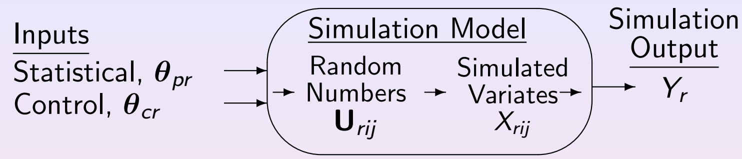
BAYESIAN IDEAS FOR DISCRETE EVENT SIMULATION: WHY, WHAT AND HOW

Stephen E. Chick[1]

[1]Technology and Operations Management
INSEAD
Fontainebleau, France

2006 Winter Simulation Conference

Why Bayesian methods in Simulation?



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

- Example: Single Server Queue (M/M/1): $\boldsymbol{\theta}_p = (\lambda, \mu_i) =$ arrival and service rates (server $i = 1, 2$)
- Output: $Y \approx \lambda/(\mu_i - \lambda)$ + noise
- Simulation: Analyze stochastic processes via sample path generation. Inform decisions: pick control parameter $\boldsymbol{\theta}_c$, to estimate or to optimize value $h(E[Y \mid \boldsymbol{\theta}_p, \boldsymbol{\theta}_c])$
- Bayesian as alternative to frequentist

Why Bayesian methods?

- Glynn (1986): *Uncertainty analysis*. Not $\alpha = h(E[Y])$, but

$$\alpha(\theta) = h(E[Y \mid \theta])$$

- Unknown parameters, $p(\theta)$, data from modeled system to update
  1. Mean $E[\alpha(\Theta)]$
  2. Distribution of $\alpha(\Theta)$ induced by Θ
  3. Credible set: θ_{lo}, θ_{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*.

INSEAD

Outline

INSEAD

Getting down to brass tacks

Probability of 7 heads in the first 10 flips?

How to approach the problem...

INSEAD

Getting down to brass tacks

Probability of 7 heads in the first 10 flips?

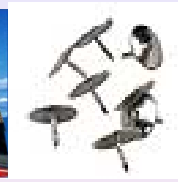
Comte d'Alembert (18th cent.)

- Indifference says maybe $1/11$?
- But wait, for one flip, probability of heads is $1/2$?

See Savage (1972) and Kreps (1988).

INSEAD

Getting down to brass tacks



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 θ for you.
- What confidence and how accurately do you need to know θ?

Hmmm. . .

Getting down to brass tacks

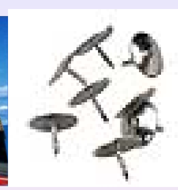
Probability of 7 heads in the first 10 flips?

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 θ for you.
- What confidence and how accurately do you need to know θ?

Let's reformulate the question . . .

Why am I a Bayesian?



Will you accept the following bet *now*? You get \$100 if there are 7 heads, but you pay \$5 if not.

more from Dwight

- I can't answer until I have a good idea of what θ is.
- Guessing wouldn't be scientific.

Why am I a Bayesian?


Ralph

Probability of 7 heads in the first 10 flips?

- 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*

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

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:

- Exchangeability (weaker than i.i.d.)

$$p(x_1, x_2, \ldots, x_n) = p(x_{s_1}, x_{s_2}, \ldots, x_{s_n})$$

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

INSEAD

Why am I a Bayesian?

Lenny

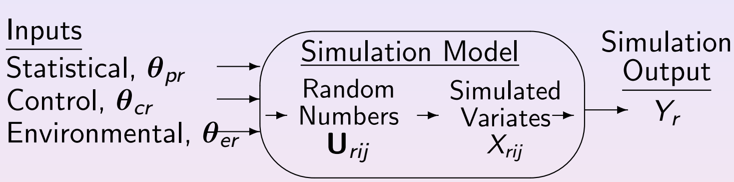
- Exchangeability plus conceptually infinite N *imply*

$$\lim_{N\to\infty} p(7 \text{ heads in first 10 flips}) = \int_0^1 \frac{10!}{7!3!}\theta^7(1-\theta)^3 dF(\theta)$$

- de Finetti (1990)-like representation
- Ralph assumed conditional i.i.d., while Lenny *derives* formula from exchangeability
- Probability defined by bet preferences, not repeated outcomes

INSEAD

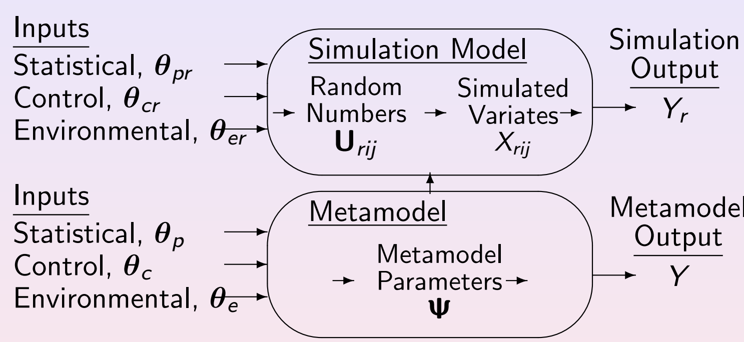
Implication for Simulation: $Y_r = g(\boldsymbol{\theta}_p, \boldsymbol{\theta}_e, \boldsymbol{\theta}_c; \mathbf{U}_r)$



- Input selection: Infinite exchangeable sequence X_{ij} from modeled system to infer ith statistical input, $\boldsymbol{\theta}_{pi}$

INSEAD

Implication for Simulation: $Y_r = g(\boldsymbol{\theta}_p, \boldsymbol{\theta}_e, \boldsymbol{\theta}_c; \mathbf{U}_r)$



- Input selection: Infinite exchangeable sequence X_{ij} from modeled system to infer ith statistical input, $\boldsymbol{\theta}_{pi}$
- Metamodeling: Infinite exchangeable Y_r to infer $\boldsymbol{\Psi}$.

INSEAD

Part of the process

- 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

INSEAD

Outline

INSEAD

Subjective Methods

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
- Kahneman, Slovic, and Tversky (1982) describe difficulties with elicitation . . .

Some seek 'automated' methods . . .

INSEAD

Principle of Indifference

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: $\lim_{N\to\infty} p(\theta_N) \stackrel{\mathcal{D}}{\to} \mathtt{uniform}[0,1]$
- Laplace (1812) used $\mathtt{uniform}[0,1]$ for his prior probability that the sun would come up tomorrow

- Coordinate dependence for continuous r.v. (X_i versus $\log X_i$)

INSEAD

Jeffrey's invariant prior

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

$$H(\theta) = E_X \left[-\frac{\partial^2 \log p(X \mid \theta)}{\partial \theta^2} \bigg|_\theta \right], \qquad (1)$$

  'uniform' with respect to the natural metric induced by the likelihood function (Kass 1989)
- *Jeffreys' prior* for Bernoulli sampling is $\texttt{beta}(1/2, 1/2)$
- For some likelihoods, Jeffreys' prior is improper (doesn't integrate to 1). Might be used formally

INSEAD

Conjugate prior distribution

Bernoulli sampling

- Set $s_n = \sum_{i=1}^n x_i$
- Likelihood: $p(\mathbf{x}_n|\theta) \propto \theta^{s_n}(1-\theta)^{n-s_n}$
- Prior: $\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, a $\texttt{beta}(\alpha, \beta)$ distribution
- Posterior: $\propto \theta^{\alpha+s_n-1}(1-\theta)^{\beta+n-s_n-1}$, a $\texttt{beta}(\alpha+s_n, \beta+n-s_n)$ (conjugate) distribution

Exponential family

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

INSEAD

'Noninformative'

- The $\texttt{uniform}[0, 1]$ distribution is conjugate for Bernoulli sampling—a $\texttt{beta}(1, 1)$ distribution.
- 'Noninformative' means 'evenly spread'—a heuristic term
- For canonical conjugate prior (for exponential family), the posterior has parameter $n_0 + n$
- Some think of $n_0 + n$ as an 'effective' number of data points
- 'Noninformative' associated with a small n_0

Others

- Jaynes (1983): maximum entropy methods
- Berger (1994), Kass and Wasserman (1996): Default rules

Useful? Actually informative?

INSEAD

Outline

INSEAD

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 θ_n given $\mathbf{x}_n = (x_1, \ldots, x_n)$, let $\tilde{\theta}_n$ be its mode (MAP), and define the $d \times d$ Bayesian observed information matrix Σ_n^{-1} by

$$\Sigma_n^{-1} = -\frac{\partial^2 \log p_n(\theta \mid \mathbf{x}_n)}{\partial \theta^2}\bigg|_{\tilde{\theta}_n}. \qquad (2)$$

Then $\phi_n = \Sigma_n^{-1/2}(\theta_n - \tilde{\theta}_n)$ 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' θ_0 (Law and Kelton 2000), *as opposed to* describing uncertainty about θ.

Outline

Decisions under uncertainty

- Uncertainty described by probability $\Rightarrow$ 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* $\mathcal{L}(d, \omega)$ that describes losses when a *decision d* is chosen when the state of nature is ω.
- Data from experiment can reduce uncertainty about ω, reduce expected loss.

INSEAD

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 σ^2) is smaller (decision $d = 1$) or larger ($d = 2$) than w_0.
- Exchangeable samples $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$, with $p(X_i) \sim \mathtt{Normal}\left(w, \sigma^2\right)$, given $W = w$, are available
- Goal: design experiment (choose n) to balance sampling cost (cn) and expected opportunity cost if wrong answer chosen

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

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

INSEAD

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

- Prior: $W \sim \mathtt{Normal}\left(\mu, 1/\tau\right)$ is conjugate. NOTE: τ is the *precision* in our uncertainty about unknown mean, W.
- Posterior: Observing $\mathbf{X}_n = \mathbf{x}_n$ would result in

$$\begin{aligned} p(w \mid \mathbf{x}_n) &\sim \mathtt{Normal}\left(z, \tau_n^{-1}\right) \\ z &= \text{posterior mean of } W = E[W \mid \mathbf{x}_n] = \frac{\tau\mu + \frac{n}{\sigma^2}\bar{x}_n}{\tau + \frac{n}{\sigma^2}} \\ \tau_n &= \text{posterior precision of } W = \tau + n/\sigma^2. \end{aligned}$$

τ_n^{-1} equals the asymptotic posterior variance approximation Σ_n from theorem

INSEAD

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 $\mathbf{X}_n$ is seen.
- Conditional distribution of $\bar{X}_n$ given w is $\mathtt{Normal}\left(w, \sigma^2/n\right)$
- *Predictive distribution* $p(z)$ of the posterior mean $Z = E[W \mid \mathbf{X}_n] = (\tau\mu + \frac{n}{\sigma^2}\bar{X}_n)/\tau_n$
- Mixing over prior $\pi(w)$ implies a predictive distribution

$$\begin{aligned} Z &\sim \mathtt{Normal}\left(\mu, \tau_z^{-1}\right) \\ \tau_z &= \tau(\tau + n/\sigma^2)/(n/\sigma^2) \end{aligned}$$

Note: $\tau_z^{-1} \to 0$ when $n \to 0$ (no new information).
If $n \to \infty$, then $\mathrm{Var}[Z] \to \sigma^2$.

INSEAD

Example: What is the mean? What is the risk. . .

To minimize risk (sampling cost + expected loss from potentially incorrect decision), pick n to minimizes a nested expectation

$$\rho(n) = cn + E_{\mathbf{X}_n}[E_W[\mathcal{L}(d(\mathbf{X}_n), W) \mid \mathbf{X}_n]].$$

- 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

$$E_W[\mathcal{L}^*(d(\mathbf{X}_n), W) \mid \mathbf{X}_n] = \begin{cases} 0 & \text{if } d = 1 \\ w_0 - Z & \text{if } d = 2. \end{cases} \quad (3)$$

- To minimize loss in Eq. 3, assert $d(\mathbf{X}_n) = 2$ ('bigger') if posterior mean exceeds threshold, $Z > w_0$, and assert $d(\mathbf{X}_n) = 1$ ('smaller') if $Z \leq w_0$.

INSEAD

Example: What is the mean? Expected loss

- Decision depends on $\mathbf{X}_n$ via Z; Z has normal distribution
- Expected loss found with standard normal loss for newsboy

$$\begin{aligned} L_N[s] &= \int_s (t - s)\phi(t)dt = \phi(s) - s(1 - \Phi(s)) \\ E[\mathcal{L}^*(d(\mathbf{X}_n), W)] &= E_{\mathbf{X}_n}[E_W[\mathcal{L}^*(d(\mathbf{X}_n), W) \mid \mathbf{X}_n]] \\ &= -\tau_z^{\frac{-1}{2}} L_N[\tau_z^{\frac{1}{2}}(w_0 - \mu)] \end{aligned}$$

- Add back $E[\mathcal{L}(1, W)]$, use prior for W

$$E[\mathcal{L}(d(\mathbf{X}_n), W)] = \tau^{\frac{-1}{2}} L_N[\tau^{\frac{1}{2}}(w_0 - \mu)] - \tau_z^{\frac{-1}{2}} L_N[\tau_z^{\frac{1}{2}}(w_0 - \mu)]$$

- EVI and EVPI

INSEAD

Example: What is the mean? What is the risk. . .

- First-order optimality condition

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

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

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

- Asymptotic approximations are a second useful tool to identify criteria-based sampling plans.
- Extensions of these ideas used to derive VIP procedures (Chick and Inoue 2001b).

INSEAD

Alternate approximation

- Regular exponential family: asymptotic variance approximation Σ_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

$$\mathtt{Normal}\left(\tilde{\theta}_n, \Sigma_n \frac{n_0 + n}{n_0 + n + m}\right).$$

- 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)

INSEAD

Outline

INSEAD

Another loss function: discrepancy

- Kullback-Leibler discrepancy: difference between distributions
- Discrete distributions $\tilde{p}$ and p,

$$\delta(p \,||\, \tilde{p}) = \sum \tilde{p}_i \log(\tilde{p}_i / q_i).$$

- Continuous r.v. X with densities $\tilde{f}$ and $f_\theta = f(x \mid \theta)$,

$$\delta(f_\theta \,||\, \tilde{f}) = \int \tilde{f}(x) \log \frac{\tilde{f}(x)}{f(x \mid \theta)} dx.$$

- One use: loss function for eliciting probability. If you believe the distribution is $\tilde{f}$, and you lose $\delta(f \,||\, \tilde{f})$ if you provide a distribution f, then you should honestly report $\tilde{f}$ (Bernardo 1979)

INSEAD

Discrepancy: Other uses

- Select design matrix $\mathbf{d}_\Theta$ of r vectors of inputs $(\boldsymbol{\theta}_{pi}, \boldsymbol{\theta}_{ei}, \boldsymbol{\theta}_{ci})$ for $i = 1, 2, \ldots, r$ with output $\mathbf{Y}$ in order to best differentiate the posterior distribution of the response parameters ψ from the prior distribution for ψ (Bayesian D-optimal, Bernardo 1979; Smith and Verdinelli 1980; Ng and Chick 2004)

$$\int p(\mathbf{Y} \mid \mathbf{d}_\Theta) \left(\int p(\psi \mid \mathbf{Y}) \log \frac{p(\psi \mid \mathbf{Y})}{p(\psi)} d\psi \right) d\mathbf{Y}$$

- Select maximum entropy prior distribution (Jaynes 1983)
- More later: input distribution selection ...

INSEAD

Outline

INSEAD

Why Uncertainty Analysis?

- Simple Example

$$\mu_j \rightarrow \boxed{\text{Simulation}} \rightarrow Y_j = 2\mu_j + e_j$$

- Input model: $X_\ell \sim \texttt{Normal}\left(\mu, \sigma_x^2\right)$, known σ_x^2
- Data: X_ℓ observed $(\ell = 1, 2, \ldots, n_0)$
- Run r replications with $\bar{X}_{n_0}$ input for μ
- Construct 90% CI:

$$\bar{Y}_r \pm z_{0.95} \frac{\hat{\sigma}_y}{\sqrt{r}}$$

Coverage?

If μ is known, expect coverage to be 90%. But μ is *not* known.

INSEAD

Why Uncertainty Analysis?

- Simple Example

$$\mu_j \rightarrow \boxed{\text{Simulation}} \rightarrow Y_j = 2\mu_j + e_j$$

- Input model: $X_\ell \sim \texttt{Normal}\left(\mu, \sigma_x^2\right)$, known σ_x^2
- Data: X_ℓ observed $(\ell = 1, 2, \ldots, n_0)$
- Run r replications with $\bar{X}_{n_0}$ input for μ
- Construct 90% CI:

$$\bar{Y}_r \pm z_{0.95} \frac{\hat{\sigma}_y}{\sqrt{r}}$$

Coverage?

If μ is known, expect coverage to be 90%. But μ is *not* known.
CI can be meaningless (several authors...)

Why Uncertainty Analysis?

- Simple Example

$$\mu_j \rightarrow \boxed{\text{Simulation}} \rightarrow Y_j = 2\mu_j + e_j$$

- Input model: $X_\ell \sim \texttt{Normal}\left(\mu, \sigma_x^2\right)$, known σ_x^2
- Data: X_ℓ observed $(\ell = 1, 2, \ldots, n_0)$
- Run r replications with $\bar{X}_{n_0}$ input for μ
- Construct 90% CI:

$$\bar{Y}_r \pm z_{0.95} \cancel{\frac{\hat{\sigma}_y}{\sqrt{r}}} \sqrt{V_{tot}}$$

Account for parameter uncertainty

$V_{tot} = \frac{\sigma_y^2}{r} + \frac{4\sigma_x^2}{n_0}$, so try $\bar{Y}_r \pm z_{0.95}\sqrt{V_{tot}}$

Why Uncertainty Analysis?

- Simple Example

$$\mu_j \rightarrow \boxed{\text{Simulation}} \rightarrow Y_j = 2\mu_j + e_j$$

- Input model: $X_\ell \sim \texttt{Normal}\left(\mu, \sigma_x^2\right)$, known σ_x^2
- Data: X_ℓ observed $(\ell = 1, 2, \ldots, n_0)$
- Run r replications with $\bar{X}_{n_0}$ input for μ
- Construct 90% CI:

$$\bar{Y}_r \pm z_{0.95} \cancel{\frac{\hat{\sigma}_y}{\sqrt{r}}} \sqrt{V_{tot}}$$

For *known* mean response g, input parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$

$V_{tot} \approx \frac{\sigma_y^2}{r} + \sum_{i=1}^k \frac{\boldsymbol{\beta}_i H(\tilde{\theta}_i)^{-1} \boldsymbol{\beta}_i^{\mathrm{T}}}{n_i}$ where gradient $\boldsymbol{\beta}_i = \frac{\partial g(\tilde{\boldsymbol{\theta}})^{\mathrm{T}}}{\partial \theta_i}\Big|_{\tilde{\boldsymbol{\theta}}}$
(asymptotics. Cheng & Holland; Ng & Chick; Wilson & Zouaoui)

Why Uncertainty Analysis? Estimate response

Inputs

$\theta_1 = (\mu_1, \lambda_1)$ →
$\theta_2 = (\mu_2, \lambda_2)$ → Simulation →
$\theta_3 = (\mu_3, \lambda_3)$ →

$$\begin{aligned} Y_r &= g(\boldsymbol{\theta}) + \sigma Z \\ &= \beta_0 + \left(\sum_{i=1}^{k} \beta_{2i-1}\mu_i + \beta_{2i}\lambda_i\right) \\ &\quad + \beta_7\mu_1\mu_2 + \beta_8\mu_1^2 + \sigma Z \end{aligned}$$

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

$$V_{tot} = \frac{\hat{\sigma}_y^2}{r_0} + \underbrace{\sum_{i=1}^{k} \frac{\partial g(\tilde{\boldsymbol{\theta}}_n, \tilde{\boldsymbol{\beta}}_{r_0})}{\partial \theta_i} \Sigma_{n_i} \frac{\partial g(\tilde{\boldsymbol{\theta}}_n, \tilde{\boldsymbol{\beta}}_{r_0})^{\mathrm{T}}}{\partial \theta_i}}_{V_{par}=O(n_i^{-1})} + \underbrace{\Sigma_{\boldsymbol{\beta}} \otimes \Sigma_{\boldsymbol{\theta}}}_{V_{resp}=O(n_i^{-1}r^{-1})}$$

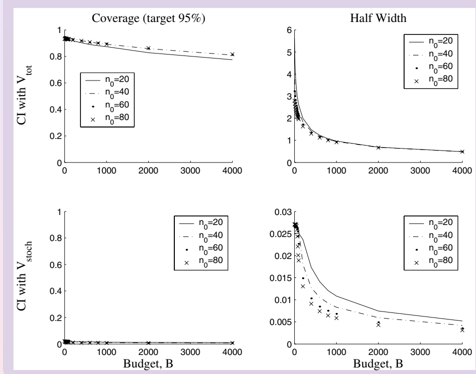
Uncertainty Analysis: Uncertainty Reduction

- Goal: Reduce uncertainty, not just quantify
  - r more replications
  - m_i more samples from source of randomness i
- Optimization:

$$\min_{r, m_i} \quad \frac{\hat{\sigma}_y^2}{r_0 + r} + \sum_{i=1}^{k} \frac{\xi_i}{n_i + m_i} + \frac{\zeta_i}{(r_0 + r)(n_i + m_i)}$$

$$r, m_i \quad \geq \quad 0$$

CI Coverage (Ng and Chick 2006)



Uncertainty Analysis

- Sensitivity analysis: $E[g(\theta)]$ as a function of θ (average out stochastic uncertainty from $\mathbf{u}$)
- Uncertainty analysis: $E[Y \mid \mathcal{E}]$, with both stochastic and parameter uncertainty, given all information $\mathcal{E}$
- Bayesian Model Average (BMA) estimates $E_Y[Y \mid \mathcal{E}]$

for $r = 1, \ldots, R$ replications
  sample parameter $\theta_r \sim p(\theta \mid \mathcal{E})$
  for $i = 1, 2, \ldots, n$
    generate output y_{ri} given input θ_r
  end loop
end loop
Generate estimate $\bar{y} = \sum_{r=1}^{R} \frac{1}{R} \sum_{i=1}^{n} \frac{y_{ri}}{n}$.

Variations on Estimating $E[Y \mid \mathcal{E}]$

- 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 $E[Y \mid \theta]$; quasi-random numbers; quadrature to select inputs θ_i
- Estimate distribution of conditional expectation $E[Y \mid \Theta, \mathcal{E}]$. Steckley and Henderson (2003) derive asymptotically optimal ways selecting r and n in BMA to produce a kernel density estimator (some conditions apply)

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

INSEAD

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 (χ-squared, 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)

INSEAD

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 $p(m, \Theta_m|\mathcal{E})$, 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).

INSEAD

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

INSEAD

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:
    $P(CS) \geq 1 - \alpha$, δ^*, 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' $\sigma^2/r_0 \rightarrow \sigma^2/(r_0 + r)$
  - Economic. Chick and Gans (2005) propose a new economic approach, includes costs of replications and discounting, maximizes E[NPV] of decisions with simulation.

VIP procedures

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

Procedure $\mathcal{LL}(\mathcal{B})$, for opportunity cost (linear loss)

1. Specify the first-stage sample size r_0. Take independent replications $y_{i1}, \ldots, y_{ir_0}$, for each system, $i = 1, \ldots, k$
2. Compute all first-stage sample means $\bar{x}_i = \sum_{j=1}^{r_0} y_{ij}/r_0$ 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 $\lambda_{i,k} = r_0/(\hat{\sigma}_{[k]}^2 + \hat{\sigma}_{[i]}^2)$
3. If sampling budget is B, run r_i more independent replications,

$$r_{[i]} = \frac{B + \sum_{j \in \mathcal{S}} r_0 c_j}{\sum_{j \in \mathcal{S}} \left(\frac{c_j c_{[i]} \hat{\sigma}_j^2 \eta_j}{\hat{\sigma}_{[i]}^2 \eta_{[i]}} \right)^{1/2}} - r_0$$

$\eta_{[i]} = (\lambda_{i,k})^{1/2} \frac{(r_0-1)+\lambda_{i,k}(\bar{x}_{[k]}-\bar{x}_{[i]})^2}{(r_0-1)-1} \phi_{r_0-1}[(\lambda_{i,k})^{1/2}(\bar{x}_{[k]} - \bar{x}_{[i]})]$ if $[i] \neq [k]$ and $\eta_{[k]} = \sum_{j=1}^{k-1} \eta_{[j]}$

4. Select system with largest $\bar{\bar{x}}_i = \sum_{j=1}^{r_0+r_i} y_{ij}/(r_0 + r_i)$

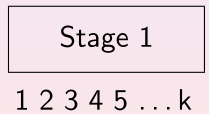
Sample Comparison with Combined Procedure, $\mathcal{C}$ (MDM)

Figure of merit	Proc.	Number of systems, k: 2	5	10	100
ANR	All	738	3,429	8,784	42,862
Empirical P(CS)	$\mathcal{C}$	0.8363	0.9140	0.9323	0.9763
	0-1($\mathcal{B}$)	0.8527*	0.9117	0.9480*	0.9937*
	$\mathcal{LL}(\mathcal{B})$	0.8500	0.9293*	0.9660*	0.9987*
Empirical frac. 'good' selections	$\mathcal{C}$	1.0000	0.9943	0.9990	1.0000
	0-1($\mathcal{B}$)	1.0000	0.9953	0.9977	0.9997
	$\mathcal{LL}(\mathcal{B})$	1.0000	0.9953	0.9993	1.0000
Expected posterior PCS	$\mathcal{C}$	0.8336	0.8379	0.8649	0.9318
	0-1($\mathcal{B}$)	0.8446*	0.8339	0.8821*	0.9717*
	$\mathcal{LL}(\mathcal{B})$	0.8470*	0.8462*	0.9022*	0.9842*
Expected bound, opp. cost	$\mathcal{C}$	0.0176	0.0138	0.0104	0.0037
	0-1($\mathcal{B}$)	0.0157*	0.0128*	0.0075*	0.0012*
	$\mathcal{LL}(\mathcal{B})$	0.0154*	0.0110*	0.0055*	0.0005*

Monotone decreasing means (MDM). * indicates statistically significant difference

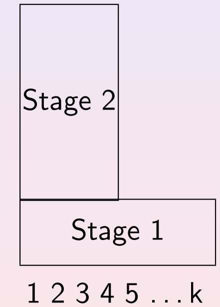
VIP: Common Random Numbers

- 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



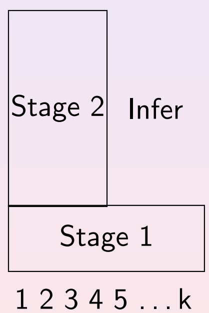
VIP: Common Random Numbers

- 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: Common Random Numbers

- 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



Summaries

- 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)

Outline

INSEAD

Metamodels

- Normal linear model

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

- Gaussian random function (GRF) - kriging

INSEAD

Normal linear model

$$Y = \sum_{\ell=1}^{p} g_\ell(\boldsymbol{\theta})\beta_\ell + Z(\boldsymbol{\theta};\mathbf{U}) = \mathbf{g}^T(\boldsymbol{\theta})\boldsymbol{\beta} + Z(\boldsymbol{\theta};\mathbf{U}), \quad (5)$$

- Model
  - Known: regression functions $g_1, \ldots, g_p$
  - Unknown: coefficients $\boldsymbol{\beta}$, variance of zero-mean noise $Z(\cdot)$
- Conjugate prior $p(\boldsymbol{\beta}, \sigma^2)$ (if all factors active)
  - Inverted gamma distribution for unknown variance σ^2
  - Multivariate normal distribution for $\boldsymbol{\beta}$ given σ^2,
  - Raftery, Madigan, and Hoeting (1997) describe a relatively 'uninformative' prior distribution, 'good' results
- Identifying important factors like input distribution selection
  - 2^p candidate models
  - KL discrepancy-based design criterion balances factor identification and parameter estimation (Ng and Chick 2004)

INSEAD

Gaussian random functions (GRF)

- Well-known in deterministic simulations, particularly in geostatistics (Cressie 1993; Santner et al. 2003)
- Provide flexibility that the linear model does not, and are useful when g takes a long time to compute.
- GRF for unknown nonstochastic g (no random numbers $\mathbf{u}$) is

$$Y(\boldsymbol{\theta}) = \sum_{\ell=1}^{p} g_\ell(\boldsymbol{\theta})\beta_\ell + Z(\boldsymbol{\theta}) = \mathbf{g}^T(\boldsymbol{\theta})\boldsymbol{\beta} + Z(\boldsymbol{\theta}) \quad (6)$$

for *known* regression functions $g_1, \ldots, g_p$ of $\mathbb{R}^d$, unknown regression coefficients $\boldsymbol{\beta}$, and a *zero-mean random second-order process* so that for any distinct inputs $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$, the vector $(Y_1, \ldots, Y_m)$ has (nonindependent) multivariate normal distribution, conditional on $\boldsymbol{\beta}$.

INSEAD

Inference with GRFs

- GRF: Determined by mean $\mathbf{g}^T(\boldsymbol{\theta})\beta$ and covariance function

$$C^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathrm{Cov}(Y(\boldsymbol{\theta}_1), Y(\boldsymbol{\theta}_2))$$

- Common to assume strong stationarity, so

$$C^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = C(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$$

- Inference for $g(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_{r+1}$ not yet input to simulation model with correlation function $R(\mathbf{h}) = C(\mathbf{h})/C(0)$ for $\mathbf{h} \in \mathbb{R}^d$. Example: power exponential $R(\mathbf{h}) = \prod \exp[-|h_i/\gamma_i|^{p_i}]$ for $p_i \in [0, 2]$.
- Kriging (geostatistics) is best linear unbiased prediction (BLUP) for $g(\boldsymbol{\theta}_{r+1})$

More GRF

- Assessment of uncertainty in $g(\boldsymbol{\theta}_{r+1}) \Rightarrow$ experimental design technique to choose inputs to reduce response uncertainty (Santner et al. 2003)
- Also see tutorial by van Beers and Kleijnen (2004)
- Deterministic simulation: (Sacks et al. 1989; O'Hagan et al. 1999; Kennedy and O'Hagan 2001; Santner et al. 2003; van Beers and Kleijnen 2003).
- More work is needed for GRFs in the stochastic simulation context.

Implementation

Issues:

- Maximize (MLE $\hat{\theta}$ or MAP $\tilde{\theta}$)
- Integrate (marginal distribution $p(\theta_1 \mid \mathbf{x}_n)$ from $p(\theta_1, \theta_2 \mid \mathbf{x}_n)$), or proportionality constant $c^{-1} = \int f(\mathbf{x}_n \mid \theta) d\pi(\theta)$)
- Simulate (sample from $p(\theta \mid \mathbf{x}_n)$ 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)

Evans and Swartz 1995; Tanner 1996; Gilks et al. 1996; Devroye 2006, ...

Metropolis-Hastings: An MCMC Algorithm

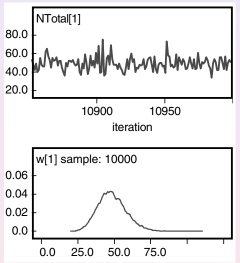
Target: Sample from $p(\theta \mid \mathcal{E})$
Capable: Easily sample from $q(\cdot \mid \theta_{t-1})$

Initialize $t = 0$, θ_0
for $t = 1, 2, \ldots$
  sample a candidate $\theta \sim q(\cdot \mid \theta_{t-1})$
  compute acceptance probability

$$\alpha(\theta_{t-1}, \theta) = \min\left\{1, \frac{p(\theta|\mathcal{E}) \cdot q(\theta_{t-1}|\theta)}{p(\theta_{t-1}|\mathcal{E}) \cdot q(\theta|\theta_{t-1})}\right\}$$

  generate an independent $u \sim \mathtt{uniform}[0, 1]$
  if $u \le \alpha(\theta_{t-1}, \theta)$ then set $\theta_t \leftarrow \theta$
    otherwise set $\theta_t \leftarrow \theta_{t-1}$
  set $t \leftarrow t + 1$
end loop

Sample path



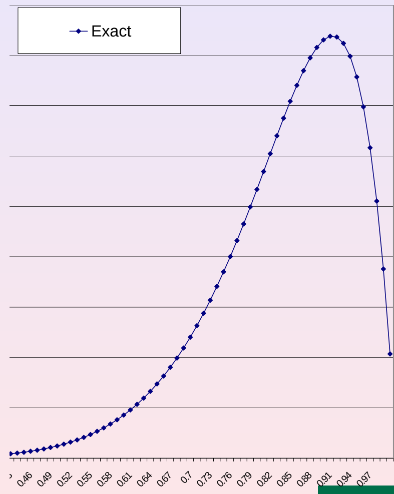
Density Estimate

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)

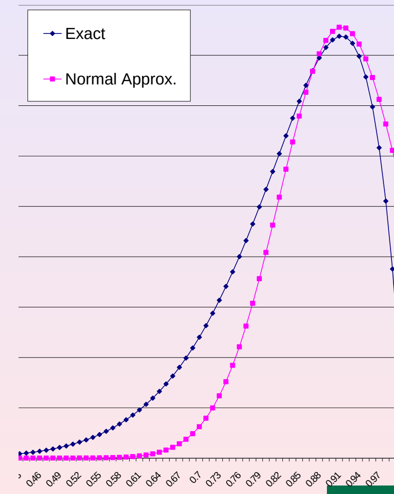


Getting down To brass tacks Applications **Implementation** S

Approximating Posterior Distributions: Pros and Cons

Asymptotic normality (from theorem)

- $\texttt{Normal}\left(\tilde{\theta}_n, \Sigma_n\right)$
- Relatively easy to compute
- Requires many data points ($n > 20d$)
- Does not model skew, etc.



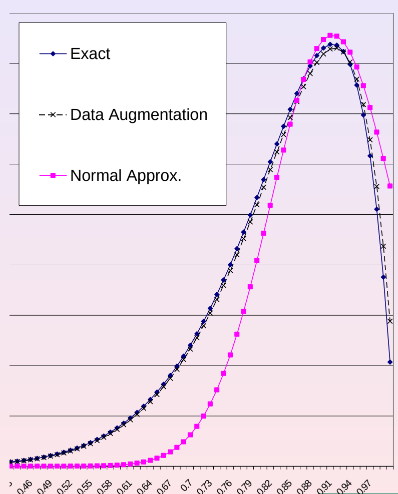
WSC'06 Bayesian Ideas for Simulation

Getting down To brass tacks Applications **Implementation** S

Approximating Posterior Distributions: Pros and Cons

Data Augmentation (IP algorithm)

- $p(\theta|\mathcal{E}) = \int p(\theta|Z,\mathcal{E})p(Z|\mathcal{E})$, average over Z = '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. $g_{i+1}(\theta) = \sum_{j=1}^{m} p(\theta|z_j, \mathcal{E})/m$



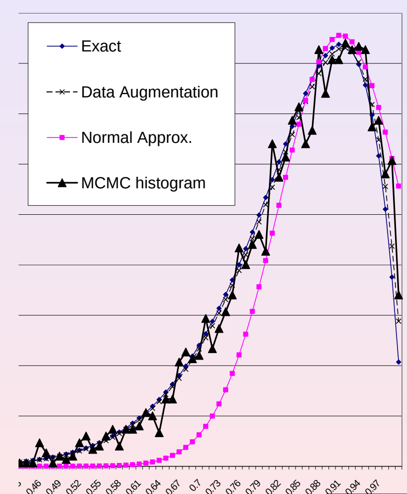
WSC'06 Bayesian Ideas for Simulation

Getting down To brass tacks Applications **Implementation** S

Approximating Posterior Distributions: Pros and Cons

MCMC with histogram

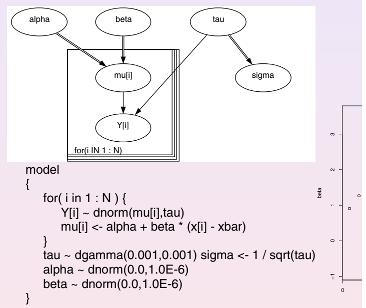
- General tool to sample (approximately) from posterior
- Complicated models possible
- Output analysis issues (convergence)
- Time-average (IP was average of distributions)
- MCMC can average distributions, too



WSC'06 Bayesian Ideas for Simulation

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



INSEAD

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

INSEAD

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)

INSEAD

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

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

Berger, J. O., and L. R. Pericchi. 1996.
The intrinsic Bayes factor for model selection and prediction.
Journal of the American Statistical Association 91:109–122.

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

Bernardo, J. 1979.

INSEAD

Expected information as expected utility.
Annals of Statistics 7:686–690.

Bernardo, J. M., and A. F. M. Smith. 1994.
Bayesian theory.
Chichester, UK: Wiley.

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

Branke, J., S. E. Chick, and C. Schmidt. 2005.
Selecting a selection procedure.
Technology and Operations Management Area, INSEAD, Working Paper.

Chen, C.-H. 1996.
A lower bound for the correct subset-selection probability and its application to discrete event simulations.

INSEAD

IEEE Transactions on Automatic Control 41 (8): 1227–1231.

Chick, S. E. 1997.
Bayesian analysis for simulation input and output.
In *Proc. 1997 Winter Simulation Conference*, ed. S. Andradóttir, K. Healy, D. Withers, and B. Nelson, 253–260.
Piscataway, NJ: IEEE, Inc.

Chick, S. E. 2001.
Input distribution selection for simulation experiments: Accounting for input uncertainty.
Operations Research 49 (5): 744–758.

Chick, S. E., and N. Gans. 2005.
The economics of simulation selection procedures.
Technical report, INSEAD, Technology and Operations Management Area working paper.

Chick, S. E., and K. Inoue. 2001a.

INSEAD

New procedures for identifying the best simulated system using common random numbers.
Management Science 47 (8): 1133–1149.

Chick, S. E., and K. Inoue. 2001b.
New two-stage and sequential procedures for selecting the best simulated system.
Operations Research 49 (5): 732–743.

Cressie, N. A. 1993.
Statistics for spatial data.
New York: J. Wiley.

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

de Groot, M. H. 1970.
Optimal statistical decisions.

INSEAD

New York: McGraw-Hill.

Devroye, L. 2006.
Nonuniform random variate generation.
In *Handbook in Operations Research and Management Science: Simulation*, ed. S. G. Henderson and B. L. Nelson. Elsevier.

Dmochowski, J. 1999.
Intrinsic priors via Kullback-Leibler geometry.
In *Bayesian Statistics 5*, 543–549.
Oxford University Press.

Evans, M., and T. Swartz. 1995.
Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems.
Statistical Science 10 (3): 254–272.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996.
Markov chain monte carlo in practice.

INSEAD