ESUP Accept/Reject Sampling

Brian Caffo
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

January 20, 2003

Home Page

Title Page

Page 2 of 32

Go Back

Full Screen

Close

Quit

Monte Carlo in general

- Use of simulated random variables to approximate unknown distributional quantities that are difficult to approximate analytically
- For example if we can simulate independent draws, X_j, from target distribution F we can use the SLLN

$$E_F[h(X)] \approx \sum_{i=1}^{m} h(X_i)/m$$

- Statistical (frequentist) approach to solving numerical problems

$\longrightarrow$ Probabilistic estimates of the error in approximation

$\longrightarrow$ Central limit theorem

Home Page

Title Page

Page 3 of 32

Go Back

Full Screen

Close

Quit

Accept/reject sampling

- Simulates from the *target* distribution by thinning out samples from a *candidate* distribution
- Target Distribution F with density f
- Candidate distribution G with density g
- Support of G contains the support of F
- Let $C = \sup_x f(x)/g(x) < \infty$

Home Page
Title Page
Page 4 of 32
Go Back
Full Screen
Close
Quit

Accept/reject algorithm

1	Generate $X \sim G$
2	Generate $U \sim$Uniform$(0,1)$
3	Accept X if $U \leq \frac{f(X)}{Cg(X)}$
4	Goto 1

- Accepted values will all have distribution F
- Number of candidates to obtain one target is geometric with success probability $1/C$ (Acceptance Rate)

Number of candidates to obtain one target is geometric with success probability $1/C$

$$
\begin{aligned}
P(X \text{ is accepted}) &= P\left(U \le \frac{f(X)}{Cg(X)}\right) \\
&= E_G\left[P\left(U \le \frac{f(X)}{Cg(X)} \middle| X\right)\right] \\
&= E_G\left[\frac{f(X)}{Cg(X)}\right] \\
&= 1/C
\end{aligned}
$$

Accepted values will all have distribution F

$$P(X \leq x | X \text{ is accepted})$$

$$\begin{aligned}
&= P(X \leq x, X \text{ is accepted}) C \\
&= P\left(X \leq x, U \leq \frac{f(X)}{Cg(X)}\right) C \\
&= E_G\left[P\left(X \leq x, U \leq \frac{f(X)}{Cg(X)} \middle| X\right)\right] C \\
&= E_G\left[1_{\{X \leq x\}} P\left(U \leq \frac{f(X)}{Cg(X)} \middle| X\right)\right] C \\
&= E_G\left[1_{\{X \leq x\}} \frac{f(X)}{g(X)}\right] \\
&= E_F\left[1_{\{X \leq x\}}\right] \\
&= F(x)
\end{aligned}$$

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 6 of 32
Go Back
Full Screen
Close
Quit

Home Page
Title Page
Page 7 of 32
Go Back
Full Screen
Close
Quit

Notes

- C represents the average number of candidates to obtain 1 target
- If $f^* \propto f$ and $g^* \propto g$ and $C^* = \sup_x f^*(x)/g^*(x)$ then

$$\frac{f(x)}{Cg(x)} = \frac{f^*(x)}{C^*g^*(x)}$$

  That is f and g need only be known up to constants of proportionality
- Inaccurate candidates often lead to easy calculation of C
- Accurate candidates often have forms inconvenient (impossible) for calculating C
- General case, assume $F << G$ and let $C = \sup_x dF(x)/dG(x)$ Replace f/g in the algorithm with dF/dG

Home Page

Title Page

◀◀ ▶▶

◀ ▶

Page 8 of 32

Go Back

Full Screen

Close

Quit

ESUP Accept/Reject Algorithm

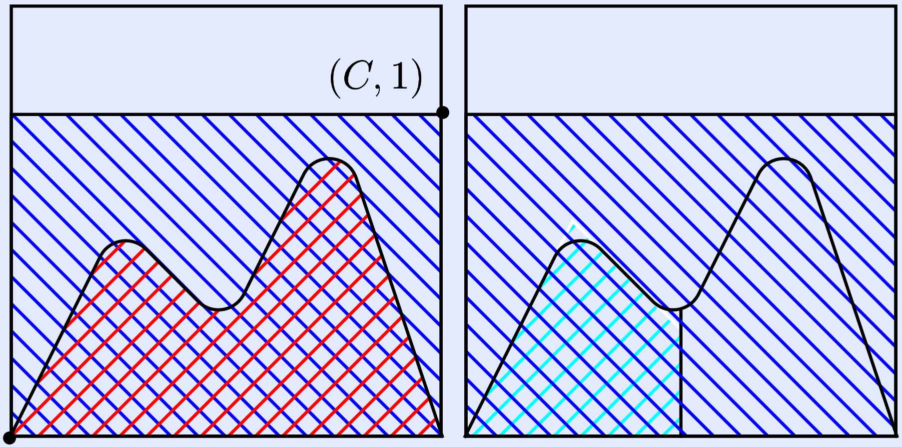
Instead of calculating C we estimate it as the algorithm progresses

$$C = \sup_x f(x)/g(x)$$

$$T = \text{maximum observed value of } f(X)/g(X)$$

Suppose we have a starting value for T

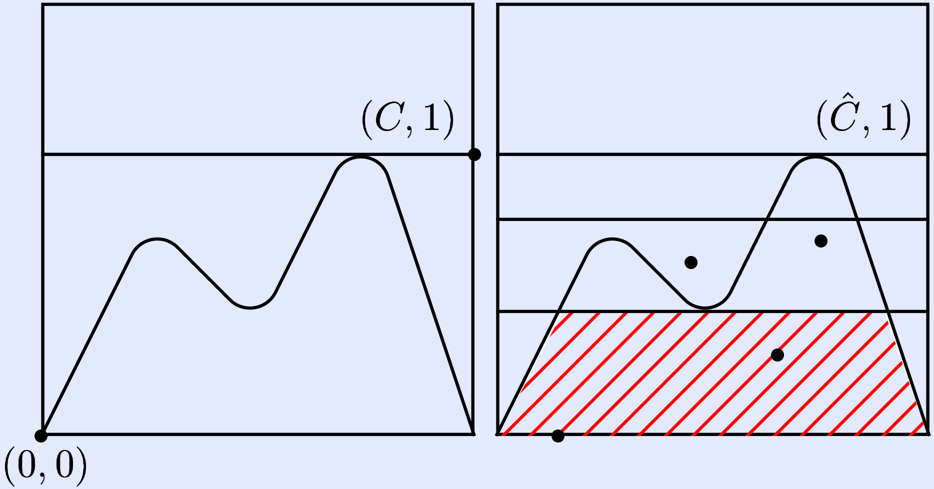
1. Generate $X \sim G$
2. Generate $U \sim$Uniform$(0, 1)$
3. Accept X if $U \leq \frac{f(X)}{Tg(X)}$
4. Update $T = \max\left(T, \frac{f(X)}{g(X)}\right)$
5. Goto 1



$P(\text{Point accepted})$ =  /  $= 1/C$

$P(\text{Pt acpt and } X < x)$ =  /  $= F(x)/C$

$P(X < x | \text{Pt acpt})$ =  /  $= F(x)$


$(C, 1)$
$(\hat{C}, 1)$
$(0, 0)$

Notes

- As $\frac{f(X)}{Cg(X)} \leq \frac{f(X)}{Tg(X)}$ the ESUP algorithm always accepts a candidate that the known sup (KSUP) algorithm would accept
- Worst cases $\frac{f(X)}{Cg(X)} \leq U \leq \frac{f(X)}{Tg(X)}$

  $\rightarrow$ ESUP algorithm accepts a candidate it should not have
- T is an extreme order statistic based on i.i.d. observations

  $\rightarrow$ T converges almost surely to C
- Sequence of accepted values are no longer identically distributed or independent, however, our results show they may be treated as if they were

Recap

- Accurate candidates often have forms inconvenient for calculating C
- ESUP rejection sampling estimates C with the largest observed value of $f(X)/g(X)$ from the sequence of $X's \sim G$
  We label this estimate T
- Claim: T converges so rapidly to C that the sequence of accepted values may be treated as if they were i.i.d.
- Prove this claim by comparing the ESUP algorithm and the KSUP algorithm run with the same candidates and uniforms
  ESUP algorithm will tend to accept the same candidates as the KSUP algorithm

Home Page
Title Page
Page 13 of 32
Go Back
Full Screen
Close
Quit

More Notation

- Y_i accepted value from classical C known rejection sampling
- $\tilde{Y}_i$ accepted value from ESUP rejection sampler using the same candidates and uniforms used to obtain Y_i
- Obtain all of our results by comparing Y_i and $\tilde{Y}_i$ We show that eventually $Y_i = \tilde{Y}_i$
- Key quantity $P(Y_i \neq \tilde{Y}_i)$
- Added assumption $C = f(x_C)/g(x_C)$ for some x_C in the support of F

Home Page

Title Page

Page 14 of 32

Go Back

Full Screen

Close

Quit

Theorem

If the support of F is discrete then

$$\sum_{i=1}^{\infty} P(Y_i \neq \tilde{Y}_i) < \infty$$

The sequences are the same for all but finitely many i

Argument

- For each candidate generated there is a positive probability $T = C$
- Notice then the number of iterations until $T = C$ is finite with probability one

$$P(Y_i \neq \tilde{Y}_i) \leq P(\text{A geometric random variable } \geq i)$$

- It is then easy to show the right hand side of the inequality sums.

Home Page
Title Page
Page 15 of 32
Go Back
Full Screen
Close
Quit

Notes

- Practically, previous theorem covers continuous supports as well, as we generally are only concerned with T reaching a certain decimal accuracy
- Formally, however, we prove $P(Y_i \neq \tilde{Y}_i) = \mathcal{O}(i^{-1})$ (does not sum)
- This rate is fast enough to prove sequence from the ESUP algorithm obeys the same CLT and SLLN as the sequence from the KSUP algorithm
- Added assumption $C = f(x_C)/g(x_C)$ for some x_C in the support of F

Main Theorem

- $\{Y_i\}$ the i.i.d. sequence of F variates from the KSUP algorithm
- $\{\tilde{Y}_i\}$ the sequence from the ESUP algorithm
- If Y is an F variate, let the mean of $h(Y) = \mu_h$ and the variance of $h(Y)$ be σ_h^2

$$(i.)\frac{\sum_{i=1}^{n} h(Y_i)}{n} \rightarrow \mu_h \qquad (ii.)\frac{\sqrt{n}(\frac{\sum_{i=1}^{n} h(Y_i)}{n} - \mu_h)}{\sigma_h} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)$$

Theorem If h is continuous and $E|h(Y)|^{2+\delta} < \infty$ for some $\delta > 0$ then (i.) and (ii.) hold with Y_i replaced with $\tilde{Y}_i$

Home Page

Title Page

Page 17 of 32

Go Back

Full Screen

Close

Quit

Notes

- The fact that $P(Y_i \neq \tilde{Y}_i) = \mathcal{O}(i^{-1})$ crucial result

  Establish this by relating $P(Y_i \neq \tilde{Y}_i)$ to the $\mathcal{L}_1$ rate of convergence of T to C

  $\mathcal{L}_1$ rate of convergence of sample extrema is generally $\mathcal{O}(i^{-1})$

- Holder inequality
- Kronecker's lemma
- When using Monte Carlo to approximate an expectation, the ESUP sequence may be treated as if it were i.i.d.
- We used the most straightforward method of estimating C, all theorems apply when better estimates are employed

Home Page

Title Page

◀◀ ▶▶

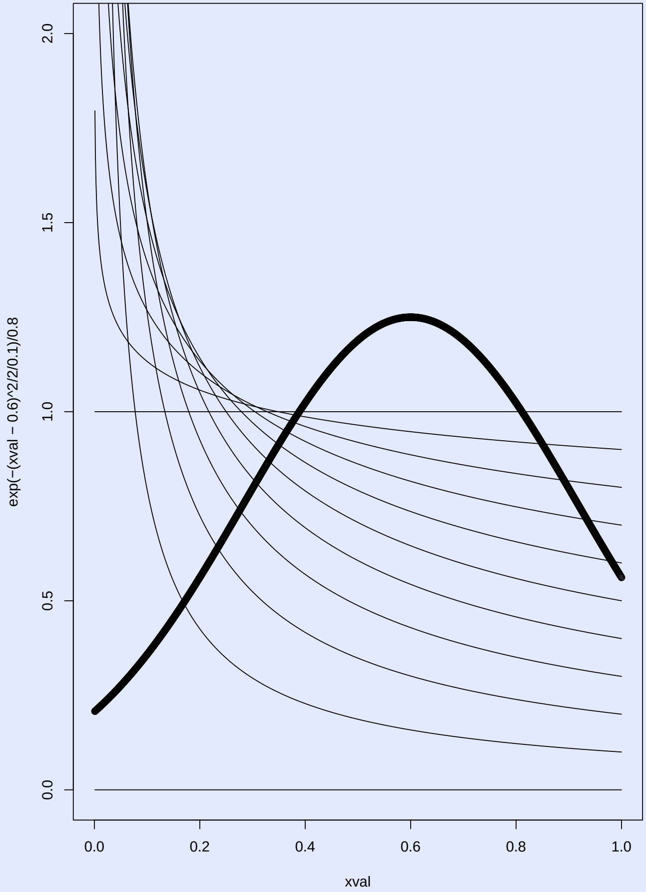
◀ ▶

Go Back

Full Screen

Close

Quit



A more accurate estimate of C

- Under smoothness assumptions, $P(Y_i \neq \tilde{Y}_i) = \mathcal{O}(i^{-2})$
- These same assumptions give us an asymptotic distribution for T
- We use this distribution to calculate an assymptotic $(1-\alpha)$ confidence upper bound for C
- As mentioned before, since the upper bound is larger than T all of the previous theory applies

Accounting for maximally lazy users

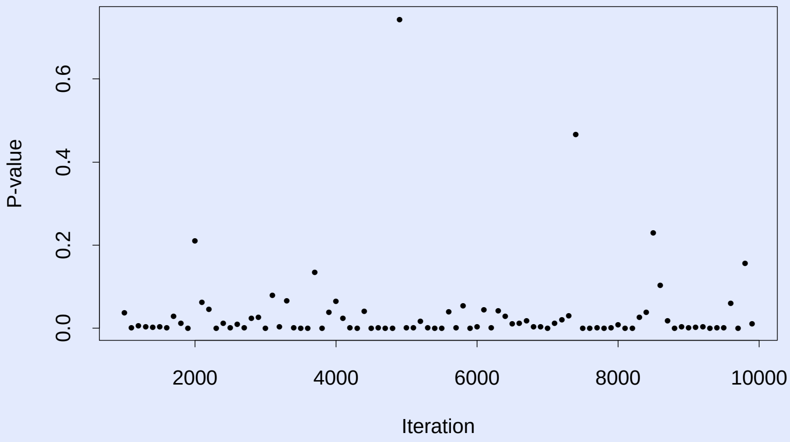
- A potential abuse is to use ESUP when in fact $C = \infty$
- Large sample behavior of *exceedences* follows a generalized Pareto distribution

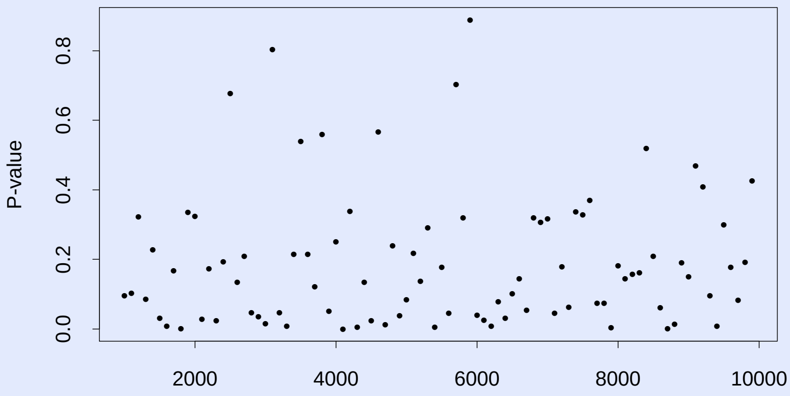
$$H(w) = \begin{cases} 1 - (1 + \kappa w/\sigma)_+^{-1/\kappa}, & \kappa \neq 0, \\ 1 - \exp(-w/\sigma), & \kappa = 0, \end{cases}$$

- κ is determined by whether or not $C = \infty$
- $\kappa < 0$ implies finite C
- Score for $\kappa = 0$ can be used to diagnose an infinite C
- Test statistic turns out to equivalent to Greenwood's statistic

$$G_s = \sum_{j=1}^{k} S_j^2$$

where $S_j = U_j - U_{j-1}$ for $U_1, \ldots, U_{k-1}$ a random sample from the $U(0,1)$ distribution


P-value
0.0
0.2
0.4
0.6
2000
4000
6000
8000
10000
Iteration


P-value
0.0
0.2
0.4
0.6
0.8
2000
4000
6000
8000
10000

Home Page
Title Page
Page 22 of 32
Go Back
Full Screen
Close
Quit

Example

- $Z|X \sim \text{Binomial}(n, p)$ with $\text{logit}(p) = X$
- $X \sim \text{N}(\alpha, \sigma^2)$
- Want to simulate from $X|Z$ which has density

$$f(x|z) \propto p^z(1-p)^{n-z}e^{(x-\alpha)^2/2\sigma^2}$$

- When several Z's represent proportions from several "small areas" random effects (X's) may be used to obtain shrinkage estimates (Agresti, Booth, Hobert and Caffo 2001)
- Sampling from $X|Z$ allows one to approximate an intractable "E" step of the EM algorithm

Home Page

Title Page

Page 23 of 32

Go Back

Full Screen

Close

Quit

Marginal Candidate

- Want to simulate from $X|Z$ which has density

$$f(x|z) \propto p^z(1-p)^{n-z}e^{(x-\alpha)^2/2\sigma^2}$$

- Recall $\text{logit}(p) = x$
- Why not choose $g(x) \propto e^{(x-\alpha)^2/2\sigma^2}$

$$C \propto \sup_{p\in[0,1]} p^z(1-p)^{n-z} = \left(\frac{z}{n}\right)^z \left(1-\frac{z}{n}\right)^{n-z}$$

- Turns out to be a very inaccurate candidate

Laplace Approximation

- Let $h_z(x) = \log f(x, z)$
- Let μ be such that $h'_z(\mu) = 0$.

$$
\begin{aligned}
E[X|Z = z] &= \frac{\int x f(x, z) dx}{\int f(x, z) dx} \\
&= \frac{\int x \exp(\log f(x, z)) dx}{\int \exp(\log f(x, z)) dx} \\
&\approx \frac{\int x \exp[h_z(\mu) + (x - \mu)^2 h''_z(\mu)/2] dx}{\int \exp[h_z(\mu) + (x - \mu)^2 h''_z(\mu)/2] dx} \\
&= \mu
\end{aligned}
$$

- Same thing to find $\mathrm{Var}(X|Z = z) \approx -(h''_z(\mu))^{-1}$
- Scaling and shifting a t distribution by the Laplace standard deviation and mean provides an extremely accurate candidate
- Difficult to calculate C

M		5		10	
Error		A	B	A	B
	t_3/Laplace candidate distribution				
ESUP	Mean	.05	0	0.03	0
	Median	0	0	0	0
ESUP UCL	Mean	0.03	0.24	0.01	0.23
	Median	0	0.2	0	0.2
	$\mathrm{N}(\alpha, \sigma^2)$ candidate distribution				
ESUP	Mean	0.49	0	0.33	0
	Median	0.4	0	0.3	0
ESUP UCL	Mean	0.20	0.01	0.10	0.01
	Median	0.2	0	0.1	0

Average and median error rates for various candidate distributions and sample sizes, M, for $z = 10$, $n = 30$ and 1,000 replications of M simulations from the new algorithm (ESUP) without and with upper confidence limit (ESUP UCL). A type A error occurs when a sampler incorrectly accepts a candidate it should have rejected and a type B error occurs when a sampler rejects a candidate it should have accepted.

Home Page

Title Page

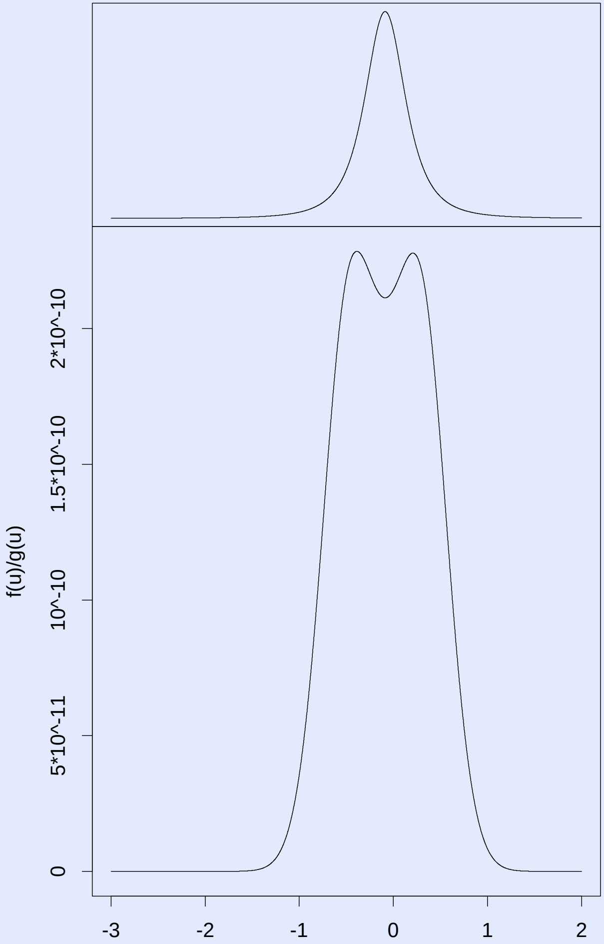
Page 26 of 32

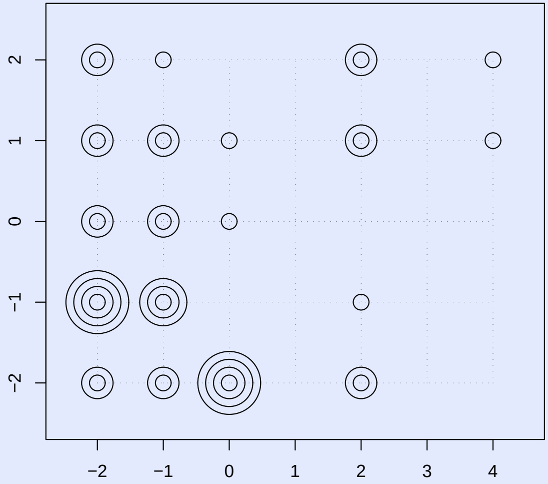
Go Back

Full Screen

Close

Quit




2
1
0
−1
−2
−2
−1
0
1
2
3
4

Home Page

Title Page

Page 28 of 32

Go Back

Full Screen

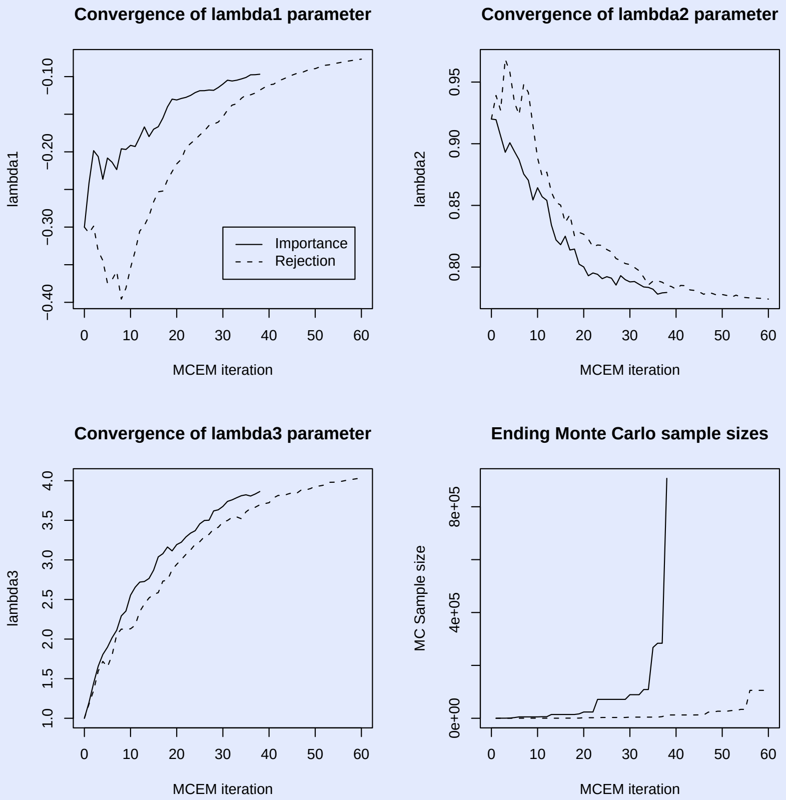
Close

Quit

Model

- Response at grid point x_i is labeled y_i for $i = 1, \dots, n$
- $y_i | u_i \sim \text{Poisson}(\mu_i)$
- $\log(\mu_i) = \lambda_1 + u_i$
- $\text{Cov}(u_i, u_j) = \lambda_2 \exp\{-\lambda_3 \|x_i - x_j\|\}$
- Monte Carlo EM estimation of $(\lambda_1, \lambda_2, \lambda_3)$ requires simulating from the joint distribution of the u_i given the y_i
- Efficient candidate distributions based on the Laplace approximation can be constructed but the exact supremum for such approximations cannot be calculated
- For higher dimensional data variants of the variable-at-a-time Metropolis/Hastings algorithm may be the only option

Home Page
Title Page
Page 29 of 32
Go Back
Full Screen
Close
Quit



Home Page

Title Page

Page 30 of 32

Go Back

Full Screen

Close

Quit

Other areas its worked so far

- MCEM for binary response GLMM
- Exploring posteriors via sequential simulation

$$\begin{array}{rc} X_1 & \text{Using ESUP} \\ X_2|X_1 & \text{Exactly} \\ X_3|X_2, X_1 & \text{Exactly} \end{array}$$

- MC exact conditional analysis of contingency table and logistic regression data

Future research

- Applications
- ?