BAYESIAN IDEAS FOR DISCRETE EVENT SIMULATION: WHY, WHAT AND HOW

Stephen E. Chick

1Technology and Operations Management
INSEAD
Fontainebleau, France

2006 Winter Simulation Conference

Why Bayesian methods?

- Glynn (1986): Uncertainty analysis. Not $\alpha = h(E[Y])$, but $\alpha(\theta) = h(E[Y | \theta])$
- Unknown parameters, $p(\theta)$, data from modeled system to update
  - Mean $E[\alpha(\Theta)]$
  - Distribution of $\alpha(\Theta)$ induced by $\Theta$
  - Credible set: $\theta_{lo}, \theta_{hi}$ so $p([h(\theta_{lo}), h(\theta_{hi})]) = 95$
- Chick (1997): Reviewed work to that date.
  - Suggested broader range of application.
    1 Ranking and selection
    2 Response surface modeling
    3 Experimental design

The Point of Today

- Review some basic concepts of subjective probability, Bayesian statistics, decision theory.
- Identify several applications to simulation experiments.
- Summarize some implementation issues.
- Identify some areas for future work.
Related work

See the WSC (2006) paper and chapter in Henderson and Nelson book for a long (but incomplete) citation list for work over the last 10 years on:

- Formal Bayes or decision theoretic theory
- Applications: scheduling, insurance, finance, traffic modeling, public health, waterway safety, supply chain and other areas
- Bayes and deterministic simulations
- Favorite books on subjective and Bayesian probability and decision theory

Public Policy and Health Economics: increasingly uses simulation (in addition to decision trees, Markov chains), and increasingly requires probabilistic sensitivity analysis.

Outline

1. Getting down to brass tacks
   - Subjective and Bayesian methods
   - Assessing prior probability
   - Asymptotic Theorems
   - Decisions, loss, and value of information
   - Entropy and Kullback-Leibler Discrepancy

2. Applications
   - Uncertainty Analysis
   - Selecting from Multiple Candidate Distributions
   - Selecting the Best System
   - Metamodels

3. Implementation

4. Summary

Getting down to brass tacks

Probability of 7 heads in the first 10 flips?
How to approach the problem...

Comte d’Alembert (18th cent.)
Indifference says maybe 1/11?
But wait, for one flip, probability of heads is 1/2?
Probability of 7 heads in the first 10 flips?

**Dwight (an unreconstructed frequentist)**
- $\frac{10!}{7!3!} \theta^7 (1 - \theta)^3$, where $\theta = \lim_{n \to \infty} \frac{X_1 + \ldots + X_n}{n}$ (a.e.).
- If we rent Madison Square garden and flip the tack repeatedly, I can estimate $\theta$ for you.
- What confidence and how accurately do you need to know $\theta$?

Hmmm...

If we rent Madison Square garden and flip the tack repeatedly, I can estimate $\theta$ for you.

What confidence and how accurately do you need to know $\theta$?

Let’s reformulate the question...

**Ralph**
- I’m willing to use probability for personal judgments
- $\int_0^1 \frac{10!}{7!3!} \theta^7 (1 - \theta)^3 \pi(\theta) d\theta$, where $\pi(\theta)$ is a prior probability.
- I’ll update with Bayes’ rule, to get posterior probability

$$p(\theta | x_n) = \frac{\pi(\theta)p(x_n | \theta)}{p(x_n)} = \frac{\pi(\theta) \prod_{i=1}^n p(x_i | \theta)}{\int p(x_n | \theta) d\pi(\theta)}$$
Why am I a Bayesian?

Lenny

Probability of 7 heads in the first 10 flips?

- Fair bets: I set \( p(E_1) > p(E_2) \) if I prefer the first bet:
  
  \[
  \begin{array}{c|c}
  E_1 & E_2 \\
  \$100 & \$100 \\
  \$0 & \$0 \\
  
  \end{array}
  \]

1) \( E_1 \) \$0
2) \( E_2 \) \$0

- Exchangeability (weaker than i.i.d.)
  
  \[ p(x_1, x_2, \ldots, x_n) = p(x_{s1}, x_{s2}, \ldots, x_{sn}) \]

for permutations \( s \) on \( \{1, 2, \ldots, n\} \) for arbitrary \( n \).

Implication for Simulation: \( Y_r = g(\theta_p, \theta_e, \theta_c; U_r) \)

- Input selection: Infinite exchangeable sequence \( X_{ij} \) from modeled system to infer \( i \)th statistical input, \( \theta_{pi} \)

Effect of \( \theta \): Metamodeling

- Input selection: Infinite exchangeable sequence \( X_{ij} \) from modeled system to infer \( i \)th statistical input, \( \theta_{pi} \)
- Metamodelling: Infinite exchangeable \( Y_r \) to infer \( \Psi \).
Establish exchangeability arguments, posit potential likelihood functions for observables, given unknown quantities
- Assess prior distributions for unknown quantities
- Relevant asymptotic theorems
- Decisions, loss, and value of information
  - ranking and selection,
  - other experimental design issues

We need a prior distribution for unknown parameters. For a Bernoulli outcome ... de Finetti (1990), Savage (1972) require You to assess your personal belief, $\pi(\theta)$ to describe $p(\Theta \leq \theta)$
- Important gain in flexibility
- Consistent with expected value decision theory
Some seek ‘automated’ methods . . .

For finite exchangeable sequence, set $\theta_N = \frac{X_1 + \ldots + X_N}{N} \in \{0/N, 1/N, \ldots, (N - 1)/N, 1\}$
- Indifference: discrete uniform for finite $N$
- Limit: $\lim_{N \to \infty} p(\theta_N) \xrightarrow{D} \text{uniform}[0,1]$
- Laplace (1812) used uniform[0,1] for his prior probability that the sun would come up tomorrow
- Coordinate dependence for continuous r.v. ($X_i$ versus log $X_i$)
Jeffrey's invariant prior

- $\pi(\theta) \propto |H(\theta)|^{1/2} d\theta$, where $H$ is the expected information in one observation,
  \[ H(\theta) = E_X \left( -\frac{\partial^2 \log p(X | \theta)}{\partial \theta^2} \right) \bigg|_{\theta} \]
  'uniform' with respect to the natural metric induced by the likelihood function (Kass 1989)
- Jeffrey's prior for Bernoulli sampling is beta(1/2, 1/2)
- For some likelihoods, Jeffrey's prior is improper (doesn't integrate to 1). Might be used formally

Conjugate prior distribution

Bernoulli sampling
- Set $s_n = \sum_{i=1}^{n} x_i$
- Likelihood:
  \[ p(x_n | \theta) \propto \theta^{s_n} (1 - \theta)^{n - s_n} \]
- Prior:
  \[ \pi(\theta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \text{ a beta}(\alpha, \beta) \text{ distribution} \]
- Posterior:
  \[ \propto \theta^{\alpha + s_n - 1} (1 - \theta)^{\beta + n - s_n - 1}, \text{ a beta}(\alpha + s_n, \beta + n - s_n) \text{ (conjugate) distribution} \]

Exponential family
- Likelihood:
  \[ p(x | \theta) = a(x) h_0(\theta) \exp \left[ \sum_{j=1}^{d} c_j \phi_j(\theta) t_j \right] \]
- Canonical conjugate prior:
  \[ p(\theta) = K(t)[h_0(\theta)]^{n_0} \exp \left[ \sum_{j=1}^{d} c_j \phi_j(\theta) t_j \right] \]
- Posterior, given $n$ data points: has parameters $n_0 + n$ and sum of $t = (t_1, t_2, \ldots, t_d)$ and sufficient statistics (Bernardo and Smith 1994)

Outline

1. Getting down to brass tacks
   - Subjective and Bayesian methods
   - Assessing prior probability
   - Asymptotic Theorems
     - Decisions, loss, and value of information
     - Entropy and Kullback-Leibler Discrepancy

2. Applications
   - Uncertainty Analysis
   - Selecting from Multiple Candidate Distributions
   - Selecting the Best System
   - Metamodels

3. Implementation

4. Summary
Classic Analogs for Infinite Exchangeable Sequences

Classical asymptotic theorems (e.g. Billingsley 1986)...
- laws of large numbers (LLN)
- central limit theorem (CLT)
- law of iterated logarithm (LIL)

...have Bayesian interpretations if considered conditional on mean and standard deviation of an infinite exchangeable sequence.

Bayesian LLN

A Bayesian extension of the LLN allows for sample averages to converge to random variables rather than to ‘true’ means.

**Theorem (Bayesian LLN)**

If \( \bar{X}_n \) and \( \bar{Y}_m \) are respectively the averages of \( n \) and \( m \) exchangeable random quantities \( X_i \) (the two averages may or may not have some terms in common), the probability that

\[
|\bar{X}_n - \bar{Y}_m| > \epsilon
\]

may be made arbitrarily small by taking \( n \) and \( m \) sufficiently large (de Finetti 1990, p. 216 assumes a finite variance).

**Theorem (Posterior Normality)**

For each \( n \), let \( p_n(\cdot) \) be the posterior pdf of the \( d \)-dimensional parameter \( \theta_n \) given \( x_n = (x_1, \ldots, x_n) \), let \( \bar{\theta}_n \) be its mode (MAP), and define the \( d \times d \) Bayesian observed information matrix \( \Sigma_n^{-1} \) by

\[
\Sigma_n^{-1} = -\left. \frac{\partial^2 \log p_n(\theta \mid x_n)}{\partial \theta^2} \right|_{\bar{\theta}_n}.
\]

Then \( \phi_n = \Sigma_n^{-1/2}(\theta_n - \bar{\theta}_n) \) converges in distribution to a standard (multivariate) normal random variable (Bernardo and Smith 1994, Prop 5.14 needs regularity conditions).

Frequentist analog: asserts that MLE is asymptotically normally distributed about a ‘true’ \( \theta_0 \) (Law and Kelton 2000), as opposed to describing uncertainty about \( \theta \).
Decisions under uncertainty

- Uncertainty described by probability ⇒ modeler can assess expected value of information (EVI) of additional data.
- EVI is useful in experimental design.
- EVI: value of resolving uncertainty with respect to a loss function $L(d, \omega)$ that describes losses when a decision $d$ is chosen when the state of nature is $\omega$.
- Data from experiment can reduce uncertainty about $\omega$, reduce expected loss.

Example: What is the mean? Setup

Example adapted from de Groot (1970) illustrates key concepts used for VIP procedures (Chick and Inoue 2001b)

- Decide if unknown mean $W$ of normal distribution (known $\sigma^2$) is smaller (decision $d = 1$) or larger ($d = 2$) than $w_0$.
- Exchangeable samples $X_n = (X_1, X_2, \ldots, X_n)$, with $p(X_i) \sim \text{Normal}(w, \sigma^2)$, given $W = w$, are available
- Goal: design experiment (choose $n$) to balance sampling cost ($cn$) and expected opportunity cost if wrong answer chosen

$$L(1, w) = \begin{cases} 0 & \text{if } w \leq w_0 \\ w - w_0 & \text{if } w > w_0, \end{cases}$$

$$L(2, w) = \begin{cases} w_0 - w & \text{if } w \leq w_0 \\ 0 & \text{if } w > w_0. \end{cases}$$

Example: What is the mean? If we knew...

- Prior: $W \sim \text{Normal}(\mu, 1/\tau)$ is conjugate. NOTE: $\tau$ is the precision in our uncertainty about unknown mean, $W$.
- Posterior: Observing $X_n = x_n$ would result in

$$p(w \mid x_n) \sim \text{Normal}(z, \tau_n^{-1})$$

$$z = \text{posterior mean of } W = E[W \mid x_n] = \frac{\tau \mu + \frac{n}{\sigma^2} x_n}{\tau + \frac{n}{\sigma^2}}$$

$$\tau_n = \text{posterior precision of } W = \tau + n/\sigma^2.$$  

$\tau_n^{-1}$ equals the asymptotic posterior variance approximation $\Sigma_n$ from theorem

Example: What is the mean? How much to know...

- Posterior mean $z$ for unknown $w$ influences the decision, but depends upon $n$, which is selected before $X_n$ is seen.
- Conditional distribution of $\bar{X}_n$, given $w$ is $\text{Normal}(w, \sigma^2/n)$
- Predictive distribution $p(z)$ of the posterior mean $Z = E[W \mid X_n] = (\tau \mu + \frac{n}{\sigma^2} \bar{X}_n)/\tau_n$
- Mixing over prior $\pi(w)$ implies a predictive distribution

$$Z \sim \text{Normal}(\mu, \tau_z^{-1})$$

$$\tau_z = \tau(\tau + n/\sigma^2)/(n/\sigma^2)$$

Note: $\tau_z^{-1} \to 0$ when $n \to 0$ (no new information). If $n \to \infty$, then $\text{Var}[Z] \to \sigma^2$. 
To minimize risk (sampling cost + expected loss from potentially incorrect decision), pick \( n \) to minimizes a nested expectation

\[
\rho(n) = cn + E_{X_n}[E_W[\mathcal{L}(d(X_n), W) | X_n]].
\]

- General technique: set \( \mathcal{L}^*(d, w) = \mathcal{L}(d, w) - \mathcal{L}(1, w) \), which is 0 if \( d = 1 \) and is \( w_0 - w \) if \( d = 2 \). Then

\[
E_W[\mathcal{L}^*(d(X_n), W) | X_n] = \begin{cases} 
0 & \text{if } d = 1 \\
 w_0 - Z & \text{if } d = 2.
\end{cases}
\]  

(3)

- To minimize loss in Eq. 3, assert \( d(X_n) = 2 \) (‘bigger’) if posterior mean exceeds threshold, \( Z > w_0 \), and assert \( d(X_n) = 1 \) (‘smaller’) if \( Z \leq w_0 \).

Alternate approximation

- Regular exponential family: asymptotic variance approximation \( \Sigma_n \) from theorem simplifies to

\[
H^{-1}(\theta)/(n_0 + n),
\]

where \( H \) is the expected information from one observation (Eq. 1), if canonical conjugate prior distribution is used

- Approximate effect of \( m \) new samples is to change posterior to

\[
\text{Normal}(\tilde{\theta}_n, \Sigma_n \frac{n_0 + n}{n_0 + n + m}).
\]

- Used for OCBA (Chen 1996), sampling plans for field data (Ng and Chick 2001). Frequentist result to obtain CI of desired size (Law and Kelton 2000)
Another loss function: discrepancy

- Kullback-Leibler discrepancy: difference between distributions
  \[ \delta(p \mid \hat{p}) = \sum \hat{p}_i \log(\hat{p}_i / q_i). \]
- Continuous r.v. \( X \) with densities \( \tilde{f} \) and \( f_\theta = f(x \mid \theta) \),
  \[ \delta(f_\theta \mid \tilde{f}) = \int \tilde{f}(x) \log \frac{\tilde{f}(x)}{f(x \mid \theta)} \, dx. \]
- One use: loss function for eliciting probability. If you believe the distribution is \( \tilde{f} \), and you lose \( \delta(f \mid \tilde{f}) \) if you provide a distribution \( f \), then you should honestly report \( \tilde{f} \) (Bernardo 1979)

Discrepancy: Other uses

- Select design matrix \( d_\Theta \) of \( r \) vectors of inputs \((\theta_{pi}, \theta_{ei}, \theta_{ci})\) for \( i = 1, 2, \ldots, r \) with output \( Y \) in order to best differentiate the posterior distribution of the response parameters \( \psi \) from the prior distribution for \( \psi \) (Bayesian D-optimal, Bernardo 1979; Smith and Verdinelli 1980; Ng and Chick 2004)
  \[ \int p(Y \mid d_\Theta) \left( \int p(\psi, Y) \log \frac{p(\psi \mid Y)}{p(\psi)} \, d\psi \right) \, dY \]
- Select maximum entropy prior distribution (Jaynes 1983)
- More later: input distribution selection . . .
Why Uncertainty Analysis?

- Simple Example
  \[ \mu_j \rightarrow \text{Simulation} \rightarrow Y_j = 2\mu_j + e_j \]

- Input model: \( X_\ell \sim \text{Normal} \left( \mu, \sigma_X^2 \right) \), known \( \sigma_X^2 \)
- Data: \( X_\ell \) observed (\( \ell = 1, 2, \ldots, n_0 \))
- Run \( r \) replications with \( X_{n_0} \) input for \( \mu \)
- Construct 90\% CI:
  \[ \bar{Y}_r \pm z_{0.05} \frac{X}{\sqrt{V_{tot}}} \]

Coverage?

If \( \mu \) is known, expect coverage to be 90\%. But \( \mu \) is not known.

\[ V_{tot} = \frac{\sigma_y^2}{r} + \frac{4\sigma^2}{n_0}, \] so try \( \bar{Y}_r \pm z_{0.05} \sqrt{V_{tot}} \)

---

Bayesian Ideas for Simulation

Getting down To brass tacks

Why Uncertainty Analysis?

- Simple Example
  \[ \mu_j \rightarrow \text{Simulation} \rightarrow Y_j = 2\mu_j + e_j \]

- Input model: \( X_\ell \sim \text{Normal} \left( \mu, \sigma_X^2 \right) \), known \( \sigma_X^2 \)
- Data: \( X_\ell \) observed (\( \ell = 1, 2, \ldots, n_0 \))
- Run \( r \) replications with \( X_{n_0} \) input for \( \mu \)
- Construct 90\% CI:
  \[ \bar{Y}_r \pm z_{0.05} \frac{X}{\sqrt{V_{tot}}} \]

Coverage?

If \( \mu \) is known, expect coverage to be 90\%. But \( \mu \) is not known.

Cl can be meaningless (several authors. . .)

\[ \bar{Y}_r \pm z_{0.05} \frac{X}{\sqrt{V_{tot}}} \]

\[ V_{tot} \approx \frac{\sigma_y^2}{r} + \sum_{i=1}^{k} \frac{\beta_i \left( H(\tilde{\beta})^{-1} \right)_{ij}}{n_i} \]
where gradient \( \beta_j = \frac{\partial g}{\partial \theta_j} \bigg|_\theta \)

(asymptotics. Cheng & Holland; Ng & Chick; Wilson & Zouaoui)
Why Uncertainty Analysis? Estimate response

Inputs
\[\theta_1 = (\mu_1, \lambda_1), \theta_2 = (\mu_2, \lambda_2), \theta_3 = (\mu_3, \lambda_3)\]

Simulation

\[Y_r = g(\theta) + \sigma Z = \beta_0 + \left(\sum_{i=1}^{k} \beta_{2i-1} \mu_i + \beta_{2i} \lambda_i\right) + \beta_1 \mu_1 \mu_2 + \beta_2 \mu_2^2 + \sigma Z\]

- Assess \(X_{it} \sim \text{Normal}(\mu_i, \lambda_i^{-1})\) for \(i\)th source of randomness
- Observe data \(X_{it}\) are observed \((i = 1, 2, 3; \ell = 1, 2, \ldots, n_i)\)
- Estimate unknown \(\beta\) with \(r_0\) runs (e.g. CCD from DOE)

\[V_{\text{tot}} = \frac{\sigma^2}{r_0} + \sum_{i=1}^{k} \frac{\partial g(\theta_n, \beta_n)}{\partial \theta_i} \sum_{n_i} \frac{\partial g(\theta_n, \beta_n)}{\partial \theta_i} + \sum_{\beta} \otimes \sum_{\theta}
\]

\[V_{\text{par}} = O(n_i^{-1}) \qquad V_{\text{resp}} = O(n_i^{-1}r^{-1})\]

Variations on Estimating \(E[Y | \mathcal{E}]\)

- Zouaoui and Wilson (2003): decouple stochastic, parameter uncertainty; update estimate as new data becomes available with variations on BMA
- Andradóttir and Glynn (2004): biased estimates of \(E[Y | \theta];\) quasi-random numbers; quadrature to select inputs \(\theta_i\)
- Estimate distribution of conditional expectation \(E[Y | \Theta, \mathcal{E}].\) Steckley and Henderson (2003) derive asymptotically optimal ways selecting \(r\) and \(n\) in BMA to produce a kernel density estimator (some conditions apply)
Basic Problem

- Which distribution/parameter to pick?
- ‘Usual’:
  1. Pick $q$ candidate input distributions (e.g. exponential, gamma, Weibull, lognormal)
  2. Find MLE $\hat{\theta}_i$ for candidates $i = 1, \ldots, q$
  3. Goodness-of-fit ($\chi^2$, K-S, A-D) tests
- Concerns:
  1. CI coverage if MLE/best distribution selected
  2. How to select among unrejected models? . . . (Lindley 1957; Berger and Pericchi 1996)

Bayesian Input Selection

- BMA applies without change for $q$ candidates
  1. Put prior on $(M = m, \theta_m)$, where $m \in \{1, 2, \ldots, q\}$
  2. Compute posterior $p(m, \Theta_m | E)$, sample from it in BMA
- Chick (2001): stochastic process simulation context; moment matching method for commensurate prior distributions
- Zouaoui and Wilson (2004): decouple stochastic uncertainty from two types of structural uncertainty (candidate model & parameters); variance reduction for BMA; numerical analysis
- Model selected closest to the true model in sense of Kullback-Leibler divergence (Berk 1966; Bernardo and Smith 1994; Dmochowski 1999).
Ranking and selection procedures differ?

- What is it for?
  - Select the ‘best’ of a finite set
  - Best identified by mean simulation response
- What are the approaches?
  - IZ, Indifference zone: $P(C) \geq 1 - \alpha$, $\delta^*$, repeated applications of procedure
  - VIP, Bayesian value of information procedures:
    - Like selection of mean bigger or smaller than threshold, Bayesian inference, loss, EVI
    - OCBA, Chen et al.:
      - Heuristic, allocates samples to improve Bayesian PCS using ‘thought experiment’ $\sigma^2/r_0 \rightarrow \sigma^2/(r_0 + r)$

VIP procedures

- Motivated by ‘statistical conservativeness’ of IZ approaches
- Two-stage: Unknown means of several systems.
  - Opportunity cost and 0-1 loss ($P(C)$)
  - Variances also unknown, different (conjugate prior, student marginal for mean)
  - Optimal solution unknown except special cases
- Sequential:
  - In theory, should improve things
  - Behrens-Fisher
  - Seems to work quite well

Procedure $LL(B)$, for opportunity cost (linear loss)

- Specify the first-stage sample size $r_0$. Take independent replications $y_{i1}, \ldots, y_{ir_0}$, for each system, $i = 1, \ldots, k$
- Compute all first-stage sample means $\bar{x}_i = \frac{\sum_{j=1}^{r_0} y_{ij}}{r_0}$ and sample variances $\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{r_0} (y_{ij} - \bar{x}_i)^2}{r_0 - 1}$, order statistics $\bar{x}[1] \leq \ldots \leq \bar{x}[k]$, and $\lambda_{i,k} = r_0 / (\hat{\sigma}_k^2 + \hat{\sigma}_i^2)$
- If sampling budget is $B$, run $r_i$ more independent replications,
  $$\eta[i] = \frac{B + \sum_{j \in S} r_0 c_j}{\sum_{j \in S} (c_j \hat{\sigma}_j^2 \eta_j)^{1/2} - r_0}$$
  $\eta[i] = (\lambda_{i,k})^{1/2} \phi_{i,k} (\bar{x}_i - \bar{x}[k])^2$ if $[i] \neq [k]$ and $\eta[k] = \sum_{j=1}^{k-1} \eta[j]$
- Select system with largest $\bar{x}_i = \sum_{j=1}^{r_0 + \eta_i} y_{ij} / (r_0 + r_i)$

Sample Comparison with Combined Procedure, $C$ (MDM)

<table>
<thead>
<tr>
<th>Figure of merit</th>
<th>Proc.</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANR</td>
<td>All</td>
<td>738</td>
<td>3,429</td>
<td>8,784</td>
<td>42,862</td>
</tr>
<tr>
<td>Empirical P(C)</td>
<td>$C - 0-1(B)$</td>
<td>0.8363</td>
<td>0.9140</td>
<td>0.9323</td>
<td>0.9763</td>
</tr>
<tr>
<td>Empirical frac. ‘good’ selections</td>
<td>$LL(B)$</td>
<td>0.8527*</td>
<td>0.9117</td>
<td>0.9480*</td>
<td>0.9937*</td>
</tr>
<tr>
<td>Expected posterior PCS</td>
<td>$C - 0-1(B)$</td>
<td>1.0000</td>
<td>0.9943</td>
<td>0.9990</td>
<td>1.0000</td>
</tr>
<tr>
<td>Expected bound, opp. cost</td>
<td>$LL(B)$</td>
<td>1.0000</td>
<td>0.9953</td>
<td>0.9997</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Monotone decreasing means (MDM). * indicates statistically significant difference
Common random numbers (CRN) can sharpen contrasts between systems (e.g. same simulated demand pattern)

Two-stage with screening (Chick and Inoue 2001a)
- Run subset of systems in stage 2
- Use 'missing data' formulas to update
- Select from even screened systems

Matrix intensive, heuristic provided

VIP procedures have solid basis, perform numerically quite well Chick and Inoue (2001, 2001a, 2002).
- Matlab: All procedures (0-1 or opportunity cost loss; two-stage or independent sequential; two-stage CRN)
- C: Independent replications, two-stage or sequential, both loss functions, plus variants (work in progress) to achieve Bayesian predictive targets for PCS and EOC

Can show asymptotic relation between certain VIP and OCBA procedures.

Specific Bayesian procedures with new stopping rules highly effective (Branke, Chick, and Schmidt 2005)
Uncertainty Analysis

Input Distribution Selection

Ranking/Selection

Metamodels

Normal linear model

\[ Y = \sum_{\ell=1}^{p} g_{\ell}(\theta)\beta_\ell + Z(\theta; U) = g^T(\theta)\beta + Z(\theta; U), \quad (4) \]

Gaussian random function (GRF) - kriging

\[ Y(\theta) = \sum_{\ell=1}^{p} g_{\ell}(\theta)\beta_\ell + Z(\theta) = g^T(\theta)\beta + Z(\theta), \quad (6) \]

for known regression functions \( g_1, \ldots, g_p \) of \( \mathbb{R}^d \), unknown regression coefficients \( \beta \), and a zero-mean random second-order process so that for any distinct inputs \( \theta_1, \ldots, \theta_m \), the vector \( (Y_1, \ldots, Y_m) \) has (nonindependent) multivariate normal distribution, conditional on \( \beta \).
Inference with GRFs

- GRF: Determined by mean $g^T(\theta)\beta$ and covariance function
  
  $$C^*(\theta_1, \theta_2) = \text{Cov}(Y(\theta_1), Y(\theta_2))$$
  
- Common to assume strong stationarity, so
  
  $$C^*(\theta_1, \theta_2) = C(\theta_1 - \theta_2)$$
  
- Inference for $g(\theta)$ at $\theta_{r+1}$ not yet input to simulation model with correlation function $R(h) = C(h)/C(0)$ for $h \in \mathbb{R}^d$. Example: power exponential $R(h) = \prod \exp[-|h_i/\gamma_i|^\rho_i]$ for $\rho_i \in [0, 2]$. 
- Kriging (geostatistics) is best linear unbiased prediction (BLUP) for $g(\theta_{r+1})$


More GRF

- Assessment of uncertainty in $g(\theta_{r+1})$ ⇒ experimental design technique to choose inputs to reduce response uncertainty (Santner et al. 2003)
- Also see tutorial by van Beers and Kleijnen (2004)
- More work is needed for GRFs in the stochastic simulation context.

Implementation

**Issues:**
- Maximize (MLE $\hat{\theta}$ or MAP $\tilde{\theta}$)
- Integrate (marginal distribution $p(\theta_1 | x_n)$ from $p(\theta_1, \theta_2 | x_n)$), or proportionality constant $c^{-1} = \int f(x_n | \theta)d\pi(\theta)$
- Simulate (sample from $p(\theta | x_n)$ to estimate $E[\alpha(\Theta)]$)

**Tools:**
- Newton-Raphson, Nelder-Mead, expectation-maximization (EM) algorithm, . . .
- Quadrature, normal approximation, data augmentation (IP algorithm), importance sampling (IS)
- Inversion, importance sampling (IS), Markov Chain Monte Carlo (MCMC)

Target: Sample from $p(\theta | \mathcal{E})$
Capable: Easily sample from $q(\cdot | \theta_{t-1})$

Initialize $t = 0, \theta_0$
for $t = 1, 2, . . .$
  sample a candidate $\theta \sim q(\cdot | \theta_{t-1})$
  compute acceptance probability $\alpha(\theta_{t-1}, \theta) = \min \left\{ 1, \frac{p(\theta | \mathcal{E})q(\theta_{t-1} | \theta)}{p(\theta_{t-1} | \mathcal{E})q(\theta | \theta_{t-1})} \right\}$
  generate an independent $u \sim \text{uniform}[0, 1]$
  if $u \leq \alpha(\theta_{t-1}, \theta)$ then set $\theta_t \leftarrow \theta$
  otherwise set $\theta_t \leftarrow \theta_{t-1}$
  set $t \leftarrow t + 1$
end loop
Approximating Posterior Distributions: Pros and Cons

**Exact Posterior**
- Exact
- Good if simple closed form known
- May be hard in general (mixtures, missing data, marginal distribution, curse of dimensionality?)

Here, gene linkage example (Tanner 1996)

**Asymptotic normality (from theorem)**
- $\text{Normal}(\hat{\theta}_n, \Sigma_n)$
- Relatively easy to compute
- Requires many data points ($n > 20d$)
- Does not model skew, etc.

**Data Augmentation (IP algorithm)**
- $p(\theta|\mathcal{E}) = \int p(\theta|z, \mathcal{E})p(z|\mathcal{E})\,dz$,
  average over $Z = \text{‘missing data’}$

1. Set $i = 0$; $g_0(\theta) =$ current estimate of $p(\theta|\mathcal{E})$
2. Generate $z_1, z_2, \ldots, z_m$ from $g_i(\theta)$ by
   1. Sample $\theta_1, \ldots, \theta_m$ from $g_i(\theta)$
   2. Sample $z_1, \ldots, z_m$ from $p(z|\theta_i, \mathcal{E})$
3. $g_{i+1}(\theta) = \sum_{j=1}^{m} p(\theta|z_j, \mathcal{E})/m$

**MCMC with histogram**
- General tool to sample (approximately) from posterior
- Complicated models possible
- Output analysis issues (convergence)
- Time-average (IP was average of distributions)
- MCMC can average distributions, too
Some Tools

- Handcode: Matlab, C, Gauss
- WinBUGS (Spiegelhalter et al. 1996) (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml)
- R (http://www.r-project.org/), S-PLUS packages, BOA add-on (http://www.public-health.uiowa.edu/boa/)
- Uncertainty analysis in spreadsheet Monte Carlo applications are available (e.g. Winston 2000).
- Most DEDS Commercial Tools: cumbersome to implement BMA

Bayesian Methods for Stochastic Process Simulation

- Themes:
  - Represent *all* uncertainty with probability, update with Bayes’ rule, expected value of information for sampling decisions
  - Use simulation to efficiently estimate quantities of interest for a Bayesian analysis
  - Bayesian, decision theory fits well with managerial/economic mindset
  - Applications: input distribution selection, uncertainty analysis, experimental design, ranking and selection
  - Asymptotic approximations helpful when exact optimal solutions are hard to find

Research opportunities include

- More links to economics of simulation analysis
- Input modeling and uncertainty analysis (kernel estimation of conditional means; the effect of different candidate distributions on uncertainty; prior distributions elicitation; calibration/inverse problem)
- Response modeling (extend the Gaussian random field work for stochastic simulation; nonasymptotic sampling plans for input parameter inference to optimally reduce output uncertainty; reasoning about models; calibration...)
- Experimental design (quantiles, non-expected value goals; CRN for unknown input parameters for ranking and selection; non-Gaussian output for ranking and selection, GRFs)
- Improved computational tools (e.g. MCMC, software interop)


Expected information as expected utility.


*Bayesian theory.*
Chichester, UK: Wiley.

Billingsley, P. 1986.
*Probability and Measure.* 2nd ed.
New York: John Wiley & Sons, Inc.

Selecting a selection procedure.

A lower bound for the correct subset-selection probability and its application to discrete event simulations.


Chick, S. E. 1997.
Bayesian analysis for simulation input and output.
Piscataway, NJ: IEEE, Inc.

Input distribution selection for simulation experiments: Accounting for input uncertainty.

The economics of simulation selection procedures.
Technical report, INSEAD, Technology and Operations Management Area working paper.

New procedures for identifying the best simulated system using common random numbers.

New two-stage and sequential procedures for selecting the best simulated system.

*Statistics for spatial data.*
New York: J. Wiley.

de Finetti, B. 1990.
*Theory of Probability, v. 2.*
New York: John Wiley & Sons, Inc.

*Optimal statistical decisions.*