

# Monte Carlo Statistical Methods

STA 6934

George Casella

[casella@stat.ufl.edu](mailto:casella@stat.ufl.edu)

# 1 Introduction

## 1.1 Statistical Models

## 1.2 Likelihood Methods

## 1.3 Bayesian Methods

## 1.4 Deterministic Numerical Methods

## 1.5 Simulation versus numerical analysis

- Experimenters choice before fast computers:
  - Describe an accurate model which usually precludes computation of explicit answers
  - **or** Choose a standard model which would allow such computations, but may not be a close representation of a realistic model.
- Such problems contributed to the development of **simulation-based inference**

## 1.1 Statistical Models

---

**Example 1 – Censored data models**– Missing data models where densities are not sampled directly.

Typical simple statistical model: we observe

$$Y_1, \dots, Y_n \sim f(y|\theta).$$

The distribution of the sample given by the product

$$\prod_{i=1}^n f(y_i|\theta)$$

Inference about  $\theta$  based on this likelihood.

With *censored* random variables, actual observations:

$$Y_i^* = \min\{Y_i, \bar{u}\}$$

where  $\bar{u}$  is censoring point.

Inference about  $\theta$  based on the censored likelihood.

For instance, if

$$X \sim \mathcal{N}(\theta, \sigma^2) \quad \text{and} \quad Y \sim \mathcal{N}(\mu, \rho^2),$$

the variable

$$Z = X \wedge Y = \min(X, Y)$$

is distributed as

$$\begin{aligned} & \left[ 1 - \Phi\left(\frac{z - \theta}{\sigma}\right) \right] \times \rho^{-1} \varphi\left(\frac{z - \mu}{\rho}\right) \\ & + \left[ 1 - \Phi\left(\frac{z - \mu}{\rho}\right) \right] \sigma^{-1} \varphi\left(\frac{z - \theta}{\sigma}\right) \end{aligned}$$

where  $\varphi$  and  $\Phi$  are the density and cdf of the normal  $\mathcal{N}(0, 1)$  distribution.

Similarly, if

$$X \sim \text{Weibull}(\alpha, \beta),$$

with density

$$f(x) = \alpha\beta x^{\alpha-1} \exp(-\beta x^\alpha)$$

the censored variable

$$Z = X \wedge \omega, \quad \omega \text{ constant,}$$

has the density

$$f(z) = \alpha\beta z^\alpha e^{-\beta z^\alpha} \mathbf{1}_{z \leq \omega} + \left( \int_\omega^\infty \alpha\beta x^\alpha e^{-\beta x^\alpha} dx \right) \delta_\omega(z),$$

where  $\delta_a(\cdot)$  Dirac mass at  $a$ .

## Example 2 – Mixture models—

Models of *mixtures of distributions*:

$X \sim f_j$  with probability  $p_j$ ,

for  $j = 1, 2, \dots, k$ , with overall density

$$X \sim p_1 f_1(x) + \dots + p_k f_k(x).$$

For a sample of independent random variables  $(X_1, \dots, X_n)$ , sample density

$$\prod_{i=1}^n \{p_1 f_1(x_i) + \dots + p_k f_k(x_i)\}.$$

Expanding this product involves  $k^n$  elementary terms: prohibitive to compute in large samples.

## 1.2 Likelihood Methods

---

### Maximum Likelihood Methods

- For an iid sample  $X_1, \dots, X_n$  from a population with density  $f(x|\theta_1, \dots, \theta_k)$ , the *likelihood function* is

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= L(\theta_1, \dots, \theta_k | x_1, \dots, x_n) \\ &= \prod_{i=1}^n f(x_i | \theta_1, \dots, \theta_k). \end{aligned}$$

- Global justifications from asymptotics

### Example 3 –Gamma MLE–

$X_1, \dots, X_n$  iid observations from gamma density

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta},$$

where  $\alpha$  is known.

*Log likelihood*

$$\begin{aligned} \log L(\alpha, \beta|x_1, \dots, x_n) &= \log \prod_{i=1}^n f(x_i|\alpha, \beta) \\ &= \log \prod_{i=1}^n \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-x_i/\beta} \\ &= -n \log \Gamma(\alpha) \\ &\quad -n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^n \log x_i - \sum_{i=1}^n x_i/\beta. \end{aligned}$$

Solving

$$\frac{\partial}{\partial \beta} \log L(\alpha, \beta | x_1, \dots, x_n) = 0$$

is straightforward

Yields the MLE

$$\hat{\beta} = \sum_{i=1}^n x_i / (n\alpha).$$

When  $\alpha$  also unknown, additional equation

$$\frac{\partial}{\partial \alpha} \log L(\alpha, \beta | x_1, \dots, x_n) = 0$$

is particularly nasty!

Involve difficult computations (incl. derivative of the gamma function, the *digamma function*)

Explicit solution no longer possible

**Example 4 – Student’s  $t$  distribution –**

Reasonable alternative to normal errors

$$\mathcal{T}(p, \theta, \sigma)$$

more “robust” against possible modeling errors

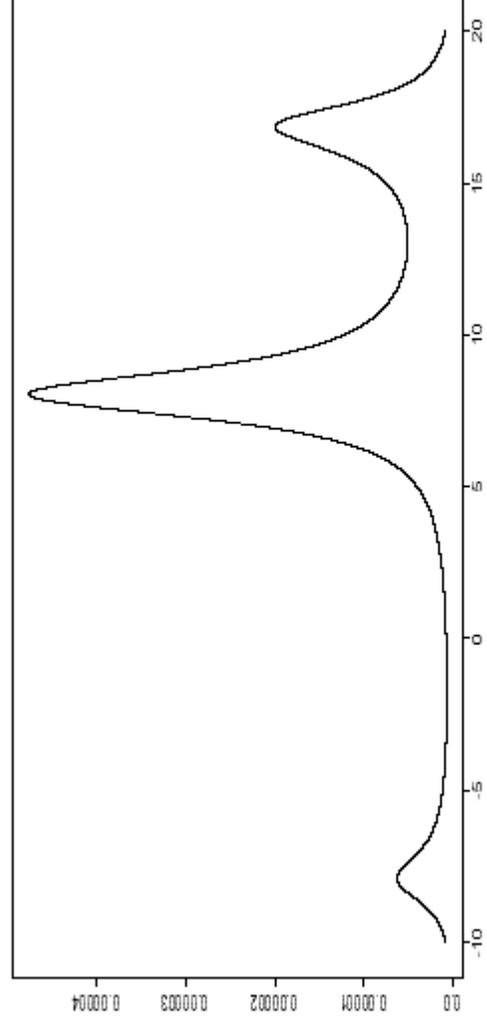
Density of  $\mathcal{T}(p, \theta, \sigma)$  proportional to

$$\sigma^{-1} \left( 1 + \frac{(x - \theta)^2}{p\sigma^2} \right)^{-(p+1)/2}$$

When  $p$  known and  $\theta$  and  $\sigma$  both unknown, the likelihood

$$\sigma^{n\frac{p+1}{2}} \prod_{i=1}^n \left( 1 + \frac{(x_i - \theta)^2}{p\sigma^2} \right)$$

may have  $n$  local minima, each of which needs to be calculated to determine the global maximum.



**Multiplicity of modes of the likelihood from  $C(\theta, 1)$  when  $n = 3$  and  $x_1 = 0, x_2 = 5, x_3 = 9$ .**

### Example 5 –Mixtures again–

For a mixture of two normal distributions,

$$p\mathcal{N}(\mu, \tau^2) + (1 - p)\mathcal{N}(\theta, \sigma^2),$$

likelihood proportional to

$$\prod_{i=1}^n \left[ p\tau^{-1} \varphi \left( \frac{x_i - \mu}{\tau} \right) + (1 - p) \sigma^{-1} \varphi \left( \frac{x_i - \theta}{\sigma} \right) \right]$$

containing  $2^n$  terms.

Standard maximization techniques often fail to find the global maximum because of multimodality of the likelihood function.

In the special case

$$f(x|\mu, \sigma) = (1-\epsilon) \exp\{(-1/2)x^2\} + \frac{\epsilon}{\sigma} \exp\{(-1/2\sigma^2)(x-\mu)^2\} \quad (1)$$

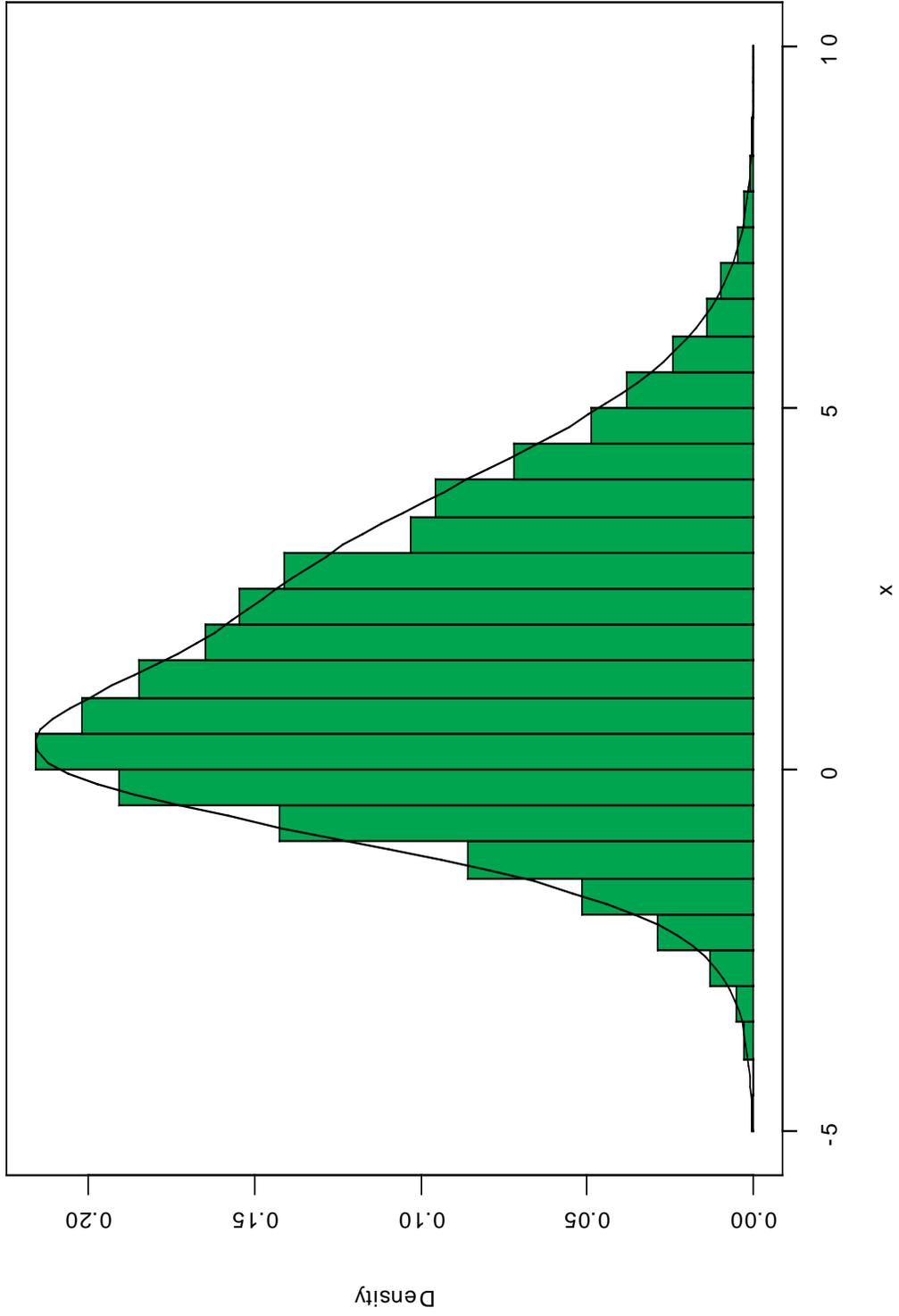
with  $\epsilon > 0$  known

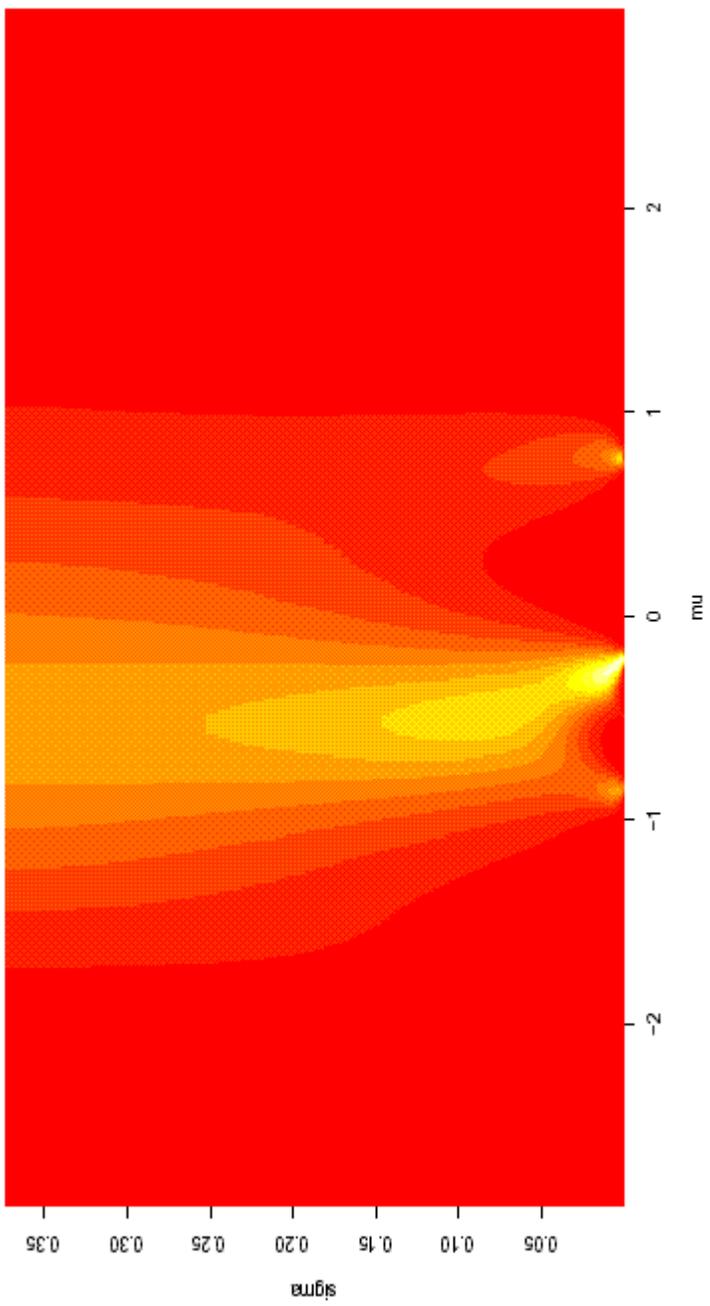
Then, whatever  $n$ , the likelihood is unbounded:

$$\lim_{\sigma \rightarrow 0} \ell(\mu = x_1, \sigma | x_1, \dots, x_n) = \infty$$

R program -> normal\_mixture1

# Mixture of normals





Likelihood of (1)

### 1.3 Bayesian Methods

---

In the Bayesian paradigm, information brought by the data  $x$ , realization of

$$X \sim f(x|\theta),$$

combined with prior information specified by *prior distribution* with density  $\pi(\theta)$

Summary in a probability distribution,  $\pi(\theta|x)$ , called the **posterior distribution**

Derived from the *joint* distribution  $f(x|\theta)\pi(\theta)$ , according to

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta},$$

[Bayes Theorem]

where

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta$$

is the *marginal density* of  $X$

### **Example 6 –Binomial Bayes Estimator–**

For an observation  $X$  from the binomial distribution  $\mathcal{B}(n, p)$  the (so-called) conjugate prior is the family of beta distributions  $\mathcal{Be}(a, b)$

The classical Bayes estimator  $\delta^\pi$  is the posterior mean

$$\begin{aligned}\delta^\pi &= \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(n-x+b)} \\ &\quad \times \int_0^1 p p^{x+a-1} (1-p)^{n-x+b-1} dp \\ &= \frac{x+a}{a+b+n}.\end{aligned}$$

## The curse of conjugate priors

---

The use of **conjugate priors** for computational reasons

- implies a restriction on the modeling of the available prior information
- may be detrimental to the usefulness of the Bayesian approach
- gives an impression of subjective manipulation of the prior information disconnected from reality.

### **Example 7 –Logistic Regression–**

Standard regression model for binary (0 – 1) responses:  
distribution of  $Y \in \{0, 1\}$  modeled by

$$P(Y = 1) = p = \frac{\exp(x^t \beta)}{1 + \exp(x^t \beta)}.$$

Equivalently, the *logit* transform of  $p$ ,

$$\text{logit}(p) = \log[p/(1 - p)]$$

satisfies  $\text{logit}(p) = x^t \beta$ .

Computation of a confidence region on  $\beta$  quite delicate when  $\pi(\beta|x)$  not explicit.

In particular, when the confidence region involves only one component of a vector parameter, calculation of  $\pi(\beta|x)$  requires the integration of the joint distribution over all the other parameters.

### Example 8 –Cauchy confidence regions–

$X_1, \dots, X_n$  an iid sample from the Cauchy distribution  $\mathcal{C}(\theta, \sigma)$ , with prior

$$\pi(\theta, \sigma) = \sigma^{-1}.$$

Confidence region on  $\theta$  then based on

$$\pi(\theta|x_1, \dots, x_n) \propto \int_0^\infty \sigma^{-n-1} \prod_{i=1}^n \left[ 1 + \left( \frac{x_i - \theta}{\sigma} \right)^2 \right]^{-1} d\sigma,$$

an integral which cannot be evaluated explicitly.

Similar computational problems with likelihood estimation in this model.

## 1.4 Deterministic Numerical Methods

---

To solve an equation of the form

$$f(x) = 0,$$

the *Newton-Raphson* algorithm produces a sequence  $x_n$ :

$$x_{n+1} = x_n - \left( \left. \frac{\partial f}{\partial x} \right|_{x=x_n} \right)^{-1} f(x_n)$$

that converges to a solution of  $f(x) = 0$ .

[Note that  $\frac{\partial f}{\partial x}$  is a matrix in multidimensional settings.]

## Example: A Simple Application of Newton Raphson

- <sup>2</sup> Find the square root of a number, <sup>1</sup>.
- <sup>2</sup> This is equivalent to finding  
the root of the simple equation

2 The first derivative is

$$\frac{\partial}{\partial x} f(x) = 2x :$$



$$\begin{aligned} x^{(j+1)} &= x^{(j)} \cdot \frac{f(x^{(j)})}{f'(x^{(j)})} \\ &= x^{(j)} \cdot \frac{x^{(j)2} \cdot 1}{2x^{(j)}} \\ &= \frac{1}{2} (x^{(j)} + \frac{1}{x^{(j)}}) : \end{aligned}$$

Optimization of smooth functions  $F$  done using the equation

$$\nabla F(x) = 0,$$

where  $\nabla F$  denotes the *gradient* of  $F$ , vector of derivatives of  $F$ . The corresponding techniques are *gradient methods*, where the sequence  $x_n$  is

$$x_{n+1} = x_n - (\nabla \nabla^t F)^{-1}(x_n) \nabla F(x_n),$$

where  $\nabla \nabla^t F$  denotes the matrix of second derivatives of  $F$ .

Numerical computation of an integral

$$I = \int_a^b h(x) dx$$

can be done by

- *Riemann integration*
- **or** by improved techniques like the *trapezoidal rule*

$$\hat{I} = \frac{1}{2} \sum_{i=1}^{n-1} (x_{i+1} - x_i)(h(x_i) + h(x_{i+1})),$$

where the  $x_i$ 's constitute an ordered partition of  $[a, b]$ ,

- **or** *Simpson's rule*

$$\tilde{I} = \frac{\delta}{3} \left\{ f(a) + 4 \sum_{i=1}^n h(x_{2i-1}) + 2 \sum_{i=1}^n h(x_{2i}) + f(b) \right\}$$

## 1.5 Simulation versus numerical analysis: when is it useful?

---

- numerical methods do not take into account the probabilistic aspects of the problem, numerical integration often focus on regions of low probability
- occurrence of local modes of a likelihood often cause more problems for a deterministic gradient method than for simulation methods

- **but** simulation methods very rarely take into account the specific analytical form of the functions (For instance, because of the randomness induced by the simulation, a gradient method yields a much faster determination of the mode of a unimodal density)
- For small dimensions, integration by Riemann sums or by quadrature converges much faster than the mean of a simulated sample.
- It is thus often reasonable to use a numerical approach when dealing with regular functions in small dimensions

- When the statistician
  - needs to study the details of a likelihood surface or posterior distribution
  - needs to simultaneously estimate several features of these functions
  - or when the distributions are highly multimodalit is preferable to use a simulation-based approach

- Fruitless to advocate the superiority of one method over the other
- More reasonable to justify the use of simulation-based methods by the statistician in terms of his/her *expertise*. The intuition acquired by a statistician in his/her every-day processing of random models can be directly exploited in the implementation of simulation techniques