, in part, due to the fact that implementation of BMA requires several elements to be specified or computed, each of which can present a difficulty. We discuss these elements in detail in the subsections that follow. They are as follows:

- The number of models in $\mathcal{M}$ can be enormous, with some models explaining the data far less than others. A practical solution is discussed (Section 2.3.3).
- The prior plausibility of each model and its parameters prior distributions should be specified. A brief literature review on a ongoing current research topic is presented (Section 2.3.4).
- The posterior model probabilities $p(M_k|\boldsymbol{x})$ should be computed. Several good approximations and some sampling-based methods are discussed (Section 2.3.5).
- The posterior predictive distribution of the quantity of interest $p(y|\boldsymbol{x}, M_k)$ in (2.11) can be unknown or formidable to evaluate (Section 2.3.6).

2.3.3 Occam's Window

The first obstacle to the BMA approach is that the size of interesting model classes often renders the exhaustive summation of Equation (2.11) impractical. Madigan and Raftery (1994) suggest the Occam's Window as a method that eliminates models that predict the data far less well than other competing models.

Two basic principles underlie this method. First, if a model predicts the data far less well than the model which provides the best predictions, then it should no longer be considered. Thus models not belonging to:

$$\mathcal{A}' = \left\{ M_k \in \mathcal{M} : \frac{\max_\ell \{p(M_\ell|\boldsymbol{x})\}}{p(M_k|\boldsymbol{x})} \leq C \right\}; \tag{2.13}$$

should be excluded from Equation (2.11), where C is chosen by the data analyst. A common choice is $C = 20$, by analogy with the popular 0.05 cutoff for P-values.

The second, optional, principle excludes complex models which receive less support from the data than their simpler counterparts. That is, if model M_l is nested within M_k and has a higher posterior probability, then model M_l is preferred. More formally, Madigan and Raftery (1994) also exclude from (2.11) models belonging to:

$$\mathcal{B} = \left\{ M_k \in \mathcal{M} : \exists M_l \in \mathcal{A}', M_l \subset M_k, \frac{p(M_l|\boldsymbol{x})}{p(M_k|\boldsymbol{x})} > 1 \right\}; \tag{2.14}$$

and Equation (2.11) is replaced by

$$p(y|\boldsymbol{x}) = \sum_{M_k \in \mathcal{A}} p(y|\boldsymbol{x}, M_k) p(M_k|\boldsymbol{x}), \tag{2.15}$$

where

$$\mathcal{A} \;=\; \mathcal{A}' \setminus \mathcal{B}.$$

This reduces considerably the number of models in the sum in Equation (2.11), and now all that is required is a search strategy to identify the models in $\mathcal{A}$. Madigan and Raftery (1994) provide a detailed description of the computational strategy for doing this. Other search methods are also suggested in the literature. The MC^3 approach (Madigan and York, 1995) uses a Markov chain Monte Carlo method to directly approximate (2.11). Another approach suggested by Volinsky et al. (1997) uses the "leaps and bounds" algorithm to rapidly identify models to be used in the summation of Equation (2.11).

2.3.4 Specification of Priors

The BMA approach requires that the model prior probabilities $\{p(M_k) : k = 1, \ldots, K\}$ and the prior distributions $\{p(\theta_k|M_k) : k = 1, \ldots, K\}$ on the parameters of each model must be specified. Historically, a major impediment to widespread use of the Bayesian paradigm has been that determination of the appropriate form of the priors is often an arduous task. We describe briefly here several methods to elicit prior distributions (see Berger (1985) for an excellent overview).

First, the specification of prior probabilities $p(M_k)$ will be typically context specific. When there is little prior information about the relative plausibility of the models considered, taking them all to be equally likely is a reasonable choice. Madigan and Raftery (1994) have found no perverse effects from putting a uniform prior over the models. Different prior probabilities can be viewed as derived from previous data and representing the relative success of the models in predicting those previous data. This may apply even when prior model probabilities represent apparently subjective expert opinion (Madigan and York, 1995).

Second, the easiest way to deal with the problem of specifying prior distributions on the model parameters is to ignore them and simply use the Schwarz criterion (see Section 2.3.5). Although this will lead to appropriate conclusions in "sufficiently large" samples, there is not much available guidance on the operational meaning of the qualifying phrase "sufficiently large." Typically, these distributions are specified based on information accumulated from past studies, or from expert opinions. But

once this is done, an important issue is the sensitivity of the predictive distributions to the choices of the priors. In order to simplify the subsequent computational burden, experimenters often limit this choice somewhat by restricting priors to some familiar distributional family. An even simpler alternative, available in some cases, is to endow the prior distribution with little information content, so that the data from the current study will be the dominant factor in determining the posterior distribution. We address each of these approaches in the subsequent sections.

Informative Priors

In the univariate case, the simplest approach to specifying a prior distribution $p(\theta_k|M_k)$ is first to limit consideration to a manageable collection of θ_k values deemed possible, and subsequently to assign probability masses to these values, reflecting the experimenter's prior beliefs as closely as possible, in such a way that their sum is 1. If θ_k is discrete-valued, such an approach may be quite natural, though perhaps quite time consuming. If θ_k is continuous, we must instead assign the masses to intervals on the real line, rather than to single points, resulting in a histogram prior for θ_k. Such a histogram may seem inappropriate, especially in concert with a continuous likelihood, but in fact may be more appropriate if the integrals required to compute the posterior distribution must be evaluated by a numerical quadrature scheme. Moreover, a histogram prior may have as many bins (classes, cells) as the patience of the elicitee and the accuracy of his prior opinion will allow. It is important, however, that the range of the histogram should be sufficiently wide, since the support of the posterior will necessarily be a subset of that of the prior.

Alternatively, we might simply assume that the prior for θ_k belongs to a parametric distributional family $p(\theta_k|\nu_k)$, choosing ν_k so that the result matches the elicitee's true prior beliefs as nearly as possible. For example, if ν_k is two-dimensional then specification of two moments (say, the mean and the variance) or two quantiles (say, the 50th and 95th quantiles) would be enough to determine its exact value. This matching approach (Berger, 1985) limits the effort required of the elicitee, and also overcomes the finite support problem inherent in the histogram approach. It may also lead to simplifications in the posterior computation.

A limitation of this approach is of course that it may not be possible for the elicitee to describe his or her prior beliefs into any of the standard parametric forms. In addition two distributions which look virtually identical may in fact have quite dif-

ferent properties. For example, Berger (1985, p. 79) points out that the Cauchy(0,1) and Normal(0,2.19) distributions have identical 25th, 50th, and 75th percentiles (-1, 0, and 1 respectively) and density functions that appear very similar when plotted, yet may lead to quite different posterior distributions.

Even when scientifically relevant prior information is available, elicitation of the precise forms for the prior distribution from experimenters can be a long and tedious process. However, prior elicitation issues tend to be application-specific, meaning that general-purpose algorithms are typically unavailable. Chaloner (1996) provides overviews of the various philosophies of prior distribution elicitation. The difficulty of prior elicitation has been ameliorated somewhat through the addition of interactive computing, especially dynamic graphics and object-oriented computer languages.

Conjugate Priors

In choosing a prior belonging to a specific distributional family, some choices may be more convenient computationally than others. In particular, it may be possible to select a member of that family which is conjugate to the likelihood—that is, one which leads to a posterior distribution belonging to the same distributional family as the prior. Morris (1983) showed that regular exponential families, from which we typically draw our likelihood functions, do in fact have conjugate priors, so that this approach will often be available in practice.

For multiparameter models, independent conjugate priors may often be specified for each parameter, leading to corresponding conjugate forms for each conditional posterior distribution. The ability of conjugate priors to produce at least unidimensional conditional posteriors in closed form enables them to retain their importance even in high-dimensional settings. This occurs through the use of Markov Chain Monte Carlo (MCMC) integration techniques (Section 2.3.5), which construct a sample from the joint posterior by successively sampling from the individual conditional posterior distributions.

Finally, while a single conjugate prior may be inadequate to reflect available prior knowledge accurately, a finite mixture of conjugate priors may be sufficiently flexible while still enabling simplified posterior calculations.

Noninformative Priors

Often no reliable prior information concerning θ_k exists, or an inference based mainly

on data is desired. Suppose we could find a distribution that contains no information about θ_k, in the sense that it does not favor one θ_k over another (provided both values were logically possible). We might refer to such a distribution as a noninformative prior for θ_k, and argue that all of the information resulting in the posterior distribution must arise from the data—and hence all resulting inferences must be completely objective, rather than subjective. Such an approach is likely to be important if Bayesian methods are to compete successfully in practice with their popular frequentist counterparts such as maximum likelihood estimation. Kass and Wasserman (1996) provides an excellent review of noninformative priors.

Simplifications involving priors should be considered carefully, because they may affect the results and yet may not be justified. Here model selection or testing is different from estimation. In frequentist theory, estimation and testing are complementary, but in the Bayesian approach the problems are completely different. Kass and Raftery (1995) discuss the main differences and highlight the problems of using improper priors in model selection. Flat priors are defined only up to undefined multiplicative constants. Thus the marginal distributions in (2.7) also contain undefined constants, which forces some of the posterior model probabilities in (2.6) to become zero.

One solution to this problem is to set aside part of the data to use as a training sample which is combined with the improper prior to produce a proper prior distribution. The marginal distributions are then computed using the remainder of the data. This idea was introduced by Lempers (1971), and other implementations have been suggested more recently under the names partial Bayes factors (O'Hagan, 1991), intrinsic Bayes factors (Berger and Perrichi, 1996), and fractional Bayes factors (O'Hagan, 1995).

Another solution to this problem is to use the cross-validation predictive densities (Gelfand, 1996) to compute the posterior model probabilities. These densities usually exist even if the prior predictive density does not. This approach is extremely important if MCMC methods are used to compute the posterior probabilities under improper priors.

Sensitivity to Prior Distributions

Posterior distributions can be sensitive to the prior specification. Setting-up a supermodel to include all possibilities and substantive knowledge is both conceptually im-

possible and computationally infeasible. It is thus necessary to examine how sensitive the resulting posterior distributions are to arbitrary specifications. One approach is to evaluate the posterior distributions over several classes of priors. This, however, makes the issue of computation more urgent, because many integrals (often multi-dimensional) must be computed. An important computational device is to use the Laplace approximation (Section 2.3.5) for quick comparisons. When there is enough information to yield initial priors with given hyperparameters, perturbation of the hyperparameters (e.g. by doubling or halving) may give a good insight on the posterior computations. Several other context-specific approaches for sensitivity analysis are discussed in the literature (see for example Kass and Vaidyanathan (1992), and Kass and Raftery (1995)).

2.3.5 Computing Posterior Model Probabilities

The posterior model probabilities $p(M_k|\boldsymbol{x})$ are computed as follows:

$$p(M_k|\boldsymbol{x}) \;=\; \frac{p(M_k)\,p(\boldsymbol{x}|M_k)}{\sum_{j=1}^{K} p(M_j)\,p(\boldsymbol{x}|M_j)},\ \text{for } k=1,\ldots,K. \tag{2.16}$$

The evaluation of these probabilities comes down to computing the marginal or prior predictive density given the model

$$p(\boldsymbol{x}|M_k) \;=\; \int p(\boldsymbol{x}|M_k,\theta_k)\,p(\theta_k|M_k)\,d\theta_k,\ k=1,\ldots,K. \tag{2.17}$$

The integral in (2.17) may be evaluated analytically for distributions in the regular exponential family with conjugate priors. However, (2.17) is generally intractable and thus must be computed by numerical methods. An excellent review for the various numerical integration strategies is provided by Kass and Raftery (1995). Here, we provide a brief presentation of these methods.

Asymptotic Approximations

A useful approximation to the marginal density of the data as given by (2.17) is the Laplace method of approximation (Tierney and Kadane, 1986). It is obtained by assuming that the posterior density $p(\theta_k|\boldsymbol{x},M_k)$, which is proportional to $p(\boldsymbol{x}|M_k,\theta_k)\times p(\theta_k|M_k)$, is highly peaked about its maximum $\tilde{\theta}_k$, which is the posterior mode. This is usually the case if the likelihood function $p(\boldsymbol{x}|M_k,\theta_k)$ is highly peaked near its maximum $\widehat{\theta}_k^{\text{mle}}$, which is the case for large samples. Let $\tilde{l}_k(\theta_k)=\ln[p(\boldsymbol{x}|M_k,\theta_k)p(\theta_k|M_k)]$.

Expanding $\tilde{l}_k(\theta_k)$ as a quadratic about $\tilde{\theta}_k$ and then exponentiating yields an approximation to $p(\boldsymbol{x}|M_k, \theta_k) \times p(\theta_k|M_k)$ that has the form of a normal density with mean $\tilde{\theta}_k$ and covariance matrix $\tilde{\Sigma}_k$ specified by

$$\left(\tilde{\Sigma}_k^{-1}\right)_{ij} = -\frac{\partial^2 \tilde{l}_k(\theta_k)}{\partial\theta_{ki}\partial\theta_{kj}}\bigg|_{\tilde{\theta}_k}.$$

Integrating this approximation gives the logarithm of the marginal density in (2.17) as follows

$$\begin{aligned}\ln[p(\boldsymbol{x}|M_k)] &= \frac{1}{2}d_k\ln(2\pi) + \frac{1}{2}\ln(|\tilde{\Sigma}_k|) + \ln[p(\boldsymbol{x}|M_k, \tilde{\theta}_k)] \\ &\quad + \ln[p(\tilde{\theta}_k|M_k)] + O(n^{-1}),\end{aligned} \tag{2.18}$$

where d_k is the dimension of θ_k, and n is the sample size of the data set $\boldsymbol{x}$.

An important variant of Laplace approximation is given by

$$\begin{aligned}\ln[p(\boldsymbol{x}|M_k)] &= \frac{1}{2}d_k\ln(2\pi) + \frac{1}{2}\ln(|\widehat{\Sigma}_k|) + \ln[p(\boldsymbol{x}|M_k, \widehat{\theta}_k^{\text{mle}})] \\ &\quad + \ln[p(\widehat{\theta}_k^{\text{mle}}|M_k)] + O(n^{-1}),\end{aligned} \tag{2.19}$$

where $\widehat{\Sigma}_k$ is the inverse of the observed information matrix

$$(\mathcal{I}_k)_{ij} = -\frac{\partial^2}{\partial\theta_{ki}\partial\theta_{kj}}\ln[p(\boldsymbol{x}|M_k, \theta_k)]\bigg|_{\hat{\theta}_k^{\text{mle}}},$$

evaluated at the Maximum Likelihood Estimator (MLE)

$$\widehat{\theta}_k^{\text{mle}} \equiv \arg\max_{\theta_k} p(\boldsymbol{x}|M_k, \theta_k).$$

Although this approximation is likely to be less accurate than the first one when the prior is informative relative to the likelihood, it has the advantage of being easily computed from any statistical package.

Finally, it is possible to avoid the introduction of the prior densities $p(\theta_k|M_k)$ by using a simpler approximation (Schwarz, 1978) given by

$$\ln[p(\boldsymbol{x}|M_k)] = -\frac{1}{2}d_k\ln(n) + \ln[p(\boldsymbol{x}|M_k, \widehat{\theta}_k^{\text{mle}})] + O(1), \tag{2.20}$$

where the term $-\frac{1}{2}d_k\ln(n)$ can be thought of as a penalty for models with a large number of parameters. This approximation is less accurate than the Laplace approximation especially in small samples.

Simple Monte Carlo, Importance Sampling and Gaussian Quadrature

The simplest Monte Carlo integration estimate of $p(\boldsymbol{x}|M_k)$ is

$$\hat{p}_1(\boldsymbol{x}|M_k) = \frac{1}{L}\sum_{l=1}^{L} p(\boldsymbol{x}|M_k, \theta_k^{(l)}), \tag{2.21}$$

where $\{\theta_k^{(l)} : l = 1, \ldots, L\}$ is a sample from the prior distribution. A major difficulty with $\hat{p}_1(\boldsymbol{x}|M_k)$ is that most of the $\theta_k^{(l)}$ have small likelihood values if the posterior is concentrated relative to the prior, so that the simulation will be quite inefficient.

The precision of simple Monte Carlo integration can be improved by importance sampling. This consists of generating a sample $\{\theta_k^{(l)} : l = 1, \ldots, L\}$ from a density $\pi^\star(\theta_k|M_k)$, known as the *importance sampling function*. Under quite general conditions, a simulation consistent-based estimator of $p(\boldsymbol{x}|M_k)$ is

$$\hat{p}_2(\boldsymbol{x}|M_k) = \frac{\sum_{l=1}^{L} w_l p(\boldsymbol{x}|M_k, \theta_k^{(l)})}{L}, \tag{2.22}$$

where $w_l = p(\theta_k^{(l)}|M_k)/\pi^\star(\theta_k^{(l)}|M_k)$.

A more efficient scheme is based on adaptive Gaussian quadrature. Using well-established methods from the numerical analysis literature, Genz and Kass (1993) showed how integrals that are peaked around a dominant mode may be evaluated.

Markov Chain Monte Carlo (MCMC) Methods

Several methods are now available for simulating from posterior distributions. In the simplest case these include direct simulation and rejection sampling. In more complex cases, MCMC methods, particularly the Metropolis-Hastings algorithm and the Gibbs sampler provide a general recipe. Another fairly general recipe is the weighted likelihood bootstrap (Newton and Raftery, 1994). Any of these methods gives us a sample approximately drawn from the posterior density $cq(\theta_k|\boldsymbol{x}, M_k)$. Then we can estimate (2.17) by

$$\hat{p}_3(\boldsymbol{x}|M_k) = \frac{\frac{1}{L}\sum_{l=1}^{L} w_l p(\boldsymbol{x}|M_k, \theta_k^{(l)})}{\frac{1}{L}\sum_{l=1}^{L} w_l}, \tag{2.23}$$

where $w_l = p(\theta_k^{(l)}|M_k)/q(\theta_k^{(l)}|\boldsymbol{x}, M_k)$, and $\sum_{l=1}^{L} w_l/L$ is a simulation-consistent estimator of c. Now, substituting $cq(\theta_k|\boldsymbol{x}, M_k) = p(\boldsymbol{x}|M_k, \theta_k)p(\theta_k|M_k)/p(\boldsymbol{x}|M_k)$ into the

above equation yields an estimate for $p(\boldsymbol{x}|M_k)$,

$$\hat{p}_4(\boldsymbol{x}|M_k) \;=\; \left\{ \frac{1}{L} \sum_{l=1}^{L} \left[p(\boldsymbol{x}|M_k, \theta_k^{(l)}) \right]^{-1} \right\}^{-1}, \tag{2.24}$$

the harmonic mean of the likelihood values. This converges almost surely to the correct value, $p(\boldsymbol{x}|M_k)$, as $L \to \infty$, but it does not generally satisfy a Gaussian Central Limit theorem(CLT). Simple modifications to (2.24) that satisfy the CLT are suggested in the literature by Meng and Weng (1993), Gelfand and Dey (1994), and Newton and Raftery (1994).

Finally, another simple estimator that performed well practice is the so called "Laplace-Metropolis" estimator of $p(\boldsymbol{x}|M_k)$, by Raftery (1996). It is obtained by using the posterior simulation output to estimate the quantities needed to compute the Laplace approximation (2.18), namely the posterior mode, $\tilde{\theta}_k$, and minus the inverse Hessian at the posterior mode, $\tilde{\Sigma}_k$.

2.3.6 Computing Posterior Predictive Distributions

The most important ingredient to the BMA formulation is the set of posterior predictive distributions of the unknown quantity of interest y given model M_k and data $\boldsymbol{x}$. This is given as

$$p(y|\boldsymbol{x}, M_k) \;=\; \int p(y|\boldsymbol{x}, M_k, \theta_k) p(\theta_k|\boldsymbol{x}, M_k) d\theta_k, \tag{2.25}$$

where $p(y|\boldsymbol{x}, M_k, \theta_k) = p(y|M_k, \theta_k)$ if the quantity of interest y is generated independently conditional on each model and its vector of parameters. Equation (2.25) creates no new computational burden, since we would have to compute it anyway as part of our parameter sensitivity analysis. However, the computation depends on whether the conditional predictive distribution $p(y|\boldsymbol{x}, M_k, \theta_k)$ is known or not. This distinction creates the main difference between the BMA applications proposed so far in the statistical literature and its applicability to discrete-event simulation.

Closed form expressions for (2.25) exist in many important statistical applications such as normal linear models (Zellner, 1971), and approximations based on Monte Carlo integration (Geweke, 1989) are also available. For large sample size n the simple approximation

$$p(y|\boldsymbol{x}, M_k) \;=\; p(y|\boldsymbol{x}, M_k, \hat{\theta}_k), \tag{2.26}$$

where $\widehat{\theta}_k^{\text{mle}}$ is the maximum likelihood estimate of θ_k under model M_k, may be sufficiently precise (Taplin, 1993). Equation (2.26) provided an excellent approximation for later applications by Draper (1995) and Raftery et al. (1996). For example, heuristic calculations by Raftery et al. (1996) suggest that in the regression variable selection problem the prediction uncertainty is $O(1)$, parameter uncertainty is $O(n^{-1})$, and model uncertainty is $O(dn^{-1})$, where d is the number of candidate independent variables. It would be reasonable here to ignore parameter uncertainty while taking account of model uncertainty when d is large.

In discrete-event simulation, the relationship between the unknown quantity of interest y and the input parameters is unknown and probably very complicated if it were known, since we are going to the trouble of simulating instead of plugging numbers into some formula. We will show in the next two chapters that the simulation output y is an unknown, stochastic and probably very messy function of the input parameters given a certain model.

2.4 Bayesian Techniques for Simulation

The literature on the application of Bayesian techniques in the field of discrete-event simulation is not extensive. Andrews and Schriber (1983) appear to be the first to discuss modeling steady state simulation output with a Bayesian formalism. They regarded the simulation output from autocorrelated batches as a stationary Gaussian process having a Gaussian prior for the mean. They later used this Bayesian approach to develop a point estimator and construct a credible interval for the mean of batch-run simulation output.

Glynn (1986) described a general framework for modeling a generalized semi-Markov process when the parameters of the input distributions are unknown. He considered the example of an $M/M/1$ queue where the service rate is unknown, and noted the dependence of the output distribution on the prior distribution of the input parameters. He also commented on potential research directions in the area. It took more than a decade for Glynn's ideas to receive some attention in the simulation community. Andradóttir and Bier (1997) discussed some possible roles of Bayesian analysis in model validation, and output analysis with normal and truncated normal distributions. They presented some results on importance sampling when the parameters of the input distributions are unknown. A number of analytical and practical

difficulties were described in their work.

Nelson et al. (1997) evaluated several techniques for combining a deterministic approximation with a stochastic simulation estimator, among them a Gaussian Bayesian analysis for a point estimator. Wang and Schmeiser (1997) formulated an optimization problem to select a prior distribution satisfying certain desirable properties. They also performed a Bayesian robustness analysis for analyzing Monte Carlo simulation output. Bayesian formulation based on the normality assumption of the output response were also used in the ranking and selection area of simulation. Inoue et al. (1999) presented a Bayesian formulation for describing the probability that a system is the best when there are two or more systems. They evaluated their procedure under a variety of scenarios including independent or common random numbers, and known or unknown variance response.

Chick (1999) addressed the problem of selecting probability models for input to stochastic simulations. His analysis lead to the Simulation Replication Algorithm given in Figure 2.1 to estimate the posterior outptut mean response $E(y|\boldsymbol{x})$. We sample an input model from the set of models $\mathcal{M}$, prior to each run or replication for a total of R replications. Given the sampled input models, we generate their vector of parameters from their posterior distributions. Finally, we run our simulation model at the sampled input models and their vector of parameters, to obtain R output responses.

for $r = 1, \ldots, R$
    sample model M^r from $p(M|\boldsymbol{x})$
    sample parameter $\theta^r_{M^r}$ from $p(\theta_{M^r}|\boldsymbol{x}, M^r)$
    run the simulation model at $(M^r, \theta^r_{M^r})$
    calculate the output response y_r
end loop
generate the estimate $\sum_{r=1}^{R} y_r/R$ as an estimate for $E(y|\boldsymbol{x})$

Figure 2.1: Chick's Simulation Replication Algorithm

Due to the following reasons, we believe that this approach has several theoretical and practical deficiencies that makes it of little use to simulation practitioners:

(a) The above algorithm calculates output responses based on a single replication at each randomly chosen input model and its parameter. This approach im-

plicitly ignores the effect of stochastic uncertainty on the performance measure of interest. For example, if the random variables M and θ_M were degenerate at M^* and $\theta^*_{M^*}$, respectively, then the algorithm would deliver just one output observation y_1 as an estimate for $E(y|\boldsymbol{x})$. Assessing the stochastic variability would also be impossible in this case.

(b) Some input models may have a very small posterior probability compared with others. If these models explain the data far less well than others, then they should be eliminated from further analysis (Occam's Window, Section 2.3.3). If they are adequate or they are kept for some other considerations, then a simple random sampling scheme will never allow the analyst to observe output responses from such models. This is the case in practice since we are generally limited by a moderate number of simulation runs.

(c) Chick's algorithm cannot accommodate more models without repeating all the simulation runs. If for some reason we decide to expand the set of models $\mathcal{M}$ to have more than K models, then we need to repeat all the runs with the new sampled models and their parameters to obtain a new estimate for the mean response.

(d) It is hard to quantify the percentage of the total variability due to each uncertainty factor using the above algorithm. This quantification is valuable to the analyst to improve the efficiency of the simulation design.

Our fundamental research task is to develop an approach that overcomes all the above difficulties, has a theoretical appeal, and most importantly performs well in practice. We divided our work in two chapters. In Chapter 3, we develop a Bayesian framework to account for parameter uncertainty as well as the usual stochastic uncertainty in a discrete-event simulation experiment. This approach will be compared to other approaches in the simulation literature and evaluated through Monte Carlo experiments. In Chapter 4, we extend our Bayesian framework to account for model uncertainty. The BMA idea and its implementation steps will prove useful to design a new Simulation Replication Algorithm that gives an excellent perfomance in estimating the mean response and assessing its variability.

Chapter 3

Accounting for Parameter Uncertainty in Simulation

Discrete-event simulations, especially those modeling complex systems, are almost all driven by random input processes. A simulation experiment therefore typically requires a number of streams of random variates drawn from specified distributions or input models. The inherent variation in the output of a simulation experiment arising from its dependence on these random inputs is often called *stochastic uncertainty* (Helton, 1998). We generally assume that the input models driving the simulation belong to known parametric families. However, uncertainty typically occurs when choosing between different models. We refer to this second source of variation as *model uncertainty* (Raftery et al., 1996). The parameters on which these models depend are usually assumed to be fixed. In practice, these parameters are estimated from subjective information (expert opinion) or from real data observed on the input processes. The estimation of unknown parameters gives rise to another source of variation often referred to in the literature as *parameter uncertainty* (Raftery et al., 1996).

Cheng and Holland (1997) consider two methods of assessing how the variation in the simulation output depends on two sources of variation: parameter uncertainty and stochastic uncertainty. The first is based on classical differential analysis, or the δ-method (Stuart and Ord, 1994). The main result of Cheng and Holland (1997) shows that under general conditions the total variation in the simulation output is composed of two distinct terms, depending respectively on the parameter uncertainty

and stochastic uncertainty. We present this result as well as the implementation issues concerning the δ-method in Section 3.2. One problem with this method is that certain sensitivity coefficients have to be estimated, and the effort needed to do this increases linearly with the number of unknown input parameters. Moreover, when the number of parameters is large, a problem can occur with spurious variation overinflating the variance estimate. Cheng and Holland (1998) consider two ways of modifying the δ-method so that much of the computing effort is concentrated on just two settings of the parameter values, irrespective of the actual number of unknown parameters. Such *two-point* methods, however, can perform very poorly in practice. We shall not discuss these methods further here. The second method that Cheng and Holland (1997) consider for assessing the variation in the simulation output is the parametric bootstrap (Efron and Tibshirani, 1993). Although computationally more expensive, this method does not suffer from the difficulties of the δ-method. It can also be more competitive on the grounds of computational efficiency if the number of unknown parameters is large. We describe this method in detail in Section 3.3.

Both of the above methods rely on the assumption that the parameters of the input models are unknown, but deterministic quantities. Moreover, the output inferences are implicitly conditional on the selected single input model. The objective of the simulation experiment is therefore to estimate the mean output response as a function of the "true" but unknown parameter values. The parameter uncertainty arises from estimating the true parameters using real observed data, and the stochastic uncertainty arises from the use of random variates generated from the selected input processes during the simulation experiment. The most fundamental problem with such approaches to input model selection is that conditional on a single input model and on given values of the parameters for that input model, the output inference underestimates the overall uncertainty in the output quantities of interest, sometimes to a dramatic extent (Kass and Raftery, 1995). Moreover, the usual approach to model selection in the simulation community is commonly guided by a series of goodness-of-fit tests (Law and Kelton, 2000). These tests can be highly misleading and very difficult to interpret in a classical statistical framework (Berger and Delampady, 1987).

All these difficulties can be avoided, if one adopts a Bayesian approach that incorporates prior information on competing models and their parameters in a rigorous manner. We can compute the posterior probabilities using Bayes' rule for all com-

peting models and their parameters; and then we can make a composite inference that takes account of model and parameter uncertainty in a formally justifiable way. Even if prior information is not readily available, there are methods to perform a full Bayesian analysis that rely on some uninformative priors and thus will give more weight to the observed data, but will still incorporate model and parameter uncertainty that is due to our lack of knowledge of the nature of the input processes driving our simulation experiment. These methods generalize the classical inferences conditional on the choice of a single input model and its parameters, and they work for both small and large sample sizes. These ideas are not new, but they were rejected for many decades because they are computationally quite expensive, if not impossible in some cases. However, with the recent development of Markov Chain Monte Carlo (MCMC) methods for computing marginal and posterior probabilities (Gilks et al., 1996), many statisticians have adopted the Bayesian approach to account for uncertainty in model selection in a broad diversity of application areas. Here we explore the use of Bayesian methods in selecting valid simulation input models and in designing simulation experiments to yield more reliable inferences based on simulation-generated outputs.

In this chapter, we use a Bayesian approach to model both stochastic uncertainty and parameter uncertainty in simulation; and we estimate the effects of these sources of uncertainty on the output quantity of interest. Hence, we assume in this chapter that the functional forms of all input models are known, perhaps based on our prior knowledge of the processes driving the simulation model—a situation that sometimes occurs in simulation applications. Another reason for fixing all input models is the computational cost associated with choosing more than one model. Finally, for some applications the assumption of no model uncertainty will also prove helpful for conducting Monte Carlo experimentation (Section 3.5) to compare the performance of the Bayesian methods with the classical and bootstrap methods. Section 3.4 gives the Bayesian framework for modeling parameter and stochastic uncertainties in discrete-event simulation. This leads to our "Bayesian Simulation Replication Algorithm" for designing simulation experiments. In the next chapter, we will extend our Bayesian framework to account also for the effects of input model uncertainty on the output quantities of interest.

3.1 The Simulation Experiment

A simulation experiment, in its basic form, consists of making m independent runs of the simulation model, and observing a single output performance measure of interest, y, from each run. Let L be the length of each simulation run measured in terms of simulation time or the number of observations of some fundamental simulation-generated output process from which the statistic y is computed on each run. For example in the jth run of a simulation of the $M/M/1$ queue, y_j might be the average delay of customers $\sum_{l=1}^{L} D_l/L$, where L is the run length representing the total number of customers, and the D_l's represent the recorded delay for each customer served during the simulation. For simplicity, we assume that the simulation model is driven by a single sequence $\{X_1, X_2, \ldots\}$ of independent and identically distributed input random variables, from which we observe the random sample $\boldsymbol{x} = (x_1, \ldots, x_n)$. Multiple independent random input sequences will be treated in a similar fashion.

During the jth simulation run, a stream of random numbers $\boldsymbol{u}_j = (u_{j1}, \ldots, u_{jT_j})$ is generated internally within the simulation model. The total number of random numbers, $\sum_{j=1}^{m} T_j$, generated during all m simulation runs constitute the entire stochastic uncertainty present in the simulation experiment. On the jth run the simulation random-number stream $\boldsymbol{u}_j$ is used to generate the input random variates $\tilde{\boldsymbol{x}}_j = (\tilde{x}_{j1}, \ldots, \tilde{x}_{jT'_j})$ by some transformation method, from which the output y_j is computed. Here T'_j represents the total number of input variates generated during run j. In principle, T_j and T'_j can be infinite, but it is more convenient to think of them as finite for a fixed simulation run length L.

One possible method for generating the input random variates $\tilde{\boldsymbol{x}}_j$, which is commonly used by the simulation software packages, is the inverse transform method. If we let u_{ji} denote the ith random number sampled on the jth simulation run, then $\tilde{x}_{ji}$ can be generated using the inverse transform method as

$$\tilde{x}_{ji} \;=\; G_M^{-1}(u_{ji}, \theta_M), \tag{3.1}$$

where $G_M^{-1}(\cdot)$ is the inverse of the distribution function $G_M(\cdot, \theta_M)$ of the simulation input model M, having θ_M as its d_M-dimensional vector of parameters. Given θ_M, we shall assume in this chapter that the conditional distributions of $\tilde{X}_{ji}$ and X_i are the same. Hence, the real data observations are assumed to have been drawn from the distribution $G_M(x, \theta_M)$.

The model and parameter uncertainties are represented by the random variables M and θ_M, respectively, both of which are assumed to depend only on the subjective information or data observed on the target input processes; and the stochastic uncertainty depends only on the randomness of $\boldsymbol{u}$. Thus the output of interest from the simulation run, y, can be regarded as an unknown complicated function of $\boldsymbol{u}$, M, and θ_M,

$$y = y(\boldsymbol{u}, M, \theta_M). \tag{3.2}$$

In this chapter, we focus on the effects of parameter uncertainty and stochastic uncertainty on the distribution of y. For simplicity we drop M from our subsequent expressions, recognizing that they are implicitly dependent on the input model M. (In the next chapter, we will relax the simplifying assumption that the input model M is known.) Thus equation (3.2) becomes

$$y = y(\boldsymbol{u}, \theta), \tag{3.3}$$

and we let

$$\eta(\theta) = \int y(\boldsymbol{u}, \theta)\, d\boldsymbol{u} \tag{3.4}$$

denote the expected value of y given θ.

The objective of a classical simulation experiment is generally to estimate $\eta(\theta_0)$, where θ_0 is the true but unknown parameter value, estimated separately from the simulation experiment using real data. It is also of interest to compute a measure of the variability of the simulation output, from which a confidence interval for $\eta(\theta_0)$ can be constructed. We consider how this can be done using the δ-method and the bootstrap method.

Our main objective in this research, however, is to estimate the mean response and assess its variability more reliably by adopting a Bayesian perspective. After modeling our uncertainty about the parameters of the input processes through prior probability distributions, and observing data on these processes, we derive in Section 3.4 an expression for the posterior mean response. We also develop a simulation replication algorithm for estimating the posterior mean response and for constructing a credible interval for the posterior mean response.

3.2 Classical Approach

From the structure of the simulation model described above, we see that the responses or outputs of the simulation runs for a fixed parameter θ can be written as

$$y_j = y(\boldsymbol{u}_j, \theta) \;\; = \;\; \eta(\theta) + e_j(\boldsymbol{u}_j, \theta), \;\; j = 1, \ldots, m. \tag{3.5}$$

The error variable e_j is the random difference between the output of the jth simulation run and $\eta(\theta)$. We generally assume that

$$E(e_j|\theta) = 0 \quad \text{and} \quad \text{Var}(e_j|\theta) = \tau^2(\theta) \;\; \text{for} \;\; j = 1, \ldots, m, \tag{3.6}$$

so that

$$E(y_j|\theta) \;\; = \;\; \eta(\theta). \tag{3.7}$$

Hence the mean of the simulation outputs,

$$\overline{y} \;\; = \;\; \frac{\sum_{j=1}^{m} y_j}{m}, \tag{3.8}$$

is an unbiased estimator of $\eta(\theta)$.

If the maximum likelihood estimator $\widehat{\theta}$ is used for θ in (3.5), then the output is

$$y_j = y(\boldsymbol{u}_j, \widehat{\theta}) \;\; = \;\; \eta(\widehat{\theta}) + e_j(\boldsymbol{u}_j, \widehat{\theta}), \;\; j = 1, \ldots, m, \tag{3.9}$$

where both $\widehat{\theta}$ and $\boldsymbol{u}_j$ are random; and in general we may assume that $\widehat{\theta}$ and $\boldsymbol{u}_j$ are stochastically independent. As $\widehat{\theta}$ is a random variable, var(y) is not τ^2, but decomposes into essentially two terms arising separately from the parameter and stochastic uncertainties. Let

$$\boldsymbol{g}(\theta) = \nabla\eta(\theta) \;\; = \;\; [\partial\eta(\theta)/\partial\theta_1, \ldots, \partial\eta(\theta)/\partial\theta_d], \tag{3.10}$$

denote the gradient of $\eta(\cdot)$ at $\theta \in \Re^d$; the components of $\boldsymbol{g}(\theta)$ are conventionally known as *sensitivity coefficients*. Cheng and Holland (1997) show that

$$\begin{aligned} \text{Var}(y) \;\; &\approx \;\; \boldsymbol{g}(\theta_0)^T \boldsymbol{V}(\theta_0) \boldsymbol{g}(\theta_0) + \tau^2(\theta_0), \\ &= \;\; V_{\text{par}} + V_{\text{sto}}, \end{aligned} \tag{3.11}$$

where: $\eta(\theta)$ is the function (3.4); V_{par} denotes the parameter variance; and V_{sto} denotes the stochastic or simulation variance.

The stochastic variance V_{sto} in (3.11) is generally easy to estimate. Based on m independent replications, the most commonly used estimator is

$$\widehat{V}_{\text{sto}} \;=\; \frac{1}{m-1}\sum_{j=1}^{m}(y_j-\overline{y})^2, \tag{3.12}$$

where $y_j = y(\boldsymbol{u}_j, \widehat{\theta})$ is the output of the jth replication, and $\overline{y}$ is the average of all y_j's and the estimate of the mean response.

The parameter variance V_{par} in (3.11) is more difficult to estimate. The method considered by Cheng and Holland (1997) is first to estimate $\boldsymbol{g}(\theta_0)$, and then to estimate V_{par} using the first term on the right-hand side of (3.11). The estimate of $\boldsymbol{g}(\theta_0)$ for a d-dimensional parameter vector θ_0 is obtained by making simulation runs in sets of $(d+1)$ runs with $\theta^1 = \widehat{\theta}$ at $\theta^{i+1} = \widehat{\theta} + \delta \boldsymbol{e}_i$, for $i = 1, \ldots, d$, where $\boldsymbol{e}_i$ is the d-dimensional unit vector with zero entries except for unity in the ith component. L'Ecuyer and Perron (1994) discuss appropriate choices for the value of the small displacement δ. The output responses from the simulation runs can then be written as

$$y(\boldsymbol{u}_j, \theta^i) = \eta(\theta^i) + e_j(\boldsymbol{u}_j, \theta^i);\; j = 1, \ldots, m \text{ and } i = 1, \ldots, d+1. \tag{3.13}$$

For each i the corresponding responses yield m estimates of the ith sensitivity coefficient g_i of $\boldsymbol{g}$,

$$\widehat{g}_{ij} \;=\; [y(\boldsymbol{u}_j, \theta^{i+1}) - y(\boldsymbol{u}_j, \theta^1)]/\delta, \;\; j = 1, \ldots, m. \tag{3.14}$$

The mean $\widehat{g}_i = \sum_{j=1}^{m} \widehat{g}_{ij}/m$ estimates g_i for $i = 1, \ldots, d$. The estimator for V_{par} is thus

$$\widehat{V}_{\text{par}} \;=\; \widehat{\boldsymbol{g}}^T \widehat{\boldsymbol{V}} \widehat{\boldsymbol{g}}, \tag{3.15}$$

where $\widehat{\boldsymbol{g}} = (\widehat{g}_1, \ldots, \widehat{g}_d)^T$, and $\widehat{\boldsymbol{V}}$ is the inverse of the observed information matrix.

The obvious method to calculate an approximate $100(1-\alpha)\%$ confidence interval for $\eta(\theta_0)$ is

$$\overline{y} \;\pm\; z_{\alpha/2}\sqrt{\frac{\widehat{V}_{\text{par}} + \widehat{V}_{\text{sto}}}{m}}, \tag{3.16}$$

where $\overline{y}$ is the maximum likelihood estimator of $\eta(\theta_0)$, and z_α is the upper α percentile of the normal distribution. We will evaluate the performance of such a confidence interval in Section 3.5 for a computer communication network application.

A problem with the δ-method is that some sensitivity coefficients might be small, which can seriously bias the estimate of variance of the simulation output. Another problem is that it is computationally expensive when the number of parameters is large. Other variants of the δ-method such as the two-point method (Cheng and Holland, 1998) are an attempt to solve some of these problems. A different approach which solves the problem of bias, but is computationally more expensive, is the bootstrap method.

3.3 Bootstrap Approach

The bootstrap was introduced by Efron (1979) as a computer-based method for estimating the variability of statistical estimators. The mostly commonly used formulation of the bootstrap approach is the *nonparametric bootstrap*. It enjoys the advantage of relieving the analyst from making distributional assumptions about the form of the underlying population, but it can be unrealistic and far too restrictive for simulation applications (Cheng and Holland, 1997). Provided that reasonable assumptions can be made about the form of the sampled probability distributions, the *parametric bootstrap* formulation seems much more convenient for simulation applications, and this is the method we now discuss.

Following the notation of Section 3.1, we substitute the estimator $\widehat{\theta}$ into $G(x, \theta)$ to obtain the fitted distribution $G(x, \widehat{\theta})$. We then draw a sample $\boldsymbol{x}_1^* = (x_{11}^*, x_{12}^*, \ldots, x_{1n}^*)$ from $G(x, \widehat{\theta})$, perhaps via the inverse transform (3.1). Corresponding to this bootstrap sample is the bootstrap estimate $\widehat{\theta}^*(1)$ of θ, computed from $\boldsymbol{x}_1^*$ in exactly the same way that $\widehat{\theta}$ was computed from $\boldsymbol{x}$. Repeating this sampling-and-estimation operation independently B times yields the estimates $\widehat{\theta}^*(1), \widehat{\theta}^*(2), \ldots, \widehat{\theta}^*(B)$. We can then carry out B bootstrap simulation experiments, one for each $\widehat{\theta}^*(i)$, with each run having the same length as in the original experiment; and altogether m' simulation runs are performed using the input parameter vector $\widehat{\theta}^*(i)$ for $i = 1, \ldots, B$. This yields B sets of simulation-generated output responses

$$\{y_{i1}^*, y_{i2}^*, \ldots, y_{im'}^*\} \text{ for } i = 1, 2, \ldots, B.$$

Let

$$\overline{y}_i^* \;=\; \frac{1}{m'} \sum_{j=1}^{m'} y_{ij}^*, \text{ for } i = 1, 2, \ldots, B \tag{3.17}$$

be the mean of the ith set of the B output responses, and let

$$\overline{\overline{y}}^* = \frac{1}{B}\sum_{i=1}^{B}\overline{y}_i^* \tag{3.18}$$

denote the grand mean of the bootstrap sample means. Using the approach of Cheng and Holland (1997) as outlined in Section 3.2, we see how the variance of $\overline{y}_i^*$ depends on both the stochastic variance and on the parameter variance. When each bootstrap is an exact replica of the original experiment (i.e. $m' = m$), Cheng and Holland suggest using the sample variance of the $\{\overline{y}_i^* : \ i = 1, 2, \ldots, B\}$,

$$S_B^2 = \frac{1}{B-1}\sum_{i=1}^{B}\left(\overline{y}_i^* - \overline{\overline{y}}^*\right)^2, \tag{3.19}$$

as an estimate for the variance of the original sample mean $\overline{y}$ specified in (3.8). Using (3.12) as an estimate for V_{sto} and (3.19), we estimate the parameter variance by

$$\widehat{V}_{\text{par}} = S_B^2 - \widehat{V}_{\text{sto}}. \tag{3.20}$$

An alternative is to use

$$\widehat{V}_{\text{sto}}^*(i) = \frac{1}{m'-1}\sum_{j=1}^{m'}\left(y_{ij}^* - \overline{y}_i^*\right)^2, \ i = 1, \ldots, B, \tag{3.21}$$

to estimate $\tau^2(\theta_0)$; and then we can use

$$\widehat{V}_{\text{sto}}^* = \frac{\widehat{V}_{\text{sto}} + \sum_{i=1}^{B} V_{\text{sto}}^*(i)}{B+1} \tag{3.22}$$

in place of $\widehat{V}_{\text{sto}}$ in (3.20).

Finally, we discuss how we can construct a bootstrap confidence interval for $\eta(\theta_0)$. One possibility is to construct a classical confidence interval similar to (3.16) and replace the variance estimates by their bootstrap estimates given in this section. However, this interval generally performs poorly in practice. Efron and Tibshirani (1993) explain this phenomenon and suggest methods to overcome such a drawback. A confidence interval that generally behaves better for such cases is the bootstrap percentile-type confidence interval. From the order statistics

$$\overline{y}_{(1)}^* \leq \overline{y}_{(2)}^* \leq \cdots \leq \overline{y}_{(B)}^*$$

of the $\{\overline{y}_i^* : \ i = 1, \ldots, B\}$, we obtain the $100(1-\alpha)\%$ percentile-type confidence interval

$$[\beta_{\mathrm{L}}^*, \beta_{\mathrm{U}}^*] \ \approx \ [\overline{y}^*_{(\lceil B\alpha/2 \rceil)}, \overline{y}^*_{(\lceil B(1-\alpha/2) \rceil)}]. \tag{3.23}$$

3.4 Bayesian Approach

We describe in this section the applicability of the Bayesian approach to our basic structure (3.3) of the discrete-event simulation model. We provide methods to estimate the posterior mean response and to construct a credible interval for that quantity which performs better in practice compared to the classical and bootstrap methods.

3.4.1 Estimating Mean Response

We observe a random sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ from our selected input model. Let θ be its d-dimensional vector of parameters with prior distribution $p(\theta)$. We assume that the hyperparameters of the prior distribution are either known or estimated using moment matching or some other empirical Bayes method (Carlin and Louis, 1996). We are not being fully Bayesian here by stopping the hierarchy at the second stage. However, we argue here that in most simulation applications in operations research and industrial engineering (as opposed, for example to applications in econometrics), prior information will be generally vague if it exists at more than one level down the hierarchy.

For our basic simulation model structure, we derive the posterior mean response given $\boldsymbol{x}$.

Theorem 3.1 *If the simulation response y has the form (3.3), then*

$$E(y|\boldsymbol{x}) \ = \ \int \eta(\theta)\, p(\theta|\boldsymbol{x})\, d\theta. \tag{3.24}$$

Proof. The law of total probability for expectations and Fubini's theorem ensure that

$$\begin{aligned} E(y|\boldsymbol{x}) \ &= \ \int \left[\int E(y(\boldsymbol{u}, \theta)|\boldsymbol{x}, \boldsymbol{u}, \theta)\, p(\theta|\boldsymbol{x})\, d\theta \right] d\boldsymbol{u} \\ &= \ \int \left[\int y(\boldsymbol{u}, \theta)\, d\boldsymbol{u} \right] p(\theta|\boldsymbol{x})\, d\theta \\ &= \ \int \eta(\theta)\, p(\theta|\boldsymbol{x})\, d\theta \ \text{ by (3.4).} \end{aligned}$$

for $r = 1, \ldots, R$
    generate the rth sample parameter vector θ^r from $p(\theta|\boldsymbol{x})$
    set the parameter vector $\theta \leftarrow \theta^r$
        for $j = 1, \ldots, m$
            set the random-number input $\boldsymbol{u} \leftarrow \boldsymbol{u}_j$
            perform the jth simulation run using $\boldsymbol{u}$ and θ
            calculate the output response $y_{rj} = y(\boldsymbol{u}, \theta)$
        end loop
    compute $\overline{y}_r = \sum_{j=1}^{m} y_{rj}/m$
end loop
compute the grand mean $\overline{\overline{y}} = \sum_{r=1}^{R} \overline{y}_r/R$ as an estimate for $E(y|\boldsymbol{x})$

Figure 3.1: Bayesian Simulation Replication Algorithm

△

Our first objective is to develop an approach that accounts fully for parameter and stochastic uncertainty and that can be extended easily to account for model uncertainty. This approach should also have a theoretical appeal, and most importantly it should perform well in practice. Figure 3.1 summarizes the Bayesian Simulation Replication Algorithm that is proposed in this research to implement a Bayesian approach to simulation input modeling. The net effect of the algorithm is to account fully for the uncertainty in the parameters of the input model as well as to account fully for the usual stochastic uncertainty. The algorithm can be seen as an uncertainty decomposition algorithm. The inner loop will be used to generate estimates for the stochastic uncertainty, whereas the outer loop will estimate the parameter uncertainty. The next subsection gives a detailed explanation of how to estimate the stochastic and parameter variances and construct a credible interval on the posterior mean response.

3.4.2 Assessing Output Variability

We try to assess the variability of the simulation output based on simple response surface models, given the objective of estimating the mean response in our simulation study. The analysis given below can be extended to more complicated models, but this is beyond the scope of our work. We propose two methods of estimating the

parameter and stochastic variances in the simulation output.

Classical Output Analysis

Following our basic simulation structure of Section 3.1, we see that the output responses from the simulation runs performed by the Bayesian Simulation Replication Algorithm of Figure 3.1 are given by

$$y_{rj} = y(\boldsymbol{u}_j, \theta^r) \;\; = \;\; \eta(\theta^r) + e_j(\boldsymbol{u}_j, \theta^r);\; r = 1, \ldots, R;\; j = 1, \ldots, m. \tag{3.25}$$

We generally assume that

$$E(e_j|\theta^r) = 0 \quad \text{and} \quad \text{Var}(e_j|\theta^r) = \tau^2, \tag{3.26}$$

where τ^2 does not depend on θ^r. Given that our main objective is to estimate the overall mean response, we further assume that

$$\eta(\theta^r) \;\; = \;\; \beta + \delta_r(\theta^r), \tag{3.27}$$

where

$$\begin{aligned} \beta = \beta(\boldsymbol{x}) \;\; &= \;\; E_{\theta^r}[\eta(\theta^r)] \\ &= \;\; \int \eta(\theta)\, p(\theta|\boldsymbol{x})\, d\theta \\ &= \;\; E(y|\boldsymbol{x}) \end{aligned}$$

from Theorem 3.1; and

$$E_{\theta^r}(\delta_r) = 0 \quad \text{and} \quad \text{Var}_{\theta^r}(\delta_r) = \sigma^2. \tag{3.28}$$

Based on these assumptions, we show in the following theorem that the posterior variance can be written as the sum of two variances measuring the stochastic and parameter uncertainty, respectively.

Theorem 3.2 *If (3.25)–(3.28) hold, then*

$$\text{Var}(y|\boldsymbol{x}) \;\; = \;\; \tau^2 + \sigma^2. \tag{3.29}$$

Proof. Using assumptions (3.26) and (3.28), we have

$$\begin{aligned}
\mathrm{Var}(y|\boldsymbol{x}) &= E_\theta[\mathrm{Var}(y|\boldsymbol{x},\theta)|\boldsymbol{x}] + \mathrm{Var}_\theta[E(y|\boldsymbol{x},\theta)|\boldsymbol{x}] \\
&= E_\theta[\mathrm{Var}(e|\theta)|\boldsymbol{x}] + \mathrm{Var}_\theta[\eta(\theta)|\boldsymbol{x}] \\
&= E_\theta[\tau^2|\boldsymbol{x}] + \mathrm{Var}_\theta[\beta + \delta(\theta)|\boldsymbol{x}] \\
&= \tau^2 + \mathrm{Var}_\theta[\delta(\theta)|\boldsymbol{x}] \\
&= \tau^2 + \sigma^2.
\end{aligned}$$

△

The response surface model given by (3.25)–(3.28) is known in the statistical literature as the classical random-effects model (Rao, 1997), where one estimates β, τ^2, and σ^2 using the simulation-generated statistics specified in Figure 3.1 as follows:

$$\widehat{\beta} = \overline{\overline{y}}, \tag{3.30}$$

$$\widehat{\tau}^2 = \frac{\sum_{r=1}^{R}\sum_{j=1}^{m}(y_{rj} - \overline{y}_r)^2}{R(m-1)}, \tag{3.31}$$

and

$$\widehat{\sigma}^2 = \frac{\sum_{r=1}^{R}(\overline{y}_r - \overline{\overline{y}})^2}{(R-1)} - \frac{\widehat{\tau}^2}{m}. \tag{3.32}$$

In many applications, especially the ones where we are totally ignorant about the output performance measure, the above formulation delivers reasonable point estimates. However, there are two important drawbacks of using such an approach. The first drawback concerns the estimate $\widehat{\sigma}^2$, which can be negative. The problem of a negative estimate for a variance component can be avoided by setting it equal to zero, but this creates new issues (Rao, 1997). The main drawback comes from the fact that substituting point estimates for the above parameters ignores our real uncertainty about them. We suggest in the next subsection a full Bayesian treatment of the same model with noninformative priors based on normally distributed simulation output observations.

In addition to point estimates, we can also construct a credible interval for β from the output of the Bayesian Simulation Replication Algorithm given in Figure 3.1.

Similar to the bootstrap approach, we can construct an approximate $100(1-\alpha)\%$ Bayesian percentile credible interval for β as

$$[\beta_{\mathrm{L}}, \beta_{\mathrm{U}}] \approx [\overline{y}_{(\lceil R\alpha/2 \rceil)}, \overline{y}_{(\lceil R(1-\alpha/2) \rceil)}], \tag{3.33}$$

where the quantities

$$\overline{y}_{(1)} \leq \overline{y}_{(2)} \leq \cdots \leq \overline{y}_{(R)}$$

denote the order statistics of the $\{\overline{y}_r : r = 1, \ldots, R\}$ specified in Figure 3.1. These intervals generally perform better in practice than t-type confidence intervals (Efron and Tibshirani, 1993).

Bayesian Output Analysis

Under the hierarchical normal model, we assume that the output data $\{y_{rj}\}$ specified in Figure 3.1 are independently normally distributed, that is

$$y_{rj} \mid \mu_r, \tau \sim \mathrm{N}(\mu_r, \tau^2),\ r = 1, \ldots, R \text{ and } j = 1, \ldots, m. \tag{3.34}$$

For $r = 1, \ldots, R$, we also assume that the parameter μ_r, which corresponds to $\eta(\theta^r)$ in equation (3.25), is also normally distributed:

$$\mu_r \mid \beta, \sigma \sim \mathrm{N}(\beta, \sigma^2),\ r = 1, \ldots, R. \tag{3.35}$$

Note that although assumption (3.34) generally holds in practice because of the simulation output usually being an average of a large number of output random variables, assumption (3.35) may not be consistent with the form of the posterior distribution $p(\theta|\boldsymbol{x})$.

To complete the Bayesian formulation, we assume a noninformative prior distribution for (β, τ, σ), with $\sigma > 0$ and $\tau > 0$; specifically we take

$$p(\beta, \tau, \sigma) \propto \tau^{-1}. \tag{3.36}$$

The joint posterior density of all the parameters is

$$p(\boldsymbol{\mu}, \beta, \tau, \sigma \mid \boldsymbol{y}) \propto \tau^{-1} \prod_{r=1}^{R} p(\mu_r \mid \beta, \sigma) \prod_{r=1}^{R} \prod_{j=1}^{m} p(y_{rj} \mid \mu_r, \tau),$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_R)$ and $\boldsymbol{y} = \{y_{rj} : r = 1, \ldots, R;\ j = 1, \ldots, m\}$.

Many numerical methods such as conditional maximization can be used to obtain posterior point estimates for the parameters of interest. However, we are also interested to compute posterior confidence intervals for these parameters; and Markov Chain Monte Carlo (MCMC) methods are appropriate for such inferences. In fact, we can even obtain a large sample from the posterior distribution of each parameter from which we can estimate its density.

The idea behind MCMC (Gilks et al., 1996) is to simulate a random walk in the space of $(\boldsymbol{\mu}, \beta, \tau, \sigma)$ which converges to a stationary distribution that is the joint posterior distribution, $p(\boldsymbol{\mu}, \beta, \tau, \sigma \mid \boldsymbol{y})$. The most widely used MCMC method is the *Gibbs Sampler* algorithm (Casella and George, 1994). Figure 3.2 summarizes briefly the steps of the algorithm. For our problem this algorithm can be easily implemented in any statistical software package, given its simplicity and the fact that we can generate variates easily from the following required conditional distributions (Gelman et al., 1995):

$$\mu_r \mid \beta, \tau, \sigma, \boldsymbol{y} \;\sim\; \mathrm{N}\left(\frac{\frac{1}{\sigma^2}\beta + \frac{m}{\tau^2}\overline{y}_r}{\frac{1}{\sigma^2} + \frac{m}{\tau^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{m}{\tau^2}}\right), \; r = 1, \ldots, R; \tag{3.37}$$

$$\beta \mid \boldsymbol{\mu}, \tau, \sigma, \boldsymbol{y} \;\sim\; \mathrm{N}\left(\frac{1}{R}\sum_{r=1}^{R}\mu_r, \frac{\sigma^2}{R}\right); \tag{3.38}$$

$$\tau^2 \mid \boldsymbol{\mu}, \beta, \sigma, \boldsymbol{y} \;\sim\; \mathrm{IG}\left(\frac{mR}{2}, \frac{1}{2}\sum_{r=1}^{R}\sum_{j=1}^{m}(y_{rj} - \mu_r)^2\right); \tag{3.39}$$

and

$$\sigma^2 \mid \boldsymbol{\mu}, \beta, \tau, \boldsymbol{y} \;\sim\; \mathrm{IG}\left(\frac{R-1}{2}, \frac{1}{2}\sum_{r=1}^{R}(\mu_r - \beta)^2\right), \tag{3.40}$$

where the density function of a random variable Z having an Inverse Gamma distribution $\mathrm{IG}(\nu, \phi)$ with shape parameter ν and scale parameter ϕ is defined as

$$p(z|\nu, \phi) \;=\; \frac{\phi^{\nu} e^{-\phi/z}}{\Gamma(\nu) z^{\nu+1}}, \; z > 0, \; \nu > 0, \; \phi > 0. \tag{3.41}$$

The classical estimates of β, τ^2, and σ^2 given by equations (3.30), (3.31), and (3.32), respectively, are good initial estimates with which to start the Gibbs sampler algorithm. If the initial estimate $\widehat{\sigma}^2$ is negative, then we can set it to a small value, say

Set β_0, τ_0, and σ_0 to their classical estimates
(3.30), (3.31), and (3.32), respectively.
For $t = 1, \ldots, T^*, T^* + 1, \ldots, T$, generate:
$\mu_{rt} \sim p\left(\mu_r \mid \beta_{t-1}, \tau_{t-1}, \sigma_{t-1}, \boldsymbol{y}\right)$ as in (3.37), $r = 1, \ldots, R$,
and take $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{Rt})$;
$\beta_t \sim p\left(\beta \mid \boldsymbol{\mu}_t, \tau_{t-1}, \sigma_{t-1}, \boldsymbol{y}\right)$ as in (3.38);
$\tau_t^2 \sim p\left(\tau^2 \mid \boldsymbol{\mu}_t, \beta_t, \sigma_{t-1}, \boldsymbol{y}\right)$ as in (3.39); and
$\sigma_t^2 \sim p\left(\sigma^2 \mid \boldsymbol{\mu}_t, \beta_t, \tau_t, \boldsymbol{y}\right)$ as in (3.40).
end loop
Generate the estimates of the parameters of interest:

$$\widehat{\beta} = \sum_{t=T^*+1}^{T} \beta_t/(T - T^*)$$
$$\widehat{\tau}^2 = \sum_{t=T^*+1}^{T} \tau_t^2/(T - T^*)$$
$$\widehat{\sigma}^2 = \sum_{t=T^*+1}^{T} \sigma_t^2/(T - T^*)$$

Figure 3.2: Gibbs Sampler Algorithm

0.001. The output inference from the Gibbs sampler algorithm is based on $T = 50000$ iterations, with a warm-up period of $T^* = 5000$ iterations. These relatively large numbers of iterations are generally required for convergence because of the high auto-correlation between successively sampled values of some parameters (Speiegelhalter et al., 1996). The point estimates of β, τ^2, and σ^2 are given in Figure 3.2. A $100(1-\alpha)\%$ credible interval for β can be constructed from the output of the Gibbs sampler algorithm as

$$[\beta_{\mathrm{L}}, \beta_{\mathrm{U}}] \quad \approx \quad [\beta_{(\lceil (T-T^*)\alpha/2 \rceil)}, \beta_{(\lceil (T-T*)(1-\alpha/2) \rceil)}], \tag{3.42}$$

where the quantities

$$\beta_{(1)} \leq \beta_{(2)} \leq \cdots \leq \beta_{(T-T^*)}$$

denote the order statistics of the $\{\beta_t : t = T^* + 1, \ldots, T\}$ generated by the Gibbs sampler algorithm of Figure 3.2. Credible intervals similar to (3.42) can also be constructed for τ^2 and σ^2.

3.5 Performance Evaluation

Which of the above methodologies promises to produce better estimates, in the sense of closeness to the target value being estimated, in real inference problems that are likely to be faced in practice? Such problems are often nonasymptotic and tend to be accompanied by some, but often not very much, prior information. In this section we propose some criteria for comparing Bayesian and frequentist (classical and bootstrap) approaches within the discrete-event simulation framework. We will test our ideas empirically by applying them to two queueing systems. Before proceeding, we will, for clarity's sake, indicate the sense in which the terms "Bayesian" and "frequentist" will be used in the sequel. We will take the view that a frequentist is an analyst who seeks to estimate an unknown parameter based only on the model that has been adopted for the observable data, whereas a Bayesian is one who seeks to estimate the parameter by appropriately combining his or her prior intuition with the information content in the data.

3.5.1 Evaluation Criteria

Point Estimation

Suppose that a sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ of independent random variables is to be drawn; and given the value of a scalar parameter θ_0, $\boldsymbol{x}$ has the conditional probability density $g_{\theta_0}(\boldsymbol{x})$. We assume that the unknown parameter θ_0 is a random variable and refer to the distribution π_0 of θ_0 as the "true prior" distribution. Here it is important to distinguish between (a) the "true" parameter vector θ_0 of the inputs $\boldsymbol{x}$ to the real system with response y_0 versus (b) the parameter θ of the input model for the simulation with response $y = y(\boldsymbol{u}, \theta)$, even when θ is sampled from the posterior distribution $p(\theta|\boldsymbol{x})$ as in the Simulation Replication Algorithm of Figure 3.1.

For a realization θ_0 obtained from π_0 (which will be unknown to the Bayesian and frequentist analysts), we assume that we can compute the true output performance measure of the real system $\beta_0(\theta_0) = E(y_0|\theta_0)$ (which will also be unknown to both analysts). The analysts will then attempt to estimate $\beta = \beta(\boldsymbol{x}) = E(y|\boldsymbol{x})$ using their respective simulation procedures based solely on observing the experimental outcomes $\{x_1, x_2, \ldots, x_n\}$ drawn from g_{θ_0}.

Our criterion for assessing the performance of a given estimator $\widehat{\beta}$ of $\beta_0(\theta_0)$ will

be the average risk $\mathcal{R}$ of $\widehat{\beta}$ relative to the true prior distribution π_0 of θ_0,

$$\begin{aligned}\mathcal{R}(\pi_0, \widehat{\beta}) &= E_{\pi_0} E_{g_{\theta_0}} \left[\widehat{\beta}(\boldsymbol{x}) - \beta_0(\theta_0)\right]^2 \\ &= \int \left\{ \int \left[\widehat{\beta}(\boldsymbol{x}) - \beta_0(\theta_0)\right]^2 g_{\theta_0}(\boldsymbol{x})\, d\boldsymbol{x} \right\} \pi_0(\theta_0)\, d\theta_0, \end{aligned} \tag{3.43}$$

where the simulation-based estimator $\widehat{\beta}(\boldsymbol{x})$ is given by $\overline{\overline{y}}$ when we adopt the Bayesian Simulation Replication Algorithm of Figure 3.1. When we adopt the frequentist approach, we compute the maximum likelihood estimator $\widehat{\theta}^{\text{mle}} = \widehat{\theta}^{\text{mle}}(\boldsymbol{x})$ based on the original input data, and then in (3.43) we take

$$\widehat{\beta}(\boldsymbol{x}) = \frac{1}{m} \sum_{j=1}^{m} y\left(\boldsymbol{u}_j, \widehat{\theta}^{\text{mle}}(\boldsymbol{x})\right) \tag{3.44}$$

based on a set of m independent simulation runs.

In the Bayesian approach we will use the "operational" prior π, chosen by the Bayesian analyst, to compute the posterior distribution $p(\theta|\boldsymbol{x})$ of θ. The operational prior will generally be noninformative or minimally informative. It is also important to stress that the true prior π_0 is entirely unknown to the Bayesian and frequentist analysts. The real contest will be to find the estimator that minimizes $\mathcal{R}(\pi_0, \widehat{\beta})$.

The criterion we propose for assessing performance makes sense and is reasonable within the Bayesian and frequentist paradigms. Although the notion of a "true prior" seemingly conflicts with the "degree of belief" interpretation of prior distributions espoused by many Bayesians, it should be noted that the Bayesians' degree of belief about θ is reflected in π, not π_0, and that the idea of a true prior distribution makes perfect sense in a computer-assisted experiment where θ_0 is generated at random from π_0, whose form is unknown to both analysts. If the true θ_0 is constant and π_0 is taken as degenerate at that constant, then the average risk reduces to the mean squared error (MSE), which is widely used by frequentists to judge estimators. In our framework, the average risk is precisely the squared error in estimating $\beta_0(\theta_0)$ by $\widehat{\beta}$, averaged over all the randomness in the problem.

Interval Estimation

The Bayesian approach can be also used to develop interval estimation procedures (Section 3.4.2) to assess its performance compared to the classical and bootstrap methods. As in the previous subsection, our strategy will be to use noninformative

priors. The main properties used to assess the performance of a Bayesian credible interval for $\beta_0(\theta_0)$ are the average interval length and variance of the interval length as well as its coverage probability. We now examine the coverage of a Bayesian credible interval. Adopting the notation

$$\mathcal{B}(\boldsymbol{x}) \equiv \left[\widehat{\beta}_L(\boldsymbol{x}), \widehat{\beta}_U(\boldsymbol{x})\right],$$

this is defined as

$$\begin{aligned} \mathcal{C}(\beta_0(\theta_0)) &= E_{\pi_0} E_{g_{\theta_0}} \left[I_{\{\beta_0(\theta_0)\in\mathcal{B}(\boldsymbol{x})\}}\right] \\ &= \int \left\{ \int I_{\{\beta_0(\theta_0)\in\mathcal{B}(\boldsymbol{x})\}}\, g_{\theta_0}(\boldsymbol{x})\, d\boldsymbol{x} \right\} \pi_0(\theta_0)\, d\theta_0. \end{aligned} \tag{3.45}$$

The average credible interval length is

$$\begin{aligned} \mathcal{L} &= E_{\pi_0} E_{g_{\theta_0}} \left[\widehat{\beta}_U(\boldsymbol{x}) - \widehat{\beta}_L(\boldsymbol{x})\right] \\ &= \int \left\{ \int \left[\widehat{\beta}_U(\boldsymbol{x}) - \widehat{\beta}_L(\boldsymbol{x})\right]\, g_{\theta_0}(\boldsymbol{x})\, d\boldsymbol{x} \right\} \pi_0(\theta_0)\, d\theta_0, \end{aligned} \tag{3.46}$$

and the variance of the credible interval length is given by

$$\begin{aligned} \mathcal{V} &= E_{\pi_0} E_{g_{\theta_0}} \left\{ \left[\widehat{\beta}_U(\boldsymbol{x}) - \widehat{\beta}_L(\boldsymbol{x}) - \mathcal{L}\right]^2 \right\} \\ &= \int \left\{ \int \left[\widehat{\beta}_U(\boldsymbol{x}) - \widehat{\beta}_L(\boldsymbol{x}) - \mathcal{L}\right]^2\, g_{\theta_0}(\boldsymbol{x})\, d\boldsymbol{x} \right\} \pi_0(\theta_0)\, d\theta_0. \end{aligned} \tag{3.47}$$

In the next two applications, we will see that in addition to producing point estimates with a small average risk, the Bayesian approach employing vague priors also achieves a high level of coverage, at a reasonable average interval length. We will also see in the next chapter that the Bayesian formalism may be the only way to develop confidence intervals that reflect all input uncertainties.

3.5.2 Experimental Design

We developed in the previous subsections point and interval estimation criteria to assess the performance of the frequentist and Bayesian approaches in a discrete-event simulation framework. However, criteria (3.43), (3.45), (3.46), and (3.47) will be generally hard to compute analytically in most of the applications and should be estimated from the output realizations. The performance evaluation of the Bayesian

and frequentist approaches is generally conducted by Monte Carlo computer-assisted simulation experiments. The output of these experiments can be used to estimate our performance criteria. Figure 3.3 summarizes the protocol we used to conduct N Monte Carlo experiments and deliver estimates for the average risk of each estimator and its standard error, the mean and the variance of the interval lengths, and their coverage probabilities.

3.5.3 Application to a Single Server Queue

To provide a practical illustration of the above performance evaluation framework, we consider a simple single server queueing system. For this example, we will just compare the performance of the Bayesian and bootstrap approaches since the computational effort is approximately the same. Moreover, the average risk for the classical and bootstrap method is the same since they both use the same estimator for the mean response. The next experiment considers a larger scale example with a degenerate true prior distribution, where the main objective is to compare the interval estimation performance of the Bayesian method with the classical and bootstrap results of Cheng and Holland (1997).

The single server system has customers enter the queue according to a Poisson process with rate λ, having a "true" Gamma prior distribution π_0 with shape parameter $\nu_1 = 50$ and scale parameter ϕ_1 that will take a range of values corresponding to increasing levels of prior traffic intensity. The customers receive an exponential service time with rate μ, having a "true" Gamma prior distribution with shape parameter $\nu_2 = 50$ and scale parameter $\phi_2 = 100$. The objective is to estimate the average time in the system. Monte Carlo experiments are based on data samples of size $n = 1000$, and are repeated $N = 500$ times.

For the Bayesian approach, we chose the operational prior distribution for the arrival rate λ to be noninformative having a density proportional to λ^{-1}. After observing a sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ of interarrival times, we can compute the posterior density of λ as

$$\begin{aligned} p(\lambda|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\lambda) \times p(\lambda) \\ &= \prod_{j=1}^{n} \lambda e^{-\lambda x_j} \times \lambda^{-1} \\ &= \lambda^{n-1} e^{-\lambda \sum_{j=1}^{n} x_j}, \end{aligned}$$

Choose the true prior distribution π_0

For $i = 1, \ldots, N$

Sample $\theta_{0,i}$ from π_0

Compute the true performance measure $\beta_{0,i} = \beta_0(\theta_{0,i})$

Sample $\boldsymbol{x}_i$ from $g_{\theta_{0,i}}(\boldsymbol{x})$

Compute the Bayes estimate $\widehat{\beta}_i^{\text{Bayes}} = \overline{\overline{y}}$ as specified in Figure 3.1 using an operational prior π

Compute the frequentist estimate $\widehat{\beta}_i^{\text{Freq}}$ as specified in (3.44) using $\widehat{\theta}^{\text{mle}}$

Compute a confidence interval (3.16) for $\beta_{0,i}$ using the classical approach

Compute a confidence interval (3.23) for $\beta_{0,i}$ using the bootstrap approach

Compute a credible interval (3.33) for $\beta_{0,i}$ using the Bayesian approach

end loop

Compute the estimated average risk for each simulation-based approach as

$$\widehat{\mathcal{R}} = \tfrac{1}{N} \sum_{i=1}^{N} \left(\widehat{\beta}_i - \beta_{0,i} \right)^2$$

with associated Monte Carlo standard error estimate

$$\widehat{\text{SE}}(\widehat{\mathcal{R}}) = \left\{ \tfrac{1}{N(N-1)} \sum_{i=1}^{N} \left[\left(\widehat{\beta}_i - \beta_{0,i} \right)^2 - \widehat{\mathcal{R}} \right]^2 \right\}^{\frac{1}{2}}$$

Compute the estimated coverage probability for each approach as

$$\widehat{\mathcal{C}} = \tfrac{1}{N} \sum_{i=1}^{N} I_{\{\beta_{0,i} \in \mathcal{B}_i(\boldsymbol{x}_i)\}}$$

Compute the estimated mean of the CI length for each approach as

$$\widehat{\mathcal{L}} = \tfrac{1}{N} \sum_{i=1}^{N} \left[\widehat{\beta}_U(\boldsymbol{x}_i) - \widehat{\beta}_L(\boldsymbol{x}_i) \right]$$

Compute the estimated variance of the CI length for each approach as

$$\widehat{\mathcal{V}} = \tfrac{1}{N-1} \sum_{i=1}^{N} \left[\widehat{\beta}_U(\boldsymbol{x}_i) - \widehat{\beta}_L(\boldsymbol{x}_i) - \widehat{\mathcal{L}} \right]^2$$

Figure 3.3: Monte Carlo Experimental Design

which is recognized as a Gamma density with shape parameter n and scale parameter $1/(\sum_{j=1}^{n} x_j)$. Similarly, choosing the same functional form for the prior distribution of the service rate μ produces a proper posterior Gamma density having a shape parameter n and a scale parameter $1/(\sum_{j=1}^{n} z_j)$, where $\boldsymbol{z} = \{z_1, z_2, \ldots, z_n\}$ is the observed sample of service times.

Table 3.1 shows the results for the average risk, coverage probability and confidence interval length. It appears that the Bayesian point estimator averages a slightly smaller risk in all cases. The main conclusion concerns the excellent performance of the Bayesian method in terms of interval estimation. The coverage probability is very close to the nominal coverage with shorter credible intervals than their bootstrap counterparts. Although the system considered here is simple and these results may not be generalized for all systems, there is strong empirical evidence in favor of the Bayesian approach. The next example will further support this conclusion.

Table 3.1: Average Risk of Classical Bootstrap and Proposed Bayesian Estimators for Average Sojourn Time in a Single Server Queue, Including Length, Coefficient of Variation, and Coverage Probability of Nominal 90% Confidence Intervals

ϕ_1	Method	$\widehat{\mathcal{R}}$	$\widehat{\mathrm{SE}}(\widehat{\mathcal{R}})$	CI Length	CV(CIL)	Coverage
10	Bootstrap	9.3×10^{-11}	6.0×10^{-12}	4.9×10^{-5}	9.5×10^{-2}	98.2
	Bayesian	9.1×10^{-11}	5.9×10^{-12}	3.0×10^{-5}	9.6×10^{-2}	88.8
30	Bootstrap	1.4×10^{-10}	1.0×10^{-11}	6.3×10^{-5}	1.0×10^{-1}	99.0
	Bayesian	1.3×10^{-10}	9.8×10^{-12}	3.9×10^{-5}	1.1×10^{-1}	89.6
50	Bootstrap	2.7×10^{-9}	2.0×10^{-10}	2.6×10^{-4}	1.7×10^{-1}	99.4
	Bayesian	2.4×10^{-9}	1.8×10^{-10}	1.5×10^{-4}	1.5×10^{-1}	90.0
70	Bootstrap	8.0×10^{-9}	1.1×10^{-9}	4.0×10^{-4}	1.8×10^{-1}	98.8
	Bayesian	5.9×10^{-9}	6.8×10^{-10}	2.3×10^{-4}	1.8×10^{-1}	90.4
90	Bootstrap	2.5×10^{-4}	1.2×10^{-4}	7.5×10^{-3}	9.3×10^{-1}	99.4
	Bayesian	1.1×10^{-5}	4.5×10^{-6}	4.3×10^{-3}	6.0×10^{-1}	91.6

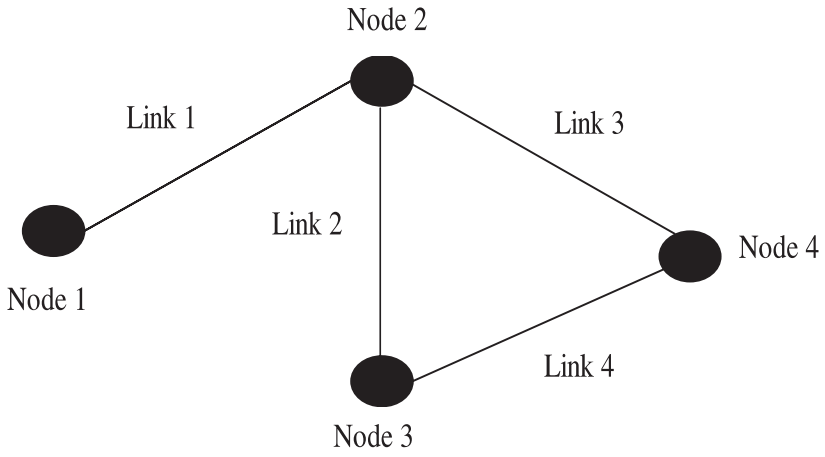

Figure 3.4: A Communication Network with $\mathbb{Q} = 4$ nodes and $\mathbb{L} = 4$ links

3.5.4 Application to a Computer Communication Network

In this example we consider a simulation of a computer communications network (Kleinrock, 1976). It is a collection of $\mathbb{Q}$ nodes consisting of computing resources which communicate with each other along a set of $\mathbb{L}$ links (the data communication channels). The aim of the simulation study is to measure the delay in messages transmitted between nodes via the communication channels. Figure 3.4 illustrates a network with $\mathbb{Q} = 4$ and $\mathbb{L} = 4$.

The $\mathbb{L}$ communication channels are assumed to be noiseless, and have a capacity of C_i bits per second for the ith channel. The $\mathbb{Q}$ nodes carry out the administration tasks such as message reassembly and routing. It is assumed that the nodal processing times are constant with value $\mathcal{T}_i$ for the ith node. In addition there are channel queueing and transmission delays. Traffic entering the network from any node forms a Poisson process with rate $\gamma(i, j)$ (messages per second) for those messages originating at node i and destined for node j. All messages are assumed to have lengths that follow an exponential distribution with mean ψ (bits). We assume that all nodes have unlimited storage capacity and that all messages are directed through the network on fixed paths. In high speed networks spanning large geographical regions, it may be important to include the propagation time H_i, which is the time required for the energy representing a single bit to propagate down the length of the ith channel. The speed of energy propagation, v miles per second, is a significant fraction of the

speed of light depending on the particular type of channel used. If the ith channel has length l_i miles, then $H_i = l_i/v$. Thus if a message has X bits then the time it occupies the ith channel will be $H_i + X/C_i$ seconds. Note that the randomness in the service time comes not from the server (the channel) but from the exponentially distributed customer message lengths.

We can see that in this model there are a large number of variables, particularly in networks with many nodes or links. Some of the parameters will either be known exactly or should be estimable to a high degree of accuracy, for example C_i, l_i and v. Other parameters, however, may only be estimated by observing samples of values. Foremost among there are ψ, the average message size, and $\gamma(i, j)$, the external nodal arrival rate at node i of messages destined for node j.

In our example of Figure 3.4 with $\mathbb{Q} = 4$ nodes and $\mathbb{L} = 4$ links, the following parameters are assumed known: $\mathcal{T}_i = 0.001$ seconds ($i = 1, \ldots, \mathbb{Q}$), C_i =275,000 bits/second ($i = 1, \ldots, \mathbb{L}$), $l_i = i \times 100$ miles ($i = 1, \ldots, \mathbb{L}$), and v =150,000 miles/second. The true message length was taken to be $\psi_0 = 300$, and the true traffic arrival rates were $\gamma_0(1, 2) = 60$, $\gamma_0(1, 3) = 40$, $\gamma_0(1, 4) = 50$, $\gamma_0(2, 1) = 80$, $\gamma_0(2, 3) = 65$, $\gamma_0(2, 4) = 20$, $\gamma_0(3, 1) = 100$, $\gamma_0(3, 2) = 22$, $\gamma_0(3, 4) = 26$, $\gamma_0(4, 1) = 40$, $\gamma_0(4, 2) = 50$, $\gamma_0(4, 3) = 60$. These were assumed unknown in the simulation, but data samples of size n were observed from the exponential distribution (n was taken as 5,000 and 50,000). A basic simulation run was of 50 seconds in length and this was repeated $m = 10$ times. For the bootstrap simulation, and to save on computing time, B was taken as 100. For a fair comparison with the bootstrap simulation, the sample size R generated from the posterior distribution of the parameters (using vague priors similar to the single server queue application) was also taken to be 100. In practice larger values of B and R are recommended, typically 1000. However, $B = R = 100$ is sufficiently large in this case because the observed coverages are stable, indicating the satisfactory behavior of the methods. In each case the experiment was repeated $N = 200$ times so that 200 confidence intervals were generated.

The "true" value β_0 of the average delay of a message in this communication network cannot be computed analytically. So we used a preliminary Monte Carlo experiment involving direct simulation of the network to compute β_0 to within $\pm 0.05\%$ of its true value with 99% confidence. For a fixed number of replications m, let $\overline{y}(m)$ and $S_y(m)$ denote the corresponding sample mean and standard deviation of the observed average message delays. Given prespecified values of the percentage

error tolerance ω, the confidence coefficient α, and the preliminary sample size m_0, we determined the final number of replications according to the following relative-precision stopping rule (Law et al., 1981)

$$m^* = \min\left\{ m : \ m \geq m_0, \ m = 0 \ (\text{mod } 10), \ S_y(m) > 0, \right.$$

$$\left. \text{and } t_{1-\alpha/2}(m-1)\frac{S_y(m)}{\sqrt{m}} \leq \omega|\overline{y}(m)| \right\}, \tag{3.48}$$

and β_0 was taken to be $\overline{y}(m)$. In our preliminary experiment, we took $m_0 = 500$, $\omega = 0.0005$, and $\alpha = 0.01$. The final sample size m^* was found to be 2450 with a final estimate of β_0 being 0.008207.

Table 3.2 shows that the Bayesian method gives the tightest confidence bands with a coverage probability closer to the nominal one. The coverage probabilities of the bootstrap method are also close to the target, but the confidence interval lengths are almost twice as large as the Bayesian credible intervals, although both results are obtained at almost the same computational effort. However, both the Bayesian and bootstrap methods required almost 9 times the computational effort required for the classical approach. But we should also note that the computational effort required for the classical method will increase linearly with the number of unknown parameters. Hence for more realistic communication networks, we may still have more accurate results using the Bayesian approach with less computational effort.

Table 3.2: Performance of Nominal 90% Confidence Intervals for Average Message Delay in the Computer Communication Network of Figure 3.4

Method	n =5,000		n =50,000	
	CI Length	Coverage	CI Length	Coverage
Classical	1.02×10^{-3}	98	3.66×10^{-4}	96
Bootstrap	7.34×10^{-4}	91	2.97×10^{-4}	94
Bayesian	3.44×10^{-4}	90	1.30×10^{-4}	93

Chapter 4

Accounting for Model Uncertainty in Simulation

The widespread application of stochastic discrete-event simulations is accompanied by a widespread concern about quantifying the uncertainties prevailing in their use. There are three main sources of uncertainty in a simulation experiment:

(a) *Stochastic uncertainty.* This source of variation arises from the dependence of the simulation output on the uniform random variates (random numbers) generated during each simulation run (Helton, 1998).

(b) *Model uncertainty.* This source of variation typically occurs when choosing between different input models that adequately fit the available sample data or subjective information (Raftery et al., 1996).

(c) *Parameter uncertainty.* This source of variation arises because the parameters of the selected input model(s) are unknown and must be estimated from available sample data or subjective information (Raftery et al., 1996).

The model and parameter uncertainty are represented by the random variables M and θ_M, respectively, both of which are assumed to depend only on the available subjective information or sample data; and the stochastic uncertainty depends only on the randomness of the uniform variates (random numbers) $\boldsymbol{u}$ generated during each simulation run. Thus an output quantity of interest from the simulation run, y, can be regarded as an unknown complicated function of $\boldsymbol{u}$, M, and θ_M,

$$y = y(\boldsymbol{u}, M, \theta_M). \tag{4.1}$$

Let

$$\eta(M, \theta_M) \;=\; E(y|M, \theta_M) \;\;=\;\; \int y(\boldsymbol{u}, M, \theta_M)\, d\boldsymbol{u}, \tag{4.2}$$

be the expected value of y, given the input model M and the corresponding parameter vector θ_M. The objective of the classical simulation experiment is generally to estimate $\eta(M_0, \theta_0)$, where θ_0 is the true but unknown value of the parameter vector θ_{M_0} under the true model M_0, estimated separately from the simulation experiment using real data. In general this approach fails to assess and propagate model and parameter uncertainty and may lead to miscalibrated uncertainty assessments about y (Draper, 1995). The δ-method (Stuart and Ord, 1994) and the bootstrap method (Cheng and Holland, 1997) are possible ways to account for parameter uncertainty. However, in addition to their failure to incorporate relevant information other than the observed data points, these methods cannot be extended to account for model uncertainty.

In this chapter, we present the Bayesian Model Averaging (BMA) approach as a coherent mechanism to account for all sources of uncertainty. The basic ingredients of the BMA approach for conducting simulation experiments are discussed in Section 4.1. After modeling our uncertainty about the input processes through prior probability distributions and observing data on those input processes, we also derive an expression for the posterior mean response. In Section 4.2 we develop a "Simulation Replication Algorithm Based on Bayesian Model Averaging" to estimate the posterior mean response and assess the variability of the resulting estimator. In Section 4.3 we use the output of the algorithm to estimate the components of this variance that are due to each source of uncertainty. We also discuss how to construct a valid credible (or confidence) interval for the overall posterior mean response. In Section 4.4 we develop a replication allocation procedure that optimally allocates simulation runs to input models so as to minimize the variance of the estimated posterior mean response subject to a budget constraint on the total amount of simulated experimentation or computer time that is available. Finally in Section 4.5, we conduct a Monte Carlo experiment on a computer communication network application to evaluate the performance of the BMA approach versus conventional techniques for estimating the posterior mean response and assessing its variability.

4.1 The BMA Approach

Assume that we have Q random inputs driving our simulation model. We observe the sample data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q)$, where $\boldsymbol{x}_q = (x_{q1}, \ldots, x_{qn_q})$ is the vector of observations based on a random sample of size n_q for the qth random input. Even though stochastic dependencies among simulation inputs will not affect our formulation, we assume for simplicity that these random inputs are independent. Let $M_{\ell q}$ represent the ℓth adequate model for the qth random input for $\ell = 1, \ldots, K_q$ and $q = 1, \ldots, Q$. In practice, most of the K_q's will be one, and only the random inputs with high model uncertainty and enough information to assess such an uncertainty will have K_q's larger than one. To simplify the notation, we define the set of candidate models $\mathcal{M}$ to consist of models $\{M_k : k = 1, \ldots, K\}$, where $K = K_1 \times \ldots \times K_Q$ is the total number of all different input model combinations each having prior probability of the form

$$p(M_k) \;=\; \prod_{q=1}^{Q} p(M_{\ell_{qk} q}), \text{ where } 1 \leq \ell_{qk} \leq K_q \text{ for } q = 1, \ldots, Q \text{ and } k = 1, \ldots, K.$$

Once $\mathcal{M}$ is chosen, we let θ_k denote the d_k-dimensional vector of parameters under model M_k with prior distribution $p(\theta_k | M_k)$, where $k = 1, \ldots, K$.

The choice of the alternative models in the set $\mathcal{M}$ is highly dependent on the specific application, but several general comments may be made in the simulation input modeling context.

- Measures based on frequentist computations, such as P-values in chi-squared goodness-of-fit testing, may be difficult to interpret and highly misleading (Berger and Delampady, 1987).

- Input modeling software packages such as ExpertFit (Law and McComas, 2000) usually have some built-in approach, which is a closely guarded trade secret and is therefore inaccessible to simulation practitioners, for selecting the best fitting model from a long list of well-known models. Analysis of nonnested models, however, is very difficult in a frequentist framework (Miller, 1990). So these choices should not be taken for granted, especially if more than one model seems to fit the data well.

- With the recent advances in Bayesian computations, we can use some new diagnostic tools for checking the goodness of fit, such as the Bayes P-value

(Gelman et al., 1995) and cross-validation predictive distributions (Gelfand, 1996).

- Estimating and predicting the simulation output response is the real goal in simulation, and Bayesian methods provide a natural approach to accounting for model uncertainty, in that they can keep all models under consideration.

With a finite set $\mathcal{M} = \{M_k : k = 1, \ldots, K\}$ of candidate models, the expected value of the output quantity of interest, y, is given by Theorem 4.1

Theorem 4.1 *If the simulation output response has the form (4.1), then*

$$E(y|\boldsymbol{X}) = \sum_{k=1}^{K} p(M_k|\boldsymbol{X}) \int \eta(M_k, \theta_k)\, p(\theta_k|\boldsymbol{X}, M_k)\, d\theta_k. \tag{4.3}$$

Proof. By conditioning on M_k and θ_k, we can write $E(y|\boldsymbol{X})$ as

$$E(y|\boldsymbol{X}) = \sum_{k=1}^{K} p(M_k|\boldsymbol{X}) \int E(y|\boldsymbol{X}, M_k, \theta_k)\, p(\theta_k|\boldsymbol{X}, M_k)\, d\theta_k.$$

Moreover, we have

$$\begin{aligned} E(y|\boldsymbol{X}, M_k, \theta) &= \int E(y(\boldsymbol{u}, M_k, \theta_k)|\boldsymbol{X}, \boldsymbol{u}, M_k, \theta_k)\, d\boldsymbol{u} \\ &= \int y(\boldsymbol{u}, M_k, \theta_k)\, d\boldsymbol{u} \\ &= \eta(M_k, \theta_k) \end{aligned}$$

using (4.2), from which the result follows. △

There are thus three ingredients for the implementation of the BMA approach to discrete-event simulations:

(a) The specification of the prior probabilities $\{p(M_k) : k = 1, \ldots, K\}$ over which model uncertainty is propagated, and the selection of the prior distributions $\{p(\theta_k|M_k) : k = 1, \ldots, K\}$ for the model parameters.

(b) The computation of the posterior distributions $\{p(\theta_k|\boldsymbol{X}, M_k) : k = 1, \ldots, K\}$.

(c) The computation of the posterior model probabilities $\{p(M_k|\boldsymbol{X}) : k = 1, \ldots, K\}$.

Each of these components is addressed in the subsections that follow.

4.1.1 Specification of Priors

The specification of the prior model probabilities $\{p(M_k)\}$ is typically context specific. When there is little prior information about the relative plausibility of the models considered, the assumption that all models are equally likely *a priori* is a reasonable choice (Madigan and Raftery, 1994). Different prior probabilities can be viewed as derived from previous data and representing the relative success of the models in predicting those previous data. Many researchers have provided a detailed analysis of the benefits of incorporating informative priors in many applications such as graphical models (Madigan and Raftery, 1994), Bayesian knowledge-based systems (Lauritzen et al., 1994), and linear models (Ibrahim and Laud, 1994).

The easiest way to deal with the problem of specifying the prior distributions $p(\theta_k|M_k)$ on the model parameters is to ignore them and simply use the Schwarz criterion (see Section 4.1.3). Although this will lead to appropriate conclusions in "sufficiently large" samples, there is not much available guidance on the operational meaning of the qualifying phrase "sufficiently large." Typically, these distributions are specified based on information accumulated from past studies, or from expert opinions. In order to simplify the subsequent computational burden, experimenters often limit this choice somewhat by restricting priors to some familiar distributional family. An even simpler alternative, available in some cases, is to endow the prior distribution with little information content, so that the data from the current study will be the dominant factor in determining the posterior distribution. We address each of these approaches in the subsequent sections.

Informative Priors

In the univariate case, the simplest approach to specifying a prior distribution $p(\theta_k|M_k)$ is first to limit consideration to a manageable collection of θ_k values deemed possible, and subsequently to assign probability masses to these values, reflecting the experimenter's prior beliefs as closely as possible, in such a way that their sum is 1. If θ_k is discrete-valued, such an approach may be quite natural, though perhaps quite time consuming. If θ_k is continuous, we must instead assign the masses to intervals on the real line, rather than to single points, resulting in a histogram prior for θ_k. Such a histogram may seem inappropriate, especially in concert with a continuous likelihood, but in fact may be more appropriate if the integrals required to compute the posterior distribution must be evaluated by a numerical quadrature scheme. Moreover, a

histogram prior may have as many bins (classes, cells) as the patience of the elicitee and the accuracy of his prior opinion will allow. It is important, however, that the range of the histogram should be sufficiently wide, since the support of the posterior will necessarily be a subset of that of the prior.

Alternatively, we might simply assume that the prior for θ_k belongs to a parametric distributional family $p(\theta_k|M_k, \nu_k)$, choosing ν_k so that the result matches the elicitee's true prior beliefs as nearly as possible. For example, if ν_k is two-dimensional, then specification of two moments (say, the mean and the variance) or two quantiles (say, the 50th and 95th percentiles) would be enough to determine its exact value. This matching approach (Berger, 1985) limits the effort required of the elicitee, and also overcomes the finite support problem inherent in the histogram approach. It may also lead to simplifications in the posterior computation.

Even when scientifically relevant prior information is available, elicitation of the precise forms for the prior distribution from experimenters can be a long and tedious process. However, prior elicitation issues tend to be application-specific, meaning that general-purpose algorithms are typically unavailable. Chaloner (1996) provides overviews of the various philosophies of prior distribution elicitation. The difficulty of prior elicitation has been ameliorated somewhat through the addition of interactive computing, especially dynamic graphics and object-oriented computer languages.

Conjugate Priors

In choosing a prior belonging to a specific distributional family, some choices may be more convenient computationally than others. In particular, it may be possible to select a member of that family which is conjugate to the likelihood—that is, one which leads to a posterior distribution belonging to the same distributional family as the prior. Morris (1983) showed that regular exponential families, from which we typically draw our likelihood functions, do in fact have conjugate priors, so that this approach will often be available in practice.

For multiparameter models, independent conjugate priors may often be specified for each parameter $\{\theta_{k,i} : i = 1, \ldots, d_k\}$, leading to corresponding conjugate forms for each conditional posterior distribution,

$$p(\theta_{k,i}|\boldsymbol{X}, M_k, \theta_{k,j}, j = 1, \ldots, d_k, j \neq i) \text{ for } i = 1, \ldots, d_k.$$

The ability of conjugate priors to produce at least unidimensional conditional posteriors in closed form enables them to retain their importance even in high-dimensional

settings. This occurs through the use of Markov Chain Monte Carlo integration techniques (Section 4.1.3), which construct a sample from the joint posterior by successively sampling from the individual conditional posterior distributions.

Finally, while a single conjugate prior may be inadequate to reflect available prior knowledge accurately, a finite mixture of conjugate priors may be sufficiently flexible while still enabling simplified posterior calculations.

Noninformative Priors

Often no reliable prior information concerning θ_k exists, or an inference based mainly on data is desired. Suppose we could find a distribution that contains no information about θ_k, in the sense that it does not favor one value of θ_k over another, but generates a proper posterior distribution. We might refer to such a distribution as a noninformative prior for θ_k, and argue that all of the information resulting in the posterior distribution must arise from the data—and hence all resulting inferences must be completely objective, rather than subjective. Such an approach is likely to be important if Bayesian methods are to compete successfully in practice with their popular frequentist counterparts such as maximum likelihood estimation. Kass and Wasserman (1996) provide an excellent review of noninformative priors.

Noninformative priors are described as vague or diffuse, and their densities are generally improper. Improper priors are defined only up to an undefined multiplicative constant. One complication with the BMA approach is that these priors cannot be used to compute the posterior model probabilities $p(M_k|\boldsymbol{X})$ even if the posterior distributions $p(\theta_k|\boldsymbol{X}, M_k)$ are proper. If the prior distribution $p(\theta_k|M_k)$ is improper, then the marginal distribution $p(\boldsymbol{X}|M_k)$ will also be improper and thus defined only up to a multiplicative constant c_k. The problem is how to calibrate these constants to compute the posterior model probabilities. Kass and Raftery (1995) discuss in detail the problem of using improper priors in model selection.

One solution to this problem is to consider part of the data, $\boldsymbol{X}^{\text{ts}}$, as a so-called *training sample*, and convert the improper prior $p(\theta_k|M_k)$ to a proper posterior distribution via

$$p(\theta_k|\boldsymbol{X}^{\text{ts}}, M_k) \;=\; \frac{p(\boldsymbol{X}^{\text{ts}}|M_k, \theta_k)\, p(\theta_k|M_k)}{p(\boldsymbol{X}^{\text{ts}}|M_k)},$$

where the training sample marginal density

$$p(\boldsymbol{X}^{\text{ts}}|M_k) \;=\; \int p(\boldsymbol{X}^{\text{ts}}|M_k, \theta_k)\, p(\theta_k|M_k)\, d\theta_k$$

is finite, but not necessarily integrable (proper). The idea is then to compute the posterior model probabilities (see Section 4.1.3) with the remainder of the data, $\boldsymbol{X}^{\text{rs}}$, using the marginal density,

$$p(\boldsymbol{X}^{\text{rs}}|M_k) \;\; = \;\; \int p(\boldsymbol{X}^{\text{rs}}|\boldsymbol{X}^{\text{ts}}, M_k, \theta_k)\, p(\theta_k|\boldsymbol{X}^{\text{ts}}, M_k)\, d\theta_k.$$

The idea of a training sample was introduced by Lempers (1971), and other implementations have been suggested more recently under the names partial Bayes factors (O'Hagan, 1991), fractional Bayes factors (O'Hagan, 1995), and intrinsic Bayes factors (Berger and Perrichi, 1996).

There are several methods proposed in the statistical literature to choose a training sample. One method proposed by Aitkin (1991) is to take the entire sample as a training sample in order to obtain $p(\theta_k|\boldsymbol{X}^{\text{ts}}, M_k)$. This double use of the data is of course not consistent with usual Bayesian logic. Another method is to use the *minimal* training sample (Berger and Perrichi, 1996), which is the smallest sample that satisfies

$$0 < p(\boldsymbol{X}^{\text{ts}}|M_k) < \infty \;\text{ for all }\; M_k.$$

Typically, a minimal training sample is one for which all parameters in all models are identifiable. Often it is a sample of size $\max_k\{d_k\}$, where d_k is the dimension of θ_k.

Another more complicated solution to the use of improper priors in model selection is to use the cross-validation predictive densities (Gelfand, 1996) to compute the posterior model probabilities. These densities usually exist even if the marginal densities do not. This approach is important if MCMC methods are used to compute the posterior probabilities under improper priors.

4.1.2 Computation of Posterior Parameter Distributions

The second ingredient for the implementation of the BMA approach is to compute the posterior distributions $\{p(\theta_k|\boldsymbol{X}, M_k),\;\; k = 1, \ldots, K\}$. For each model M_k, the joint probability density function $p(\boldsymbol{X}, \theta_k|M_k)$ can be written as the product of the prior density $p(\theta_k|M_k)$ and the sampling distribution $p(\boldsymbol{X}|M_k, \theta_k)$,

$$p(\boldsymbol{X}, \theta_k|M_k) \;\; = \;\; p(\theta_k|M_k)\, p(\boldsymbol{X}|M_k, \theta_k).$$

Conditioning on the known value of the data $\boldsymbol{X}$ and using Bayes' rule, we obtain the posterior density,

$$p(\theta_k|\boldsymbol{X}, M_k) \;=\; \frac{p(\boldsymbol{X}, \theta_k|M_k)}{p(\boldsymbol{X}|M_k)} \;=\; \frac{p(\boldsymbol{X}|M_k, \theta_k)\, p(\theta_k|M_k)}{p(\boldsymbol{X}|M_k)},$$

where $p(\boldsymbol{X}|M_k)$ is the marginal distribution of the data $\boldsymbol{X}$, given model M_k.

For some models, with a specific choice of a prior distribution such as a conjugate prior, the posterior distribution can easily be recognized from the unnormalized posterior density

$$p(\theta_k|\boldsymbol{X}, M_k) \propto p(\theta_k|M_k)\, p(\boldsymbol{X}|M_k, \theta_k).$$

This removes the burden of computing the normalizing constant $p(\boldsymbol{X}|M_k)$. However, we cannot limit the choices of priors to specific distributional families in all applications, and we will generally have some unnormalized densities that do not belong to any of the well-known distributions. As we shall see in the next section, all that we need for running simulation experiments is a random sample $\{\theta_k^r : r = 1, \ldots, R_k\}$ from the posterior distribution $p(\theta_k|\boldsymbol{X}, M_k)$. To generate a sample from the posterior distribution, we should compute the exact form of its density. This requires some high-dimensional numerical integrations or asymptotic approximations (see Section 4.1.3). These computational difficulties limited the use of Bayesian methods for more than two centuries. In the last decade, Markov Chain Monte Carlo (MCMC) methods are being increasingly used for dealing with such problems. The basic philosophy behind MCMC is to take a Bayesian approach and carry out the necessary numerical integrations using Monte Carlo simulation (see Gilks et al. (1996) for background). Instead of calculating exact or approximate estimates of the posterior density, this computer-intensive technique generates a stream of simulated values from the posterior distribution of any parameter or quantity of interest. These computations can be easily coded in the BUGS statistical package (Spiegelhalter et al., 1996) using a small set of BUGS commands. BUGS is a software that carries out Bayesian inference using a MCMC technique known as Gibbs sampling (Gelfand and Smith, 1990).

4.1.3 Computation of Posterior Model Probabilities

The posterior model probabilities $p(M_k|\boldsymbol{X})$ are computed as follows:

$$p(M_k|\boldsymbol{X}) \;=\; \frac{p(M_k)\, p(\boldsymbol{X}|M_k)}{\sum_{j=1}^{K} p(M_j)\, p(\boldsymbol{X}|M_j)}, \text{ for } k = 1, \ldots, K. \tag{4.4}$$

The evaluation of these probabilities comes down to computing the marginal data density given model M_k,

$$p(\boldsymbol{X}|M_k) \;=\; \int p(\boldsymbol{X}|M_k,\theta_k)\, p(\theta_k|M_k)\, d\theta_k,\;\; k=1,\ldots,K. \tag{4.5}$$

The integral in (4.5) may be evaluated analytically for distributions in the regular exponential family with conjugate priors. However, (4.5) is generally intractable and thus must be computed by numerical methods. An excellent review for the various numerical integration strategies is provided by Kass and Raftery (1995). Here, we provide a brief presentation of these methods.

Asymptotic Approximations

A useful approximation to the marginal density of the data as given by (4.5) is the Laplace method of approximation (Tierney and Kadane, 1986). It is obtained by assuming that the posterior density $p(\theta_k|\boldsymbol{X}, M_k)$, which is proportional to $p(\boldsymbol{X}|M_k,\theta_k)$ $\times$ $p(\theta_k|M_k)$, is highly peaked about its maximum $\tilde{\theta}_k$, which is the posterior mode. This is usually the case if the likelihood function $p(\boldsymbol{X}|M_k,\theta_k)$ is highly peaked near its maximum $\widehat{\theta}_k^{\text{mle}}$, which is the case for large samples.

Let $\tilde{l}_k(\theta_k) = \ln[p(\boldsymbol{X}|M_k,\theta_k)p(\theta_k|M_k)]$. Expanding $\tilde{l}_k(\theta_k)$ as a quadratic about $\tilde{\theta}_k$ and then exponentiating yields an approximation to $p(\boldsymbol{X}|M_k,\theta_k)$ $\times$ $p(\theta_k|M_k)$ that has the form of a normal density with mean $\tilde{\theta}_k$ and covariance matrix $\tilde{\Sigma}_k$ specified by

$$\left(\tilde{\Sigma}_k^{-1}\right)_{ij} \;=\; -\frac{\partial^2 \tilde{l}_k(\theta_k)}{\partial\theta_{ki}\partial\theta_{kj}}\bigg|_{\tilde{\theta}_k}.$$

Integrating this approximation gives the logarithm of the marginal density in (4.5) as follows

$$\begin{aligned}\ln[p(\boldsymbol{X}|M_k)] \;=\; & \frac{1}{2}d_k\ln(2\pi) + \frac{1}{2}\ln(|\tilde{\Sigma}_k|) + \ln[p(\boldsymbol{X}|M_k,\tilde{\theta}_k)] \\ & + \ln[p(\tilde{\theta}_k|M_k)] + O(n^{-1}),\end{aligned} \tag{4.6}$$

where d_k is the dimension of θ_k, and n is the sample size of the data set $\boldsymbol{X}$.

An important variant of the Laplace approximation is given by

$$\begin{aligned}\ln[p(\boldsymbol{X}|M_k)] \;=\; & \frac{1}{2}d_k\ln(2\pi) + \frac{1}{2}\ln(|\widehat{\Sigma}_k|) + \ln[p(\boldsymbol{X}|M_k,\widehat{\theta}_k^{\text{mle}})] \\ & + \ln[p(\widehat{\theta}_k^{\text{mle}}|M_k)] + O(n^{-1}),\end{aligned} \tag{4.7}$$

where $\widehat{\Sigma}_k$ is the inverse of the observed information matrix

$$(\mathcal{I}_k)_{ij} = -\frac{\partial^2}{\partial\theta_{ki}\partial\theta_{kj}}\ln[p(\boldsymbol{X}|M_k,\theta_k)]\Big|_{\hat{\theta}_k^{\mathrm{mle}}},$$

evaluated at the Maximum Likelihood Estimator (MLE),

$$\widehat{\theta}_k^{\mathrm{mle}} \equiv \arg\max_{\theta_k} p(\boldsymbol{X}|M_k,\theta_k).$$

Although the approximation (4.7) is likely to be less accurate than (4.6) when the prior is informative relative to the likelihood, it has the advantage of being easily computed from any statistical package.

Finally, it is possible to avoid the introduction of the prior densities $p(\theta_k|M_k)$ by using a simpler approximation (Schwarz, 1978) given by

$$\ln[p(\boldsymbol{X}|M_k)] = -\frac{1}{2}d_k\ln(n) + \ln[p(\boldsymbol{X}|M_k,\widehat{\theta}_k^{\mathrm{mle}})] + O(1), \tag{4.8}$$

where the term $-\frac{1}{2}d_k\ln(n)$ can be thought of as a penalty for models with a large number of parameters. This approximation is less accurate than the Laplace approximation, especially in small samples.

Simple Monte Carlo, Importance Sampling and Gaussian Quadrature

The simplest Monte Carlo integration estimate of $p(\boldsymbol{X}|M_k)$ is

$$\widehat{p}_1(\boldsymbol{X}|M_k) = \frac{1}{L}\sum_{l=1}^{L} p(\boldsymbol{X}|M_k,\theta_k^{(l)}), \tag{4.9}$$

where $\{\theta_k^{(l)} : l = 1,\ldots,L\}$ is a sample from the prior distribution. A major difficulty with (4.9) is that most of the $\theta_k^{(l)}$ have small likelihood values if the posterior is concentrated relative to the prior, so that the simulation will be quite inefficient.

The precision of simple Monte Carlo integration can be improved by importance sampling. This consists of generating a sample $\{\theta_k^{(l)} : l = 1,\ldots,L\}$ from a density $\pi^\star(\theta_k|M_k)$, known as the *importance sampling function*. Under quite general conditions, a simulation-based consistent estimator of $p(\boldsymbol{X}|M_k)$ is

$$\widehat{p}_2(\boldsymbol{X}|M_k) = \frac{\sum_{l=1}^{L} w_l p(\boldsymbol{X}|M_k,\theta_k^{(l)})}{L}, \tag{4.10}$$

where $w_l = p(\theta_k^{(l)}|M_k)/\pi^\star(\theta_k^{(l)}|M_k)$.

A more efficient scheme is based on adaptive Gaussian quadrature. Using well-established methods from the numerical analysis literature, Genz and Kass (1993) showed how integrals that are peaked around a dominant mode may be evaluated.

Markov Chain Monte Carlo Methods

Several methods are now available for simulating from posterior distributions. In the simplest case these include direct simulation and rejection sampling. In more complex cases, MCMC methods, particularly the Metropolis-Hastings algorithm (Chib and Greenberg, 1995) and the Gibbs sampler provide a general recipe. Another fairly general recipe is the weighted likelihood bootstrap (Newton and Raftery, 1994). Any of these methods gives us a sample approximately drawn from the posterior density $cq(\theta_k|\boldsymbol{X}, M_k)$. Then we can estimate (4.5) by

$$\widehat{p}_3(\boldsymbol{X}|M_k) \;=\; \frac{\frac{1}{L}\sum_{l=1}^{L} w_l p(\boldsymbol{X}|M_k, \theta_k^{(l)})}{\frac{1}{L}\sum_{l=1}^{L} w_l}, \tag{4.11}$$

where $w_l = p(\theta_k^{(l)}|M_k)/q(\theta_k^{(l)}|\boldsymbol{X}, M_k)$, and $\sum_{l=1}^{L} w_l/L$ is a simulation-consistent estimator of c. Now, substituting $cq(\theta_k|\boldsymbol{X}, M_k) = p(\boldsymbol{X}|M_k, \theta_k)p(\theta_k|M_k)/p(\boldsymbol{X}|M_k)$ into (4.11) yields an estimate for $p(\boldsymbol{X}|M_k)$,

$$\widehat{p}_4(\boldsymbol{X}|M_k) \;=\; \left\{ \frac{1}{L}\sum_{l=1}^{L} \left[p(\boldsymbol{X}|M_k, \theta_k^{(l)}) \right]^{-1} \right\}^{-1}, \tag{4.12}$$

the harmonic mean of the likelihood values. This converges almost surely to the correct value, $p(\boldsymbol{X}|M_k)$, as $L \to \infty$, but it does not generally satisfy a Gaussian Central Limit theorem (CLT). Simple modifications to (4.12) that satisfy the CLT are suggested in the literature by Meng and Weng (1993), Gelfand and Dey (1994), and Newton and Raftery (1994).

Finally, another simple estimator that performed well practice is the so called "Laplace-Metropolis" estimator of $p(\boldsymbol{X}|M_k)$, by Raftery (1996). It is obtained by using the posterior simulation output to estimate the quantities needed to compute the Laplace approximation (4.6), namely the posterior mode, $\tilde{\theta}_k$, and minus the inverse Hessian at the posterior mode, $\tilde{\Sigma}_k$.

4.2 Estimating the Output Mean Response

In our inference about the output quantity of interest, y, we focus on estimating its mean response given by equation (4.3) and assessing its variability. Chick (1999) proposed an algorithm for estimating the output mean response. He suggested that for the rth simulation run, we need to sample a model M^r from its discrete posterior probability mass function $\{p(M_k|\boldsymbol{X}) : k = 1, \ldots, K\}$, and then sample its vector of parameters $\theta^r_{M^r}$ from its posterior distribution $p(\theta_{M^r}|\boldsymbol{X}, M^r)$. The mean response estimate would be the average of all output responses $\{y_r : r = 1, \ldots, R\}$, computed using the randomly sampled input models $\{M^r : r = 1, \ldots, R\}$ and their corresponding randomly sampled parameter vectors $\{\theta^r_{M^r} : r = 1, \ldots, R\}$.

This algorithm gives a good estimate of the mean response for a large number of runs, but we believe that it has several deficiencies that makes it of little use to simulation practitioners. First, it is hard to estimate the variability of y due to each uncertainty factor. This quantification is valuable to the simulation analyst to improve the efficiency of the simulation design. Second, if we have a limited number of runs and some models may have small posterior weights compared to others, then a simple random sampling scheme will never allow the analyst to observe output responses from such models. Finally and most importantly, Chick's algorithm cannot accommodate more models without repeating all the simulation runs. If for some reason we decide to expand our summation in (4.3) to have more than K models, then we need to repeat all the runs with the new sampled models and their parameters to obtain a new estimate for the posterior mean response.

Figure 4.1 summarizes an algorithm which implements the BMA approach for designing simulation experiments and overcomes all the above deficiencies. The net effect of the algorithm is to account for the full extent of the model and parameter uncertainty as well as the usual stochastic uncertainty. The inner loop of the algorithm will be used to generate estimates for the stochastic uncertainty, whereas the middle loop will assess the parameter uncertainty for each input model. Finally, the outer loop will be used to estimate the model uncertainty.

Every model in the set $\mathcal{M}$ can have a different predictive inference on the output of interest y, and the composite inference will be a weighted average of the predictive distributions $p(y|\boldsymbol{X}, M_k)$ for $k = 1, \ldots, K$. This explains our idea of not resampling the input model M and its associated parameter vector θ_M prior to each replication. If

for $k = 1, \ldots, K$
    set input model $M \leftarrow M_k$
    for $r = 1, \ldots, R_k$
        generate the rth sample θ^r independently from $p(\theta|\boldsymbol{X}, M)$
        set the parameter vector $\theta \leftarrow \theta^r$
        for $j = 1, \ldots, m$
            set the random-number input $\boldsymbol{u} \leftarrow \boldsymbol{u}_j$ independently
            perform the jth simulation run using $\boldsymbol{u}$, M, and θ
            calculate the output response $y_{krj} = y(\boldsymbol{u}, M, \theta)$
        end loop
        compute $\overline{y}_{kr} = \sum_{j=1}^{m} y_{krj}/m$
    end loop
    compute the model grand mean $\overline{\overline{y}}_k = \sum_{r=1}^{R_k} \overline{y}_{kr}/R_k$
end loop
compute the weighted mean $\sum_{k=1}^{K} p(M_k|\boldsymbol{X})\overline{\overline{y}}_k$ as an estimate for $E(y|\boldsymbol{X})$

Figure 4.1: Simulation Replication Algorithm Based on Bayesian Model Averaging

the simulation model is costly in terms of computing time and the number of possible input models is large, then we can eliminate the models which explain the data far less than others using the Occam's Window method (Madigan and Raftery, 1994).

The total sample sizes $\{R_k : k = 1, \ldots, K\}$ respectively generated from the posterior distributions $\{p(\theta_k|\boldsymbol{X}, M_k) : k = 1, \ldots, K\}$ may not be necessarily the same, because the effect of parameter uncertainty on the variability of the output response, given model M_k, is usually different for different input models. Theoretically, the accuracy of our estimate of the posterior mean response improves as all the R_k's get large. However, we are usually restricted in practice by a fixed number of simulation runs N. In Section 4.4, we propose a method to specify the $\{R_k\}$ that optimally allocates the total simulation effort to the different models, based on the minimization of the variance of the estimator of the posterior mean response subject to a constraint on the total number of simulation runs N.

4.3 Assessing Output Variability

In this section, we try to assess the variability of the simulation output based on a decomposition of the posterior variance, $\text{Var}(y|\boldsymbol{X})$. To simplify the notation, we denote our posterior model probabilities as $p_k = p(M_k|\boldsymbol{X})$ for $k = 1, \ldots, K$. In view of (4.1), we see that the output response from the jth simulation run using input model M_k and the rth sample of that model's parameters θ_k^r can be written as

$$\begin{aligned} y_{krj} &= y(\boldsymbol{u}_j, M_k, \theta_k^r) \\ &= \eta(M_k, \theta_k^r) + e_j(\boldsymbol{u}_j, M_k, \theta_k^r);\ k = 1, \ldots, K;\ r = 1, \ldots, R_k;\ j = 1, \ldots, m, \end{aligned} \tag{4.13}$$

where the error variable e_j is the random difference between the simulation output response y_{krj} and $\eta(M_k, \theta_k^r)$. We generally assume that

$$E(e_j|M_k, \theta_k^r) = 0 \quad \text{and} \quad \text{Var}(e_j|M_k, \theta_k^r) = \tau_k^2, \tag{4.14}$$

so that

$$E(y_{krj}|M_k, \theta_k^r) = \eta(M_k, \theta_k^r).$$

Here we assume that τ_k^2 does not depend on θ_k^r because we are interested in obtaining a measure of the average variability in the output due to stochastic uncertainty. The effect of the randomness in θ_k^r will be captured instead by the randomness in $\eta(M_k, \theta_k^r)$, which will give a measure of the output variability due to parameter uncertainty. Moreover, given that our main objective is to estimate the overall mean response, we can further assume that

$$\eta(M_k, \theta_k^r) = \beta_k + \delta_{kr}(M_k, \theta_k^r), \tag{4.15}$$

where

$$\begin{aligned} \beta_k &= E_{\theta_k^r}\left[\eta(M_k, \theta_k^r)\right] \\ &= \int \eta(M_k, \theta_k) p(\theta_k|\boldsymbol{X}, M_k) d\theta_k \\ &= E(y|\boldsymbol{X}, M_k), \end{aligned}$$

using Theorem 4.1, and

$$E_{\theta_k^r}(\delta_{kr}|M_k) = 0 \quad \text{and} \quad \text{Var}_{\theta_k^r}(\delta_{kr}|M_k) = \sigma_k^2. \tag{4.16}$$

Based on these assumptions, we show in the following theorem that the posterior variance can be written as the sum of three variances measuring the model, parameter and stochastic uncertainty.

Theorem 4.2 *If (4.13), (4.14), (4.15), and (4.16) hold, then*

$$\mathrm{Var}(y|\boldsymbol{X}) = \sum_{k=1}^{K} p_k(\beta_k - \beta)^2 + \sum_{k=1}^{K} p_k\sigma_k^2 + \sum_{k=1}^{K} p_k\tau_k^2, \tag{4.17}$$

where

$$E\,(y|\boldsymbol{X}) = \sum_{k=1}^{K} p_k\,\beta_k = \beta. \tag{4.18}$$

Proof. Let M denote a model chosen at random from $\mathcal{M}$ according to the posterior distribution $p(M|\boldsymbol{X})$, and let θ_M denote its associated parameter vector. We have

$$\mathrm{Var}(y|\boldsymbol{X}) = \mathrm{Var}_M[E(y|\boldsymbol{X}, M)|\boldsymbol{X}] + E_M[\mathrm{Var}(y|\boldsymbol{X}, M)|\boldsymbol{X}], \tag{4.19}$$

$$E(y|\boldsymbol{X}, M) = E_{\theta_M}[E(y|\boldsymbol{X}, M, \theta_M)|\boldsymbol{X}, M], \tag{4.20}$$

and

$$\begin{aligned}\mathrm{Var}(y|\boldsymbol{X}, M) &= \mathrm{Var}_{\theta_M}[E(y|\boldsymbol{X}, M, \theta_M)|\boldsymbol{X}, M] \\ &\quad + E_{\theta_M}[\mathrm{Var}(y|\boldsymbol{X}, M, \theta_M)|\boldsymbol{X}, M].\end{aligned} \tag{4.21}$$

Substituting (4.20) and (4.21) into (4.19) gives

$$\begin{aligned}\mathrm{Var}(y|\boldsymbol{X}) &= \mathrm{Var}_M[E_{\theta_M}\{E(y|\boldsymbol{X}, M, \theta_M)|\boldsymbol{X}, M\}|\boldsymbol{X}] \\ &\quad + E_M[\mathrm{Var}_{\theta_M}\{E(y|\boldsymbol{X}, M, \theta_M)|\boldsymbol{X}, M\}|\boldsymbol{X}] \\ &\quad + E_M[E_{\theta_M}\{\mathrm{Var}(y|\boldsymbol{X}, M, \theta_M)|\boldsymbol{X}, M\}|\boldsymbol{X}] \\ &= V_{\mathrm{mod}} + V_{\mathrm{par}} + V_{\mathrm{sto}},\end{aligned} \tag{4.22}$$

where V_{mod}, V_{par}, and V_{sto} respectively measure the model, parameter and stochastic uncertainty in the simulation experiment. Now using (4.15) and (4.16), we have

$$E\,(y|\boldsymbol{X}) = E_M[E(y|\boldsymbol{X}, M)|\boldsymbol{X}] = \sum_{k=1}^{K} p_k\,\beta_k = \beta. \tag{4.23}$$

Then from (4.15), (4.16), and (4.23), we have

$$\begin{aligned} V_{\text{mod}} &= \text{Var}_M[E_{\theta_M}\{\eta(M,\theta_M)\}|\boldsymbol{X}] \\ &= \text{Var}_M[\beta_M|\boldsymbol{X}] \\ &= \sum_{k=1}^{K} p_k(\beta_k-\beta)^2, \end{aligned}$$

and

$$\begin{aligned} V_{\text{par}} &= E_M[\text{Var}_{\theta_M}\{\eta(M,\theta_M)\}|\boldsymbol{X}] \\ &= E_M[\text{Var}_{\theta_M}\{\beta_M+\delta(M,\theta_M)\}|\boldsymbol{X}] \\ &= E_M[\text{Var}_{\theta_M}\{\delta(M,\theta_M)\}|\boldsymbol{X}] \\ &= E_M[\sigma_M^2|\boldsymbol{X}] \\ &= \sum_{k=1}^{K} p_k\sigma_k^2. \end{aligned}$$

Finally, from (4.14) we have

$$\begin{aligned} V_{\text{sto}} &= E_M[E_{\theta_M}\{\text{Var}(\eta(M,\theta_M)+e(\boldsymbol{u},M,\theta_M)|\boldsymbol{X},M,\theta_M)|\boldsymbol{X},M\}|\boldsymbol{X}] \\ &= E_M[E_{\theta_M}\{\text{Var}(e(\boldsymbol{u},M,\theta_M)|\boldsymbol{X},M,\theta_M)|\boldsymbol{X},M\}|\boldsymbol{X}] \\ &= E_M[E_{\theta_M}\{\tau_M^2|\boldsymbol{X},M\}|\boldsymbol{X}] \\ &= E_M[\tau_M^2|\boldsymbol{X}] \\ &= \sum_{k=1}^{K} p_k\tau_k^2. \end{aligned}$$

△

The response surface model given by (4.13)–(4.16) for each input model M_k is known in the statistical literature as the classical random-effects model (Rao, 1997), where one estimates β_k, τ_k^2, and σ_k^2 from the output of the algorithm in Figure 4.1 as follows:

$$\widehat{\beta}_k = \overline{\overline{y}}_k, \tag{4.24}$$

$$\widehat{\tau}_k^2 = \frac{\sum_{r=1}^{R_k}\sum_{j=1}^{m}(y_{krj}-\overline{y}_{kr})^2}{R_k(m-1)}, \tag{4.25}$$

and

$$\widehat{\sigma}_k^2 = \frac{\sum_{r=1}^{R_k}(\overline{y}_{kr} - \overline{\overline{y}}_k)^2}{(R_k - 1)} - \frac{\widehat{\tau}_k^2}{m}. \tag{4.26}$$

From the above estimates, we can estimate the three variance components of Theorem 4.2 as

$$\widehat{V}_{\text{mod}} = \sum_{k=1}^{K} p_k(\widehat{\beta}_k - \widehat{\beta})^2, \tag{4.27}$$

$$\widehat{V}_{\text{par}} = \sum_{k=1}^{K} p_k\widehat{\sigma}_k^2, \tag{4.28}$$

and

$$\widehat{V}_{\text{sto}} = \sum_{k=1}^{K} p_k\widehat{\tau}_k^2, \tag{4.29}$$

where

$$\widehat{\beta} = \sum_{k=1}^{K} p_k\widehat{\beta}_k. \tag{4.30}$$

In addition to point estimates, we can also construct a posterior confidence interval for the overall mean response β from the output of the Simulation Replication Algorithm given in Figure 4.1. We propose two methods for interval estimation.

The first method is based on the percentile method (Efron and Tibshirani, 1993), which assumes that the sample sizes R_k from the posterior distributions of the model parameters are the same (i.e. $R_k = R$ for $k = 1, \ldots, K$). We can then deliver an approximate $100(1-\alpha)\%$ percentile interval for β as

$$[\beta_{\text{L}}, \beta_{\text{U}}] \approx [\overline{y}_{(\lceil R\alpha/2 \rceil)}, \overline{y}_{(\lceil R(1-\alpha/2) \rceil)}], \tag{4.31}$$

where the quantities

$$\overline{y}_{(1)} \leq \overline{y}_{(2)} \leq \cdots \leq \overline{y}_{(R)}$$

denote the order statistics of the $\{\overline{y}_r : r = 1, \ldots, R\}$ defined as

$$\overline{y}_r = \sum_{k=1}^{K} p_k\overline{y}_{kr} \text{ for } r = 1, \ldots, R,$$

and $\overline{y}_{kr}$'s are specified in Figure 4.1.

In Subsection 4.4.3, we present a second method for constructing an approximate $100(1-\alpha)\%$ confidence interval for the posterior mean response under any scheme for allocating the sample sizes $\{R_k : k = 1, \ldots, K\}$ among the input models.

4.4 Replication Allocation Procedures

We describe in this section two methods to determine the sample sizes $\{R_k : k = 1, \ldots, K\}$ respectively allocated to the models $\{M_k : k = 1, \ldots, K\}$ based on the practical assumption that the total computational effort is generally limited by a fixed number of simulation replications N. We assume further that the stochastic variability can be assessed by a small number of replications m that are fixed prior to the simulation experiment.

The first replication allocation method is based on minimizing the variance of our posterior mean response estimate. Assuming that all the simulation replications are independent, we have the following result.

Theorem 4.3 *If (4.13), (4.14), (4.15), and (4.16) hold, then the Simulation Replication Algorithm of Figure 4.1 yields*

$$\mathrm{Var}(\widehat{\beta}) \;=\; \sum_{k=1}^{K} p_k^2 \frac{\sigma_k^2}{R_k} + \sum_{k=1}^{K} p_k^2 \frac{\tau_k^2}{m R_k}. \tag{4.32}$$

Proof. Given that our mean response estimate is $\widehat{\beta} = \sum_{k=1}^{K} p_k \overline{\overline{y}}_k$ as specified in Figure 4.1, we have

$$\mathrm{Var}(\widehat{\beta}) \;=\; \sum_{k=1}^{K} p_k^2 \,\mathrm{Var}(\overline{\overline{y}}_k). \tag{4.33}$$

Since in Figure 4.1 the parameters are sampled independently on each replication of each input model, we can write

$$\mathrm{Var}(\overline{\overline{y}}_k) \;=\; \frac{\mathrm{Var}(\overline{y}_{kr})}{R_k}. \tag{4.34}$$

Moreover, we have

$$\mathrm{Var}(\overline{y}_{kr}) \;=\; \sigma_k^2 + \frac{\tau_k^2}{m} \tag{4.35}$$

using (4.14) and (4.16) since the algorithm in Figure 4.1 requires that the m simulation runs are independent, given a model and its sampled parameter. Finally substituting (4.34) and (4.35) into (4.33), we obtain the desired result. $\triangle$

4.4.1 Optimal Allocation Procedure

To minimize the variance (4.32) subject to a budget constraint on the total number of runs, we must solve the following optimization problem,

$$\left.\begin{array}{ll} \min\limits_{\{R_k:1\leq k\leq K\}} & \sum\limits_{k=1}^{K} \dfrac{p_k^2}{R_k}\left[\sigma_k^2+\tau_k^2/m\right] \\ \\ \text{subject to:} & \sum\limits_{k=1}^{K} R_k = N/m = N'. \end{array}\right\} \tag{4.36}$$

We reformulate (4.36) as an unconstrained optimization problem using the method of Lagrange multipliers to show that, modulo rounding, the optimal sample sizes are given by

$$R_k^* \;=\; \frac{N' p_k \sqrt{\vartheta_k}}{\sum_{i=1}^{K} p_i \sqrt{\vartheta_i}} \text{ for } k = 1, \ldots, K, \tag{4.37}$$

where

$$\vartheta_k \;=\; \sigma_k^2 + \tau_k^2/m \text{ for } k = 1, \ldots, K. \tag{4.38}$$

Note that R_k^* depends on the ϑ_k values, which are unknown and usually estimated after observing the actual output responses. We suggest a two-phase replication allocation procedure that exploits the above result. In the first phase, we can make a small, equal number of pilot runs at each model M_k; and then for $k = 1, \ldots, K$, we estimate ϑ_k by $\widehat{\vartheta}_k = \widehat{\sigma}_k^2 + \widehat{\tau}_k^2/m$, where $\widehat{\tau}_k^2$ and $\widehat{\sigma}_k^2$ are estimated using (4.25) and (4.26), respectively. In the second phase, we allocate the rest of the runs according to (4.37). Assuming that the variance estimates are constant from phase one to phase two, the two-phase replication allocation procedure delivers a smaller variance for the mean response compared to the equal allocation scheme

$$R_k \;=\; \frac{N'}{K} \text{ for } k = 1, \ldots, K, \tag{4.39}$$

used to construct the percentile-type confidence interval (4.31) on the posterior mean response.

4.4.2 Proportional Allocation Procedure

One feasible solution to the optimization problem (4.36) is the proportional allocation procedure

$$R_k \;=\; p_k N \text{ for } k = 1, \ldots, K, \tag{4.40}$$

which is also optimal if all the ϑ_k's in (4.39) happen to be equal. This allocation scheme can be easily implemented prior to making the simulation runs, and it overcomes the problem of having to estimate the variances in the optimal allocation procedure (4.37). Moreover, the mean estimator $\widehat{\beta}_{\text{pa}}$, computed from the Simulation Replication Algorithm given in Figure 4.1 and the allocation scheme (4.40), has a smaller variance compared to the mean estimator $\widehat{\beta}_{\text{srs}}$ computed using the Simple Random Sampling (SRS) procedure. The SRS procedure is similar to Chick's (1999) approach described in Section 4.2, where we randomly sample a new input model and its vector of parameters from their posterior distributions prior to each run, and then perform m independent runs for a total of N' runs. We formally state and prove this result in the following theorem.

Theorem 4.4 *If (4.13), (4.14), (4.15), and (4.16) hold, then with the proportional allocation scheme (4.40) we obtain the following reduction in variance of the posterior mean estimator versus simple random sampling:*

$$\text{Var}(\widehat{\beta}_{\text{srs}}) - \text{Var}(\widehat{\beta}_{\text{pa}}) \;=\; \frac{1}{N'} \sum_{k=1}^{K} p_k (\beta_k - \beta)^2 > 0. \tag{4.41}$$

Proof. Let y_{lj} denote the output response on the jth simulation run computed using the lth randomly sampled input model and its randomly sampled vector of parameters. Then, the mean estimator under the SRS procedure is given by

$$\widehat{\beta}_{\text{srs}} \;=\; \frac{1}{N'} \sum_{l=1}^{N'} \overline{y}_l,$$

where

$$\overline{y}_l \;=\; \frac{1}{m} \sum_{j=1}^{m} y_{lj} \text{ for } l = 1, \ldots, N'.$$

From Theorem 4.2 of Section 4.3, the variance of $\widehat{\beta}_{\text{srs}}$ is given by

$$\begin{aligned}\text{Var}(\widehat{\beta}_{\text{srs}}) &= \frac{1}{N'}\sum_{k=1}^{K} p_k(\beta_k-\beta)^2 + \frac{1}{N'}\sum_{k=1}^{K} p_k\sigma_k^2 + \frac{1}{N'}\sum_{k=1}^{K} p_k\frac{\tau_k^2}{m} \\ &= \frac{1}{N'}\sum_{k=1}^{K} p_k(\beta_k-\beta)^2 + \frac{1}{N'}\sum_{k=1}^{K} p_k\sigma_k^2 + \frac{1}{N}\sum_{k=1}^{K} p_k\tau_k^2. \qquad (4.42)\end{aligned}$$

To obtain the variance of $\widehat{\beta}_{\text{pa}}$, we substitute $R_k = p_k N'$ into equation (4.32). This yields

$$\begin{aligned}\text{Var}(\widehat{\beta}_{\text{pa}}) &= \sum_{k=1}^{K} p_k^2\frac{\sigma_k^2}{p_k N'} + \sum_{k=1}^{K} p_k^2\frac{\tau_k^2}{m p_k N'} \\ &= \frac{1}{N'}\sum_{k=1}^{K} p_k\sigma_k^2 + \frac{1}{N}\sum_{k=1}^{K} p_k\tau_k^2. \qquad (4.43)\end{aligned}$$

Finally, using equations (4.42) and (4.43) we obtain the desired result. △

4.4.3 Confidence Interval for the Posterior Mean with Any Allocation Procedure

In this subsection we derive a t-type confidence interval for β based on estimating the variance of our posterior mean estimate $\widehat{\beta} = \sum_{k=1}^{K} p_k\overline{\overline{y}}_k$. This interval does not assume that the sample sizes $\{R_k\}$ are equal as in the percentile-type confidence interval (4.31), and thus can be used with any of the allocation schemes described in the above subsections. We proved in Theorem 4.3 that the variance of our mean estimate is given by

$$\text{Var}(\widehat{\beta}) = \sum_{k=1}^{K} p_k^2\frac{\sigma_k^2}{R_k} + \sum_{k=1}^{K} p_k^2\frac{\tau_k^2}{mR_k},$$

so that we have the following estimator for $\text{Var}(\widehat{\beta})$

$$\begin{aligned}\widehat{\text{Var}}(\widehat{\beta}) &= \sum_{k=1}^{K} p_k^2\left(\frac{\widehat{\sigma}_k^2}{R_k} + \frac{\widehat{\tau}_k^2}{mR_k}\right), \\ &= \sum_{k=1}^{K} p_k^2\widehat{\mathbb{V}}_k, \qquad (4.44)\end{aligned}$$

where

$$\widehat{\mathbb{V}}_k = \frac{\sum_{r=1}^{R_k}(\overline{y}_{kr} - \overline{\overline{y}}_k)^2}{R_k(R_k-1)} \text{ for } k = 1, \ldots, K,$$

using equations (4.25) and (4.26). Assuming that $\{R_k : k = 1, \ldots, K\}$ are fixed quantities, we can use the approximation of Satterthwaite (1946), who showed that the complex variance estimator (4.44) has a distribution that is approximately chi-squared with "effective" degrees of freedom given by

$$f_{\text{eff}} = \left\lfloor \frac{\left(\sum_{k=1}^{K} p_k^2 \widehat{\mathbb{V}}_k\right)^2}{\sum_{k=1}^{K} \frac{p_k^4 \widehat{\mathbb{V}}_k^2}{(R_k-1)}} \right\rfloor .$$

Thus an approximate $100(1-\alpha)\%$ confidence interval for $\beta = \sum_{k=1}^{K} p_k \beta_k$ is

$$\sum_{k=1}^{K} p_k \overline{\overline{y}}_k \pm t_{1-\alpha/2, f_{\text{eff}}} \left(\sum_{k=1}^{K} p_k^2 \widehat{\mathbb{V}}_k\right)^{1/2}. \tag{4.45}$$

We will use (4.45) to evaluate the performance of the proportional allocation procedure (4.40) and the optimal allocation procedure (4.37) empirically using a Monte Carlo experiment of a computer communication network.

4.5 Application to a Computer Communication Network

4.5.1 Description

In this example we consider a simulation of a computer communications network (Kleinrock, 1976). It is a collection of $\mathbb{Q}$ nodes consisting of computing resources which communicate with each other along a set of $\mathbb{L}$ links (the data communication channels). The aim of the simulation study is to measure the delay in messages transmitted between nodes via the communication channels. Figure 4.2 illustrates a network with $\mathbb{Q} = 4$ and $\mathbb{L} = 4$.

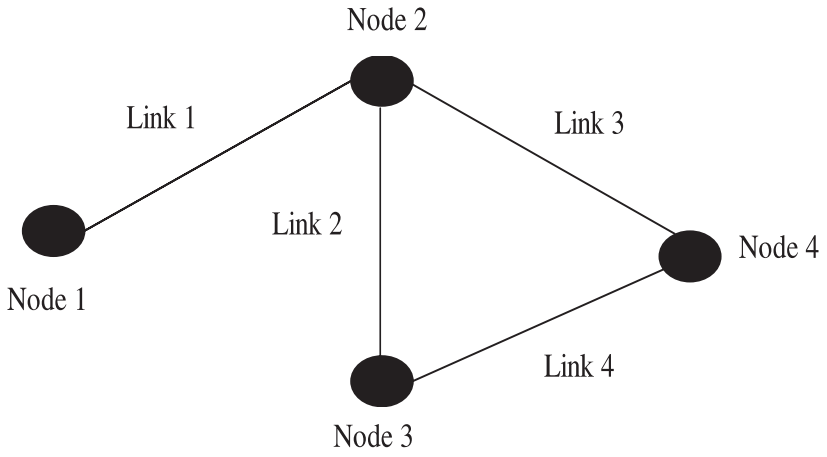

Figure 4.2: A Communication Network with $\mathbb{Q} = 4$ nodes and $\mathbb{L} = 4$ links

The $\mathbb{L}$ communication channels are assumed to be noiseless, and have a capacity of C_i bits per second for the ith channel. The $\mathbb{Q}$ nodes carry out the administration tasks such as message reassembly and routing. It is assumed that the nodal processing times are constant with value $\mathcal{T}_i$ for the ith node. In addition there are channel queueing and transmission delays. Traffic entering the network from any node forms a Poisson process with rate $\gamma(i, j)$ (messages per second) for those messages originating at node i and destined for node j. All messages are assumed to have length X that follows a mixture distribution given by

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \pi_3 f_3(x) \text{ for all } x, \tag{4.46}$$

where

$$\begin{aligned} f_1(x) &= \lambda e^{-\lambda x} \quad (x \geq 0) && \text{(Exponential}(\lambda)\text{)}, \\ f_2(x) &= \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\log x - \mu)^2/2\sigma^2} && \text{(Lognormal}(\mu, \sigma)\text{)}, \\ f_3(x) &= \frac{1}{b-a} \quad (a \leq x \leq b) && \text{(Uniform}(a, b)\text{)}. \end{aligned}$$

We assume that all nodes have unlimited storage capacity and that all messages are directed through the network on fixed paths. In high speed networks spanning large geographical regions, it may be important to include the propagation time H_i, which is the time required for the energy representing a single bit to propagate down

the length of the ith channel. The speed of energy propagation, v miles per second, is a significant fraction of the speed of light depending on the particular type of channel used. If the ith channel has length l_i miles, then $H_i = l_i/v$. Thus if a message has X bits then the time it occupies the ith channel will be $H_i + X/C_i$ seconds.

Some of the parameters in the network were known exactly: $\mathcal{T}_i = 0.001$ seconds ($i = 1, \ldots, \mathbb{Q}$), $C_i = 275,000$ bits/second ($i = 1, \ldots, \mathbb{Q}$), $l_i = i \times 100$ miles ($i = 1, \ldots, \mathbb{L}$), and $v = 150,000$ miles/second. The traffic arrival rates were: $\gamma(1,2) = 60$, $\gamma(1,3) = 40$, $\gamma(1,4) = 50$, $\gamma(2,1) = 80$, $\gamma(2,3) = 65$, $\gamma(2,4) = 20$, $\gamma(3,1) = 100$, $\gamma(3,2) = 22$, $\gamma(3,4) = 26$, $\gamma(4,1) = 40$, $\gamma(4,2) = 50$, $\gamma(4,3) = 60$. In Chapter 3, we assumed that these traffic rates were unknown in the simulation, and we only observe data samples from the exponential distribution. We have illustrated that accounting for parameter uncertainty in the arrival rates using the Bayesian approach delivered a better coverage probability with shorter intervals compared to the δ and bootstrap methods. In this chapter, we assume that these parameters can be estimated to a high degree of accuracy. We focus instead on the model uncertainty in the message lengths due to the multimodal complicated structure of their true sampling distribution given by equation (4.46). The parameters of the mixture distribution are: $\pi_{0,1} = 0.6$, $\lambda_0 = 1/300$, $\pi_{0,2} = 0.3$, $\mu_0 = 5.46$, $\sigma_0 = 0.7$, $\pi_{0,3} = 0.1$, $a_0 = 290$, $b_0 = 310$.

4.5.2 BMA Analysis

The true distribution of the message lengths was unknown in the simulation, and only data samples of size $n = 1000$ were observed. Three models were entertained for the (assumed independent) message lengths, $\{X_i : i = 1, \ldots, n\}$:

$$
\begin{aligned}
M_1 :\ p(x_i|M_1, \lambda_1) &= \lambda_1 e^{-\lambda_1 x_i} \quad (x_i \geq 0) && (\text{Exponential}(\lambda_1)),\\
M_2 :\ p(x_i|M_2, \mu_2, \sigma_2) &= \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(x_i-\mu_2)^2/2\sigma_2^2} && (\text{Normal}(\mu_2, \sigma_2)),\\
M_3 :\ p(x_i|M_3, \mu_3, \sigma_3) &= \frac{1}{\sqrt{2\pi}\sigma_3 x_i} e^{-(\log x_i-\mu_3)^2/2\sigma_3^2} \quad (x_i \geq 0) && (\text{Lognormal}(\mu_3, \sigma_3)).
\end{aligned}
$$

We chose the above models because they appear often in simulation applications, and they exist in all input modeling and simulation software systems. Moreover, the posterior model probabilities can be computed analytically which otherwise may be time-consuming to do using MCMC methods for each data set. Note that these three models are not nested, so that model comparison may not be conclusive in

a classical framework. They also can represent very different behavior in terms of message lengths. Finally, we assume that we do not have any prior information to favor one model over the other and assign equal probabilities to all candidate models (i.e. $p(M_k) = 1/3$ for $k = 1, 2, 3$).

To construct priors on the model parameters, we generate a training sample of size $T = 100$ from the true sampling distribution (4.46). We denote by $\boldsymbol{z} = \{z_1, \ldots, z_T\}$ the observations in the training sample, and $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ the observations in the data sample.

For M_1, the standard noninformative prior is $p(\lambda_1|M_1) = 1/\lambda_1$. This yields a Gamma posterior distribution with shape parameter T and scale parameter $1/(\sum_{t=1}^{T} z_t)$. The marginal density of the data $\boldsymbol{x}$ given the exponential model M_1 is

$$p(\boldsymbol{x}|M_1) = \frac{\Gamma(n+T)}{\Gamma(T)} \frac{(\sum_{t=1}^{T} z_t)^T}{(\sum_{t=1}^{T} z_t + \sum_{i=1}^{n} x_i)^{n+T}}.$$

Although the support of the normal distribution is the entire line, and message lengths are nonnegative, the parameters of the normal distribution will rarely give random variates that are negative. We will reject any negative values generated during the simulation. We discuss here the nontruncated normal distribution to simplify the exposition. The standard noninformative prior for M_2 is $p(\mu_2, \sigma_2^2) = 1/\sigma_2^2$. This yields an inverse-gamma posterior distribution for σ_2^2 with shape parameter $(T-1)/2$ and scale parameter $\sum_{t=1}^{T}(z_t - \overline{z})^2/2$, and a generalized student-t distribution for μ_2, having the following density

$$p(\mu_2|\boldsymbol{z}, M_2) = \frac{\Gamma(T/2)\sqrt{T}}{\Gamma((T-1)/2)\sqrt{(T-1)\pi}S_z} \left(1 + \frac{T}{T-1}\left(\frac{\mu_2 - \overline{z}}{S_z}\right)^2\right)^{-T/2},$$

where $\overline{z} = \sum_{t=1}^{T} z_t/T$ and $S_z^2 = \sum_{t=1}^{T}(z_t - \overline{z})^2/(n-1)$. The marginal density of the data $\boldsymbol{x}$ given the normal model M_2 is

$$p(\boldsymbol{x}|M_2) = \frac{\pi^{-n/2}(T-1)^{(T-1)/2}\Gamma((n+T-1)/2)S_x^{(T-1)/2}}{(T+n)^{-1/2}(n+T-1)^{(n+T-1)/2}\Gamma((T-1)/2)S^{(n+T-1)/2}},$$

where

$$S^2 = \frac{T-1}{n+T-1}S_z^2 + \frac{n-1}{n+T-1}S_x^2 + \frac{nT}{(n+T-1)(n+T)}(\overline{x} - \overline{z})^2.$$

The analysis of the lognormal model is similar to that of the normal model, by working with the logarithm of the data observations instead of the original observations.

4.5.3 Simulation Design

A basic simulation run was of 50 seconds in length and this was repeated $m = 10$ times for each model and its sampled parameters. The sample sizes $\{R_k : \ k = 1, 2, 3\}$ generated from the posterior distribution of the parameters were taken to be 100. In practice larger values of R_k are recommended, typically 1000. However, $R_k = 100$ is sufficiently large in this case because the observed coverages are stable, indicating the satisfactory behavior of the method. In each case the experiment was repeated 200 times so that 200 credible intervals were generated.

The "true" value β_0 of the average delay of a message in this communication network cannot be computed analytically. So we used a preliminary Monte Carlo experiment involving direct simulation of the network to compute β_0 to within $\pm 0.05\%$ of its true value with 99% confidence. For a fixed number of replications m, let $\overline{y}(m)$ and $S_y(m)$ denote the corresponding sample mean and standard deviation of the observed average message delays. Given prespecified values of the percentage error tolerance ω, the confidence coefficient α, and the preliminary sample size m_0, we determined the final number of replications according to the following relative-precision stopping rule (Law et al., 1981)

$$m^* = \min \Bigg\{ m : \ m \geq m_0, \ m = 0 \ (\text{mod } 10), \ S_y(m) > 0, \\ \text{and } t_{1-\alpha/2}(m-1) \frac{S_y(m)}{\sqrt{m}} \leq \omega |\overline{y}(m)| \Bigg\}, \tag{4.47}$$

and β_0 was taken to be $\overline{y}(m)$. In our preliminary experiment, we took $m_0 = 500$, $\omega = 0.0005$, and $\alpha = 0.01$. The final sample size m^* was found to be 1060 with a final estimate of β_0 being 0.006585.

4.5.4 Simulation Results

Table 4.1 summarizes the results of the BMA analysis. For each candidate model (M_1 : Exponential, M_2 : Normal, M_3 : Lognormal), we show the average posterior probability $p(M_k|\boldsymbol{x})$ over all the Monte Carlo experiments. We also present for each model M_k the mean estimate $\widehat{\beta}_k$, the stochastic variance estimate $\widehat{\tau}_k^2$, and the parameter variance estimate $\widehat{\sigma}_k^2$ of the average delay of messages in the network. In terms of posterior probabilities, the exponential model is the least favorite, but we cannot

really favor the lognormal model over the normal model. Note also that in terms of mean response estimates, the behavior of the lognormal model is completely different from the other two models. This is a situation where model uncertainty is the dominating uncertainty factor since it acccounts for about 99% of the overall uncertainty, so that a simulation analyst can have a completely different response choosing apriori one model over the other.

To study the effect of model uncertainty, we analyzed three different approaches of model selection: classical, partial Bayes, and BMA. In the classical approach, we made the simulation runs at a fixed model and a fixed parameter estimated using MLE. In the partial Bayes approach, we fixed the model but we accounted for parameter uncertainty by resampling the parameters prior to each set of m simulation runs. Finally in the BMA approach, we used the algorithm of Figure 4.1 to account for both model and parameter uncertainty. For the BMA approach, we considered three replication allocation procedures. The first procedure allocates the same number of runs to each model; the second procedure uses the Proportional Allocation Procedure (PAP) given in (4.40); and the third procedure uses the Optimal Allocation Procedure (OAP) given in (4.36). To find the optimal allocations of the simulation runs to models, we used the final variance estimates of the equal allocation scheme and we limited our computing effort to the total number of replications in a single Monte Carlo experiment.

Table 4.2 shows the performance of the mean estimate of the message delay in the communication network in terms of the Absolute Percentage Error $100|\widehat{\beta}-\beta_0|/\beta_0$ and the Mean Square Error $E[(\widehat{\beta}-\beta_0)^2]$. As expected, the classical and partial Bayes approaches show almost similar performance because of the small number of unknown parameters in the network and our choice of noninformative priors. However, both of these approaches show extremely poor performance compared to the BMA approach. The mean estimate of the BMA approach is very close to the target mean having less than 2% absolute percentage error and negligible mean square error. The optimal allocation procedure delivered the most precise mean estimate showing almost 50% reduction in mean square error compared to the equal allocation procedure. However, the performance of the proportional allocation procedure was almost as good as the optimal one. This suggests that the proportional scheme may be more applicable in practice given its simplicity.

In addition to point estimation, we studied the performance of the different ap-

proaches in terms of interval estimation. Table 4.3 summarizes the nominal 90% confidence interval lengths and coverage probabilities. Although the classical and partial Bayes approaches have much tighter confidence bands, they have zero coverage probabilities. This shows that their intervals are built around the wrong expected mean response. The BMA approach on the other hand has a much higher coverage probability, at a reasonable length, and centered at the right mean response. The replication allocation procedures deliver intervals with a higher coverage probability compared to the equal allocation BMA approach. This justifies the benefits of running more replications at models with higher posterior probability and smaller output response variance.

We conclude this analysis with some practical advice on the important issue of model expansion. In many simulation experiments, substantive knowledge on the output quantities of interest may exist to validate the simulation results. Discrepancies should be used to suggest possible expansions of the input models to improve the output inference. For example although the BMA approach showed great success in estimating the posterior mean response in our Monte Carlo experiment, the coverage probability was almost 10% lower than its nominal value. One remodelling approach suggested by Bernardo and Smith (1994) is to expand the most likely model by embedding it in a larger model, and then repeat the BMA analysis. In our experiment, the lognormal distribution was the most adequate model so that a mixture of two lognormal distributions can be a good choice to add to our set of candidate models.

Table 4.1: Posterior Probability, Mean and Variance Estimates for Each Candidate Model of Message Lengths in the Communication Network of Figure 4.2

Model	Post. Prob. $p(M_k\|\boldsymbol{x})$	Mean $\widehat{\beta}_k$	Stochastic Var. $\widehat{\tau}_k^2$	Parameter Var. $\widehat{\sigma}_k^2$
Exp.	1.23E–01	8.19E–03	6.20E–09	4.58E–08
Norm.	3.96E–01	8.57E–03	7.17E–09	2.30E–08
Logn.	4.81E–01	4.55E–03	1.08E–10	1.08E–11

Table 4.2: Absolute Percentage Error (APE) and Mean Square Error (MSE) for the Mean Estimator of Average Message Delay in the Communication Network of Figure 4.2

Approach	Model	Mean $\widehat{\beta}$	APE $100\lvert\widehat{\beta}-\beta_0\rvert/\beta_0$	MSE $E[(\widehat{\beta}-\beta_0)^2]$	SE(MSE)
Classical	Exp.	8.18E–03	24.16	2.57E–06	4.13E–08
	Norm.	8.57E–03	30.04	3.98E–06	7.22E–08
	Logn.	4.55E–03	30.98	4.17E–06	9.27E–10
Partial Bayes	Exp.	8.19E–03	24.35	2.61E–06	4.18E–08
	Norm.	8.57E–03	30.12	4.00E–06	7.35E–08
	Logn.	4.55E–03	30.86	4.17E–06	1.19E–10
BMA	Mix.	6.60E–03	1.78	2.00E–08	1.66E–09
BMA + PAP	Mix.	6.59E–03	1.35	1.20E–08	1.11E–09
BMA + OAP	Mix.	6.59E–03	1.35	1.20E–08	1.10E–09

Table 4.3: Performance of Nominal 90% Confidence Interval for the Average Message Delay in the Communication Network of Figure 4.2

Approach	Model	CIL	CV(CIL)	Coverage
Classical	Exp.	8.32E–05	2.62E–01	0
	Norm.	9.40E–05	2.93E–01	0
	Logn.	1.21E–05	2.29E–01	0
Partial Bayes	Exp.	6.90E–04	1.19E–01	0
	Norm.	4.89E–04	1.48E–01	0
	Logn.	1.05E–05	8.32E–02	0
BMA	Mix.	2.57E–04	1.31E–01	75
BMA + PAP	Mix.	2.89E–04	1.16E–01	81
BMA + OAP	Mix.	2.87E–04	1.13E–01	82

Chapter 5

Conclusions and Recommendations

This research addresses the input model selection problem for simulations of stochastic systems. We adopt a Bayesian approach which incorporates prior information into the model selection process in a formal and rigorous manner. An inference on the output quantities of interest can then be made that accounts for model, parameter, and stochastic uncertainty in a formally justifiable way. In this chapter, we summarize the main conclusions of this work, and we recommend some undertakings for future research.

5.1 Conclusions

Input models provide the driving force for a simulation model. Even if the model structure is valid, if the input modeling process is not carefully addressed, the output inference will be misleading when used for decision making. In Chapter 2, we reviewed the conventional approach to simulation input modeling. This approach rests on the use of classical statistical techniques in parameter estimation and goodness of fit testing to specify a single plausible "best" choice M^* for the input model M together with a unique estimate $\widehat{\theta}_{M^*}$ for its vector of parameters θ_M; and then we proceed as if M^* and $\widehat{\theta}_{M^*}$ were known to be correct. In general this approach fails to assess and propagate the model and parameter uncertainty fully, and it may lead to miscalibrated uncertainty assessments about the output quantity of interest y.

In Chapter 3, we presented three methods to account for the parameter uncertainty as well as the stochastic uncertainty. The first method is the δ-method, which de-

composes the variance of the simulation output into two distinct terms measuring the parameter and stochastic uncertainty, respectively. The problem with the δ-method is that certain sensitivity coefficients which can bias the variance estimate have to be estimated, and the effort needed to do this increases linearly with the number of unknown parameters. The second method we presented is the bootstrap method which is computationally more expensive, but does not suffer from the difficulties of the δ-method. Both of these methods, however, rely on large-sample approximations to show that the output response variance can be decomposed as the sum of two variances. They also rely solely on the data observations, and provide no mechanism for incorporating other types of subjective information into the modeling process.

To avoid the above difficulties, we have proposed a Bayesian approach that uses prior information and data observations for inferences on the distribution of a simulation-generated output response. We have derived an expression for the posterior mean, and developed a Bayesian Simulation Replication Algorithm to compute point and credible (confidence) interval estimators for the posterior mean response. We have also developed two response surface models to decompose the posterior variance into two components measuring the parameter and stochastic uncertainty, respectively. The first response surface model is the classical random-effects model which does not have any parametric assumptions, but may lead to estimation problems in some applications. The second response surface model is a Bayesian random-effects model which adds the assumption of normality for the observed responses, but overcomes the deficiencies of the previous model. We have finally presented two approaches to construct credible intervals around the mean response. The first is a percentile-type credible interval from the output of the Bayesian Simulation Replication Algorithm, and the second is based on the output of the Gibbs sampling algorithm for the second response surface model.

Frequentists and Bayesians approach the problem of estimation differently. Frequentists seek to estimate an unknown parameter based only on the model that has been adopted for the observable data, whereas Bayesians seek to estimate the parameter by appropriately combining their prior intuition with the information content in the data. Given the difference in views, we have developed some criteria for comparing the performance of both approaches in a simulation framework. The basic criteria are: the Bayes risk for point estimation, and the length of the confidence interval and its coverage probability for interval estimation. We have conducted two Monte Carlo

experiments on two queueing applications for our performance analysis. The first application is on a single server queue, where given the simplicity of the system we have seen clear empirical evidence in favor of the Bayesian approach mainly in terms of coverage probability. The second more realistic application supports this evidence by obtaining closer to nominal coverage probability at almost half the length of the confidence interval compared to the δ and bootstrap methods.

In Chapter 4, we have extended our Bayesian framework presented in Chapter 3 to account for model uncertainty as well as parameter and stochastic uncertainty. Note that the δ and bootstrap methods cannot be extended to account for model uncertainty. This Bayesian approach is known in the statistical literature as Bayesian Model Averaging (BMA). Although the BMA approach has been recently used in some applied statistical fields such as linear and nonlinear models, it has never been integrated and used in stochastic simulations as presented in our work. We have discussed in detail the basic ingredients of the BMA approach for conducting simulation experiments, and we have developed a Bayesian Simulation Replication Algorithm for estimating the mean response and for assessing the various sources of variability. We have also derived expressions for the mean and variance of the posterior distribution of the target simulation response; and we have presented a method to construct a credible interval on the mean response.

In practice, we are usually restricted by a limited computing time to conduct simulation experiments and make reasonable inferences about the output performance measures. We have proposed a replication allocation replication procedure to optimally assign the simulation effort to the different competing input models, based on the minimization of the variance of the overall estimator of the posterior mean response subject to a constraint on the total number of simulation runs. We have also proposed a proportional allocation scheme that can be easily implemented in practice compared to the optimal allocation scheme.

Finally, we have conducted a Monte Carlo simulation experiment on a computer communication network with 4 nodes and 4 communication channels to illustrate the application of the BMA approach in a situation for which the usual input modeling practice may be ambiguous. We have also evaluated the performance of the BMA approach in estimating the posterior mean response and assessing its variability. The results confirmed our earlier hypothesis about the possible dangers in using classical approaches for input model selection. The classical approach delivered high absolute

percentage errors and high mean squared errors in estimating the posterior mean response compared to the BMA approach. Moreover, the classical confidence bands systematically underestimated the intrinsic error in their corresponding point estimators, resulting in zero coverage probabilities for those confidence intervals. The BMA approach, on the other hand, showed a better performance in estimating the posterior mean response and in assessing the accuracy with which this quantity has been estimated. The credible interval was wider compared to the classical confidence intervals, but had a high coverage probability. The optimal and proportional allocation procedures improved further the BMA approach, resulting in posterior mean estimates with lower absolute percentage and mean squared errors and in confidence intervals with higher coverage probabilities.

5.2 Recommendations for Future Research

In this research, we have developed a Bayesian approach for simulation input modeling that takes into account all sources of uncertainty in a simulation experiment. We have illustrated the major benefits of such an approach in practice where input uncertainties driving the simulation model are large. With the recent advances in Bayesian computations, we believe that the introduction of Bayesian techniques into routine simulation practice is a much more attainable goal than previously thought. However, more work remains to be done in a number of areas to reach this goal. We provide a list of recommendations for future research in the area of Bayesian simulation input modeling.

5.2.1 Software Implementation

The implementation of a user-friendly software tool is an important step for the widespread use of the Bayesian techniques in simulation input modeling. This software should contain a wide range of distributions that are commonly used in simulation applications, including shifted versions of the exponential family of distributions and mixture distributions. The ideas behind the BUGS software can be adapted into this software to implement the MCMC techniques for sampling from posterior distributions. However, the software should also contain some graphical tools for the specification of prior distributions such as histogram construction, and statistical

tools such as quantile and moment matching estimation.

An important extension of the software would be the addition of more flexible types of distributions such as the four parameter Johnson distribution and the Beziér distribution.

5.2.2 Efficiency of the Simulation Replication Algorithm

If the simulation model can be run at a large number of input configurations, then a large random sample from the posterior distributions of the input parameters gives accurate inferences about the output quantity of interest. However, it is often not practical to evaluate a sufficient number of replications for a meaningful quantification of parameter uncertainty by this method. Thus we need to find a way of learning more about the output without having to run the simulation model a large number of times. A possible alternative approach involves the use of Latin Hypercube Sampling (LHS) described in detail by McKay et al. (1979). We briefly describe this sampling scheme, then give a simple illustrative example that shows its efficiency compared to Random Sampling (RS).

Suppose that the input model M_k consists of Q independent univariate random inputs. Let $F_{\theta_{kq}}(\cdot)$ be the posterior distribution function of parameter θ_{kq} for a single random input distribution q in model M_k where $q = \{1, \ldots, Q\}$, and let θ^r_{kq} be the r-th input parameter sampled for run $r = \{1, \ldots, R_k\}$. Define $A = (a_{rq})$ to be an $R_k \times Q$ matrix, where each column of A is an independent random permutation of $\{1, \ldots, R_k\}$. Let $U = (u_{rq})$ be a matrix of independent random numbers generated independently of A. Then θ^r_{kq} are sampled according to

$$\theta^r_{kq} \;=\; F^{-1}_{\theta_{kq}}\left(\frac{a_{rq} + u_{rq} - 1}{R_k}\right) \;\text{ for }\; q = 1, \ldots, Q. \tag{5.1}$$

Stein (1989) provides an extension for Latin Hypercube to approximate sampling of dependent variables.

Example. We consider an M/M/1 queue with arrival rate λ having a U(0,1) prior, and service rate μ having a U(1,2) prior. Let y denote the number of customers at steady state. Its expected value is given by

$$E[y] \;=\; \int_1^2 \int_0^1 E[y|\lambda,\mu]\, p(\lambda,\mu)\, d\lambda\, d\mu \;\;=\;\; \int_1^2 \int_0^1 \frac{\lambda}{\mu - \lambda}\, d\lambda\, d\mu \;=\; 0.8863.$$

Table 5.1: Percentage Errors: LHS vs. RS

R	5	10	20	50	100
RS	47.5	37.9	27.6	20.5	15.9
LHS	37.5	25.9	20.0	15.4	10.8

Table 5.2: Mean Square Errors: LHS vs. RS

R	5	10	20	50	100
RS	0.651	0.227	0.129	0.077	0.042
LHS	0.562	0.140	0.082	0.047	0.023

To illustrate the performance of LHS versus RS, we compute

$$\bar{y} \;=\; \frac{1}{R}\sum_{r=1}^{R}\frac{\lambda^{(r)}}{\mu^{(r)}-\lambda^{(r)}}$$

as an estimate for $E[y]$, where $\lambda^{(r)}$ and $\mu^{(r)}$ are sampled using LHS or RS. We compute this for several values of R and all the experiment is repeated 1000 times to generate the average absolute percentage errors (Table 5.1) and the mean square errors (Table 5.2) for both sampling schemes.

The results in Table 5.1 clearly show that the LHS scheme provides better estimates for the output of interest than the RS scheme for the same number of replications R. The percentage errors are much smaller using LHS with reasonable accuracy even for small values of R. For example, with a sample of size 20, the percentage error using LHS is around 20%. The same percentage error is achieved, however, with a sample size as large as 50 using RS. In addition to reductions in percentage errors, Table 5.1 shows that LHS achieves considerable reductions in mean square error when compared to RS.

The major problem with LHS is that the output responses computed from a LHS sample are correlated. This correlation does not affect our estimate of the mean response, but it usually affects our assessment of variability in the output response. In our work, the variance estimates were derived based on the assumption of independent replications. This is achieved under RS, but may not be the case for LHS. The challenging research problem is to investigate the effect of bias induced by using Latin hypercube sampling instead of random sampling.

5.2.3 Validity of the Response Surface Model

Among promising future improvements to the response surface models used in Chapters 3 and 4, the following items merit special attention.

- Relaxation of assumptions (3.26) and (4.14) so that τ_k^2 can be a function of θ_k.

- A more comprehensive experimental performance evaluation of the Bayesian approach on a broader range of test problems typically addressed by simulation techniques.

- Nonnormal variants of the Gibbs Sampler algorithm obtained by relaxing the normality assumption (3.35), which may not be consistent with the form of the posterior distribution $p(\theta|\boldsymbol{x})$. One possible variant of the Gibbs Sampler is to replace the normal distribution by a generalized Student-t distribution $t_\nu(\beta, \sigma^2)$ (see Section 4.5.2 for the form of the density). We can then assign a discrete prior density to the degrees of freedom parameter ν as

  $$p(\nu = k) \quad = \quad \frac{1}{K} \text{ for } k = 1, \ldots, K,$$

  where K is an integer less than 30. This variant of the Gibbs Sampler represents a more flexible family which contains the normal distribution as a special case for high values of K. Another more complicated variant is to replace the normal distribution with a mixture of two normals. This has the flexibility of detecting any bimodal mean response, but prohibits the use of improper priors (Robert, 1994). Mengersen and Robert (1996) suggest a new parameterization of the mixture model with noninformative priors which leads to a considerable improvement in the Bayesian approach to mixture models. We finally note that both of these variants of the Gibbs Sampler can be easily implemented in the BUGS statistical package (Spiegelhalter et al., 1996).

Bibliography

AbouRizk, S. M., Halpin, D. W. and Wilson, J. R. (1991). Visual Interactive Fitting of Beta Distributions. *Journal of Construction Engineering and Management*, 117 (4), 589-605.

Aitkin, M. (1991). Posterior Bayes Factors. *Journal of the Royal Statistical Society*, Ser. B, 53, 111-142.

Andradóttir, S., and Bier, V. M. (1997). Applying Bayesian Ideas in Simulation. Department of Industrial Engineering, University of Wisconsin-Madison, Technical Report 97-1.

Andrews, R. W., and Schriber, T. J. (1983). A Bayesian Batch Means Methodology for Analysis of Simulation Output. In *Proceedings of the Winter Simulation Conference*, eds. S. Roberts, J. Banks, and B. Schmeiser, 37-38. Institute of Electrical and Electronics Engineers, Inc.

Avramidis, A. N., and Wilson, J. R. (1994). A Flexible Method for Estimating Inverse Distribution Functions In Simulation Experiments. *ORSA Journal On Computing*, 6, 342-355.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, J. O., and Delampady, M. (1987). Testing Precise Hypothesis. *Statistical Science*, 3, 317-352.

Berger, J. O., and Perrichi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, 91, 109-122.

Bernardo, J. M., and Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Cario, M. C., and Nelson, B. L. (1997). *Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix*. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern

University, Evanston, Illinois.

Cario, M. C., and Nelson, B. L. (1996). *Autoregressive to Anything Time-series Input Processes for Simulation. Operations Research Letters*, 19, 51-58.

Carlin, B. P., and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall: London, UK.

Chaloner, K. (1996). The Elicitation of Prior Distributions. *Bayesian Biostatistics*, D. Berry and D. Stangl (eds), New York: Marcel Dekker.

Cheng, R. C. H., and Holland, W. (1998). Two-Point Methods for Assessing Variability in Simulation Output. *Journal of Statistical Computation and Simulation*, Vol 60, 183-205.

Cheng, R. C. H., and Holland, W. (1997). Sensitivity of Computer Simulation Experiments to Errors in Input Data. *Journal of Statistical Computation and Simulation*, Vol 57, 219-241.

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

Chick, S. E. (1999). Steps to Implement Bayesian Input Distribution Selection. In *Proceedings of the 1999 Winter Simulation Conference*, P.A. Farrington, H. B. Nembhard, D. T. Sturrock and G. W. Evans (eds), Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 317-324.

Chick, S. E. (1997). Bayesian Analysis for Simulation Input and Output. In *Proceedings of the 1997 Winter Simulation Conference*, S. Andradóttir, K. J. Healy, D. H. Withers and B. L. Nelson (eds), Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 253-260.

Cooke, R. M. (1994). Uncertainty in Dispersion and Deposition Accident Consequence Modeling Assessed with Performance-based Expert Judgement. *Reliability Engineering and System Safety*, 45, 35-46.

DeBrota, D. J., Dittus, R. S., Roberts, S. D., and Wilson, J. R. (1989). Visual Interactive Fitting of Bounded Johnson Distributions. *Simulation*, 52 (5), 199-205.

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

Dijkstra, T. K. (1988). *On Model Uncertainty and Its statistical Implications*. Springer, Berlin.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty (with discus-

sion). *Journal of the Royal Statistical Society*, Ser. B, 56, 501-514.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jacknife. *The Annals of Statistics*, 7, 1-26.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Gelfand, A. E. (1996). Model Determination Using Sampling-based Methods. In *Practical Markov Chain Monte Carlo*, eds. W. R. Gilks, D. J. Spiegelhalter, and S. Richardson, London: Chapman and Hall, 145-162.

Gelfand, A. E., and Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society*, Ser. B, 57, 45-98.

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

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

Genz, A., and Kass, R. E. (1993). Subregion Adaptive Integration of Functions Having a Dominant Peak. Technical Report, Carnegie Melon University, Department of Statistics.

George, E. I. (1999). Bayesian Model Selection. *Encyclopedia of Statistical Sciences Update*, Volume 3, Wiley, New York.

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrika*, 57, 1317-1339.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Glynn, P. (1986). Problems in Bayesian Analysis of Stochastic Simulation. In *Proceedings of the 1986 Winter Simulation Conference*, J. R. Wilson, J. O. Henriksen, and S. D. Roberts (eds), Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 376-383.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society*, Ser. B, 14, 107-114.

Gross, D., and Juttijudata, M. (1997). Sensitivity of Output Performance Measures to Input Distributions in Queueing Simulation Modeling. In *Proceedings of the 1997 Winter Simulation Conference*, S. Andradóttir, K. J. Healy, D. H. Withers and B. L. Nelson (eds), Institute of Electrical and Electronics Engineers, Piscataway,

New Jersey, 296-302.

Helton, J. C. (1998). Uncertainty and Sensitivity Analysis in the Presence of Stochastic and Subjective Uncertainty. *Journal of Statistical Computing and Simulation*, 1-74.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. Technical Report 9814, Department of Statistics, Colorado State University.

Ibrahim, J. G., and Laud, P. W. (1994). A Predictive Approach to the Analysis of Designed Experiments. *Journal of the American Statistical Association*, 89, 309-319.

Inoue, K., Chick, S. E., and Chen, C. H. (1999). An empirical Evaluation of Several Methods to Select the Best System. *ACM Transactions on Modeling and Computer Simulation* 9 (4): in press.

Jagerman, D. L., and Melamed, B. (1992). The Transition and Autocorrelation Structure of TES Processes, part I: General Theory. *Communications in Statistics: Stochastic Models*, 8 (2), 193-219.

Johnson, M. E. (1987). *Multivariate Statistical Simulation*. New York: John Wiley.

Johnson, N. L. (1949). Systems of Frequency Curves Generated By Method of Translation. *Biometrika*, 36, 297-304.

Kass, R. E., and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, Vol. 90, 773-795.

Kass, R. E., and Vaidyanathan, S. (1992). Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions. *Journal of the Statistical Royal Society*, Ser. B, 54, 129-144.

Kass, R. E., and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, 91, 1343-1370.

Kelton, W. D., Sadowski, R. P., and Sadowski, D. A. (1998). *Simulation with Arena*. McGraw-Hill, Inc.

Kleinrock, L. (1976). *Queueing Systems (Volume 2): Computer Applications*. New York: John Wiley.

Kuhl, M. E., J. R. Wilson, and Johnson, M. A. (1997). Estimating and Simulating Poisson Processes Having Trends or Multiple Periodicities. *IIE Transactions*, 29, 201-211.

Lauritzen, S. L., Thiesson, B., and Spiegelhalter, D. J. (1994). Diagnostic Systems

Created by Model Selection Methods—a Case Study. In Cheeseman, P. and Oldford, W. (eds), *Uncertainty in Artificial Intelligence 4*, Springer Verlag, 143-152.

Law, A. M., and Kelton, W. D. (2000). *Simulation Modeling & Analysis* (3rd ed.). New York: McGraw-Hill, Inc.

Law, A. M., Kelton, W. D., and Koenig, L. W. (1981). Relative Width Confidence Intervals for the Mean. *Communications in Statistics: Simulation and Computation*, B10, 29-39.

Law, A. M., and McComas, M. G. (2000). In *Proceedings of the 2000 Winter simulation Conference*, J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick (eds), Institute of Electrical and Electronics Engineers, Piscataway, NJ, 253-258.

Leamer, E. E. (1978). *Specification Searches*. Wiley, New York.

L'Ecuyer, P., and Perron, G. (1994). On the Convergence Rates of IPA and FDC Derivative Estimators. *Operations Research*, 42, 463-656.

Lee, S., Wilson, J. R., and Crawford, M. M. (1991). Modeling and Simulation of a Nonhomogeneous Poisson Process Having Cycle Behavior. *Communications in Statistics—Simulation and Computation*, 20, 777-809.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*. Rotterdam: University Press.

McKay, M. D., Conover, W. J., and Beckman, R. J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from Computer Code. *Technometrics*, 21, 239-245.

Madigan, D., and Raftery, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, Vol. 89, 1535-1546.

Madigan, D., and York, J. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review*, Vol. 63, 215-232.

Meng, X. L., and Wong, W. H. (1993). Simulating Ratios of Normalizing Constants Via a Simple Identity. Technical Report 365, University of Chicago, Department of Statistics.

Mengerson, K. L., and Robert, C. P. (1996). Testing for Mixture Via Entropy Distance and Gibbs Sampling. *Bayesian Statistics 5*, Clarendon Press-Oxford, 255–276.

Morris, C. N. (1983). Natural Exponential Families with Quadratic Variance Functions: Statistical Theory. *Annals of Statistics*, 11, 515-529.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall, London.

Nelson, B. L., Schmeiser, B. W., Taaffe, M. R., and Wang, J. (1997). Approximation-Assisted Point Estimation. *Operations Research Letters*, 20, 109-118.

Nelson, B. L., and Yamnitsky, M. (1998). Input Modeling Tools for Complex Systems. In *Proceedings of the 1998 Winter simulation Conference*, D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan (eds), Institute of Electrical and Electronics Engineers, Piscataway, NJ, 105-112.

Newton, M. A., and Raftery, A. E. (1994). Approximate Bayesian by the Weighted Likelihood Bootstrap (with Discussion). *Journal of the Royal Statistical Society*, Ser. B, 56, 3-48.

O'Hagen, A. (1991). Contribution to the Discussion of "Posterior Bayes Factors." *Journal of the Royal Statistical Society*, Ser. B, 56, 3-48.

O'Hagen, A. (1995). Fractional Bayes Factors for Model Comparison (with Discussion). *Journal of the Royal Statistical Society*, Ser. B, 56, 99-138.

Raftery, A. E. (1986). Choosing Models for Cross-Classifications. *American Sociological Review*, 51, 145-146.

Raftery, A. E. (1996). Hypothesis Testing and Model Selection with Posterior Simulation. In *Practical Markov Chain Monte Carlo*, eds. W. R. Gilks, D. J. Spiegelhalter, and S. Richardson, London: Chapman and Hall, 163-188.

Raftery, A. E., Madigan, D., and Volinsky, C. T. (1996). Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with Discussion). *Bayesian Statistics 5*, Clarendon Press-Oxford, 323–349.

Rao, P. (1997). *Variance Components Estimation*. London: Chapman and Hall.

Robert, C. P. (1994). *The Bayesian Choice*. New York: Springer-Verlag.

Roberts, H. V. (1965). Probabilistic Prediction. *Journal of the American Statistical Association*, 60, 50-62.

Roberts, S. D. (1983). *Simulation Modeling and Analysis with INSIGHT*. Regenstrief Institute for Health Care, Indianapolis, Indiana.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics*, 2, 110-114.

Schmeiser, B. W., and Deutsch, S. J. (1977). A Versatile Four Parameter Family of Probability Distributions Suitable for Simulation. *IIE Transactions*, 9, 176-181.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6, 461-464.

Scott, E. M. (1996). Uncertainty and Sensitivity Studies of Models of Environmental Systems. In *Proceedings of the 1996 Winter simulation Conference*, J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain (eds), Institute of Electrical and Electronics Engineers, Piscataway, NJ, 255-259.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). *BUGS 0.5: Bayesian Inference Using Gibbs Sampling Manual (Version ii).* MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK.

Stein, M. (1989). Large Sample Properties of Simulations Using Latin Hypercube Samples. *Technometrics*, 29 (2), 143-151.

Stuart, A., and Ord, J. K. (1994). *Kendall's Advanced Thoery of Statistics, Volume 1: Distribution Theory.* 6th ed. London: Edward Arnold.

Swain, J. J., Venkatraman, S., and Wilson, J. R. (1988). Least-squares Estimation of Distribution Functions in Johnson's Translation System. *Journal of Statistical Computation and Simulation*, 29, 271-297.

Taplin, R. H. (1993). Robust Likelihood Calculation for Time Series. *Journal of the Royal Statistical Society*, Ser. B, 55, 829-836.

Tierney, L., and Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, Vol. 81, 82-86.

Wagner, M. A. F., and Wilson, J. R. (1996). Using Univariate Bézier Distributions to Model Simulation Input Processes. *IIE Transactions*, 28, 699-711.

Wang, J., and Schmeiser, B. W. (1997). Monte Carlo Estimation of Bayesian Robustness. Forthcoming, *IIE Transactions.*

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian Model Averaging in Proportional Hazard Models. *Applied Statistics*, 46(3), 433-448.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.