Systematics in Monte Carlo

Galin Jones

University of Minnesota

April 2023

Systematics in Monte Carlo?

Systematics in Monte Carlo?

Effect of model error on MC

Effect of using nuisance parameters on MC

MC isn't producing what you think it is

MC hasn't been run long enough

Using the MC output suboptimally

Systematics in Monte Carlo?

Effect of model error on MC

Effect of using nuisance parameters on MC

MC isn't producing what you think it is

MC hasn't been run long enough

Using the MC output suboptimally

Systematics in Monte Carlo?

- Effect of model error on MC
- Effect of using nuisance parameters on MC
- MC isn't producing what you think it is
- **MC hasn't been run long enough**
- Using the MC output suboptimally

Often two sample sizes to think about:

- MC sample size, m
- Observed data sample size, n

Can we account for both sample sizes and dimension of the setting?

Monte Carlo

F is a given (complicated) probability distribution

Want to know something about θ, a vector of features of F

θ can consist of a mean, variance, median, quantiles, marginal densities, output of system, and so on

Monte Carlo

Simulate $X_1, \ldots, X_m$ (GOFMC, MCMC, ...)

Monte Carlo sample size, m, is "sufficiently large"

$\hat{\theta}_m = \hat{\theta}(X_1, \ldots, X_m) \approx \theta$

Example:

$$\theta = E_F(X) = \langle X \rangle_F = \int x F(dx)$$

Estimate with sample mean

$$\hat{\theta}_m = \frac{1}{m} \sum_{i=1}^{m} X_i \approx \theta$$

Assessing Monte Carlo Error

There is an unknown (multivariate) Monte Carlo error

$$\hat{\theta}_m - \theta$$

Use the approximate sampling distribution: For large m

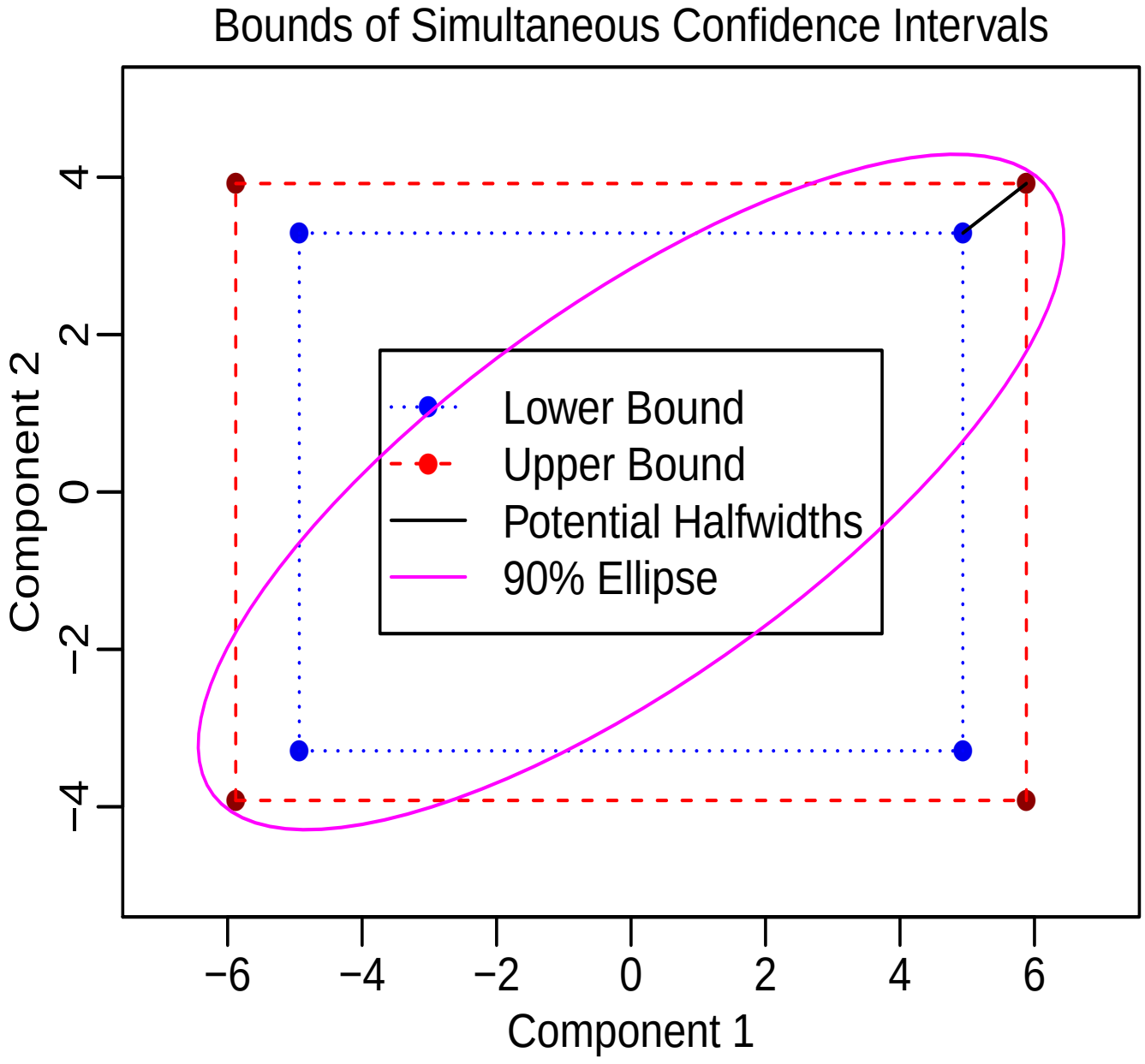
$$\sqrt{m}(\hat{\theta}_m - \theta) \approx \mathrm{N}(0, \Sigma)$$

Σ accounts for temporal dependence (MCMC) and dependence between components.

Need to estimate Σ to assess the simultaneous Monte Carlo error.

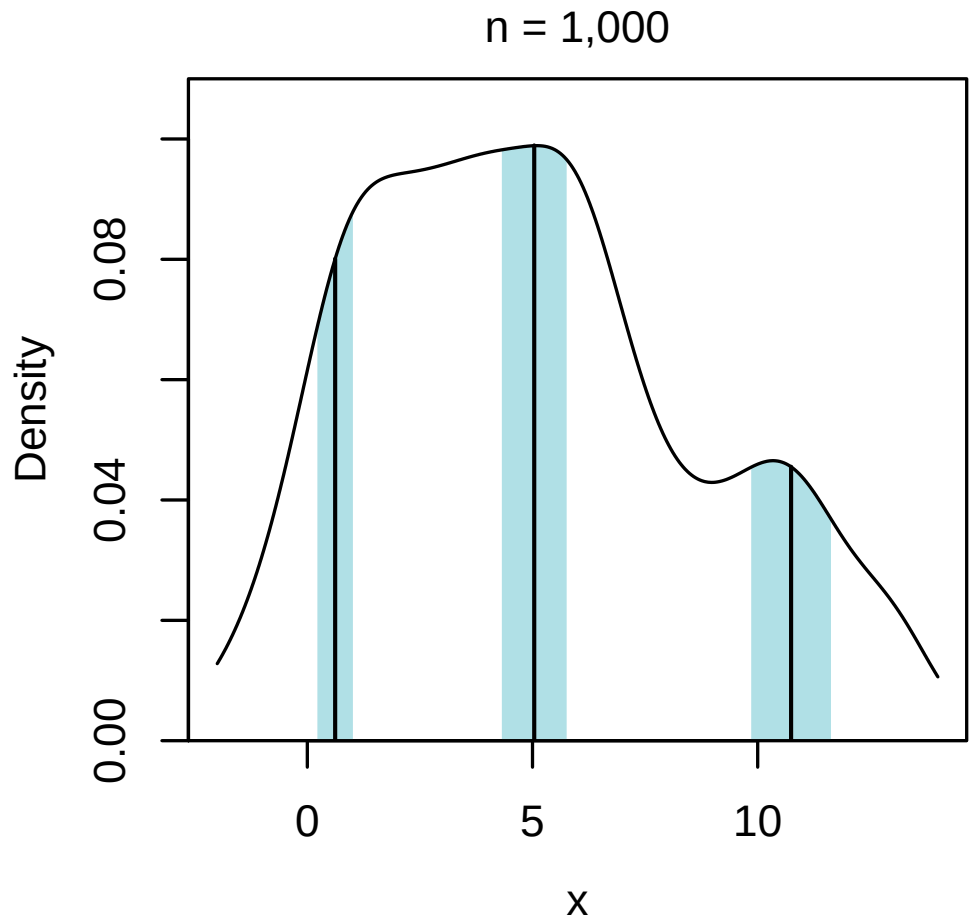
Vats et al 2019

Constructing Simultaneous Intervals



Illustrative Example: SimTools

Goal: Estimate the mean and the 0.1 and 0.9 quantiles for a marginal distribution.



Robertson et al (2020)

Producing a representative sample

GOFMC: Not really an issue almost by definition

MCMC: Not just for Bayesians

Metropolis-Hastings

Let f be a target density and q_h a proposal density.

Given $X_t = x$, draw $Y \sim q_h(\cdot \mid x)$

Draw $U \sim Unif(0, 1)$ and set $X_{t+1} = y$ if

$$u \leq \frac{f(y)q_h(x \mid y)}{f(x)q_h(y \mid x)}$$

otherwise set $X_{t+1} = x$

Choice of h is crucial to the finite-sample performance

When does MH fail?

If

$$A_h(x) = \int \left[\frac{f(y) q_h(x \mid y)}{f(x) q_h(y \mid x)} \wedge 1 \right] q_h(y \mid x) dx'.$$

then, for every x, the distance between the t-th step of the simulation and the target is bounded below by

$$[1 - A_h(x)]^t$$

Answer: Choose proposal scaling or other features so that we avoid $A_h(x) \approx 0$.

Brown and Jones (2023)

Gaussian Proposals

Suppose $\mu : \mathbb{R}^d \to \mathbb{R}^d$ and consider a proposal of the form

$$N_d(\mu(x), hC)$$

then

$$A_h(x) \leq \frac{1}{f(x)(2\pi h)^{d/2} \det(C)^{1/2}}$$

Remark: Suggests h must be small to avoid poor convergence properties and that many MH chains can have poor dimension dependence unless the scaling is chosen carefully.

If n is also large or d depends on n, then h will have to depend on both, but it's complicated–see Brown and Jones (2023)

Application: Approximating Likelihoods

Developed in statistics by Charlie Geyer in late 1980s

A density is often known up to a constant h_γ so that

$$f_\gamma(x) = \frac{1}{c(\gamma)} h_\gamma(x)$$

and the log-likelihood is

$$l(\gamma) = \log h_\gamma(x) - \log c(\gamma)$$

but

$$\nabla l(\gamma)$$

or profile likelihoods aren't available because $c(\gamma)$ is intractable

Monte Carlo Likelihood Approximation

Suppose α is arbitrary, but fixed and

$$g_\alpha(x) = \frac{1}{n(\alpha)} b_\alpha(x)$$

is a family that we can simulate from (GOFMC or MCMC). Then

$$E_\alpha \left[\frac{h_\gamma(x)}{b_\alpha(x)} \right] = \int \frac{h_\gamma(x)}{b_\alpha(x)} g_\alpha(x) dx = \frac{c(\gamma)}{n(\alpha)}$$

so estimate it with

$$\frac{1}{m} \sum_{i=1}^{m} \frac{h_\gamma(x_i)}{b_\alpha(x_i)}$$

Monte Carlo Likelihood Approximation

Monte Carlo approximation to $l(\gamma)$:

$$l_m(\gamma) = \log \frac{h_\gamma(x)}{b_\alpha(x)} - \frac{1}{m} \sum_{i=1}^{m} \frac{h_\gamma(x_i)}{b_\alpha(x_i)}$$

MC-MLE converges to MLE and is asymptotically normal

Profile MC-likelihoods converge to the profile likelihood

and so on...

Geyer (1990) and for an application see Knudson et al (2020)

Monte Carlo Likelihood Approximation

In particular, for large m, n

$$\hat{\gamma}_{m,n} \approx N\left(\gamma^*, \frac{J^{-1}VJ^{-1}}{n} + \frac{J^{-1}WJ^{-1}}{m}\right)$$

J is minus the expectation of the second derivative of the log likelihood

V is the variance of the score function

W is the variance of the deviation of the score from its Monte Carlo approximation

This does not require the model be correct, but there is no free lunch. . .

Sung and Geyer (2007)

Finally...

What can we do to assess MC results in settings where the simulation is extremely costly and only a small number of samples can be obtained?

There is a lot of current work on high-dimensional (large n and d) Monte Carlo, I've only scratched the surface.

Incorporating MC error and observational error due to m and n simultaneously seems hard in general. Again, we've only scratched the surface.