Monte Carlo Statistical Methods

George Casella
University of Florida
January 3, 2008
casella@stat.ufl.edu

Based on

- Monte Carlo Statistical Methods,
  Christian Robert and George Casella,
  2004, Springer-Verlag
- Programming in R (available as a free download from
  http://www.r-project.org
- Also WinBugs, available free from
  http://www.mrc-bsu.cam.ac.uk/bugs/
- R programs for the course available at
  http://www.stat.ufl.edu/ casella/mcsm/

Introduction

- Statistical Models
- Likelihood Models
- Bayesian Models
- Deterministic Numerical Models
- Simulation vs. Numerical Methods

1.1 Statistical Models

- In a typical statistical model we observe

$$Y_1, Y_2, \ldots, Y_n \sim f(y|\theta)$$

- The distribution of the sample is given by the product, the likelihood function

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

- Inference about θ is based on this likelihood.
- In many situations the likelihood can be complicated

Example 1.1: Censored Random Variables

- If
$$X_1 \sim N(\theta, \sigma^2), \quad X_2 \sim N(\mu, \rho^2),$$
- the distribution of $Y = \min\{X_1, X_2\}$ is
$$\left[1 - \Phi\left(\frac{y-\theta}{\sigma}\right)\right] \times \rho^{-1}\phi\left(\frac{y-\mu}{\rho}\right) + \left[1 - \Phi\left(\frac{y-\mu}{\rho}\right)\right] \times \sigma^{-1}\phi\left(\frac{y-\theta}{\sigma}\right),$$
where Φ and ϕ are the cdf and pdf of the normal distribution.
- This results in a complex likelihood.

Example 1.2: Mixture Models

- Models of *mixtures of distributions*:

  $$X \sim f_j \text{ with probability } p_j,$$

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

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

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

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

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

Example 1.2 : Normal Mixtures

- For a mixture of two normal distributions,

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

- The 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.

- **R program → normal-mixture1**

```
#This gives the distribution of the mixture of two normals#
e<-.3; nsim<-1000;m<-2;s<-1;

u<-(runif(nsim)<e);z<-rnorm(nsim)

z1<-rnorm(nsim,mean=m,sd=s)

#This plots histogram and density#
hist(u*z+(1-u)*z1,xlab="x",xlim=c(-5,5),freq=F,
    col="green",breaks=50,)

mix<-function(x)e*dnorm(x)+(1-e)*dnorm(x,mean=m,sd=s)

xplot<-c(-50:50)/10

par(new=T)
plot(xplot,mix(xplot), xlim=c(-5,5),type="l",yaxt="n",ylab="")
```

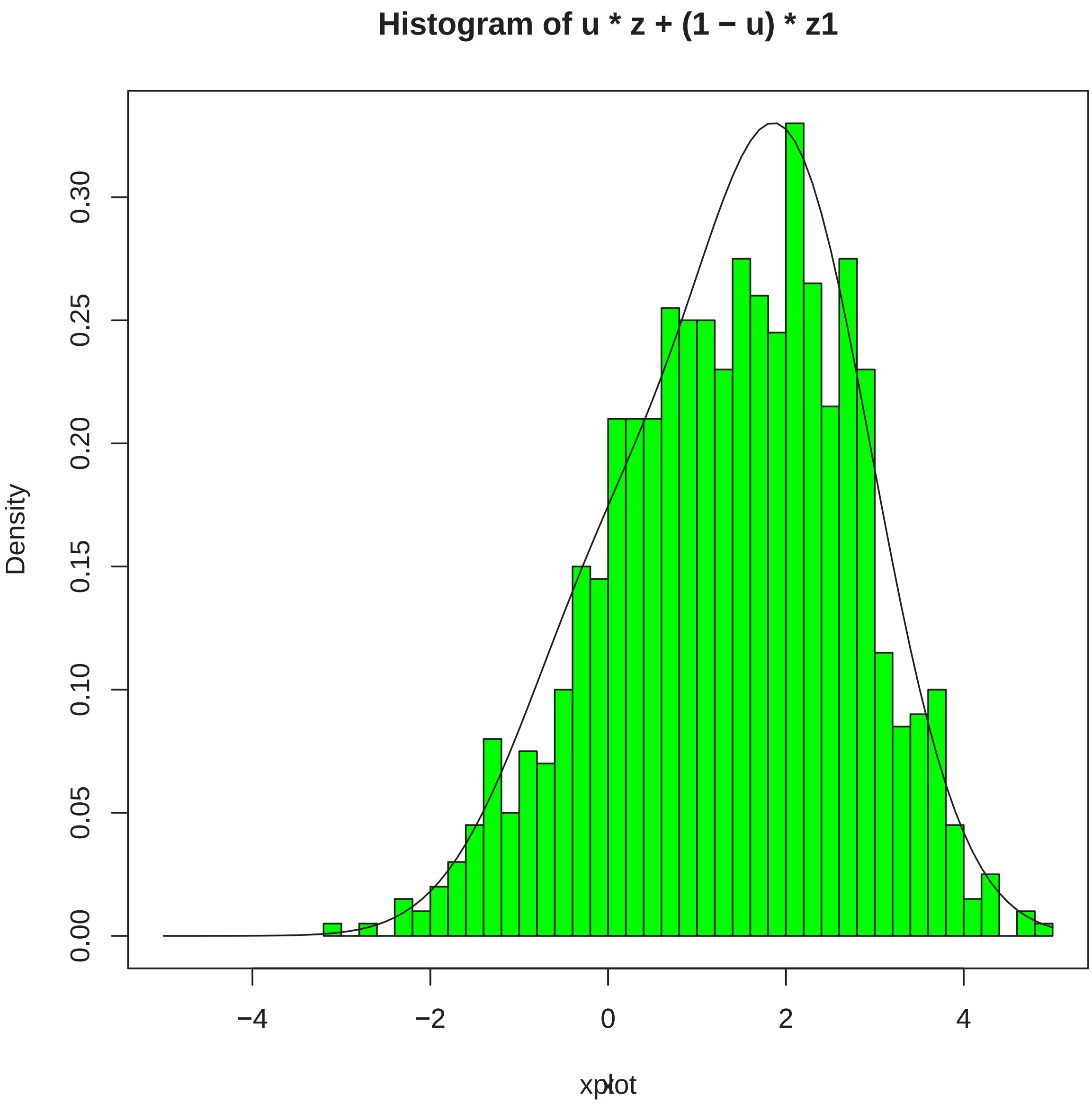


Figure 1: Histogram and density of normal mixture

1.2: Likelihood Methods

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

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

  - Global justifications from asymptotics

Example 1.9: Student's t distribution

- Reasonable alternative to normal errors is Student's t distribution, denoted by
$$\mathcal{T}(p, \theta, \sigma)$$
more "robust" against possible modelling errors

- Density of $\mathcal{T}(p, \theta, \sigma)$ proportional to
$$\sigma^{-1} \left(1 + \frac{(x-\theta)^2}{p\sigma^2}\right)^{-(p+1)/2} \quad ,$$

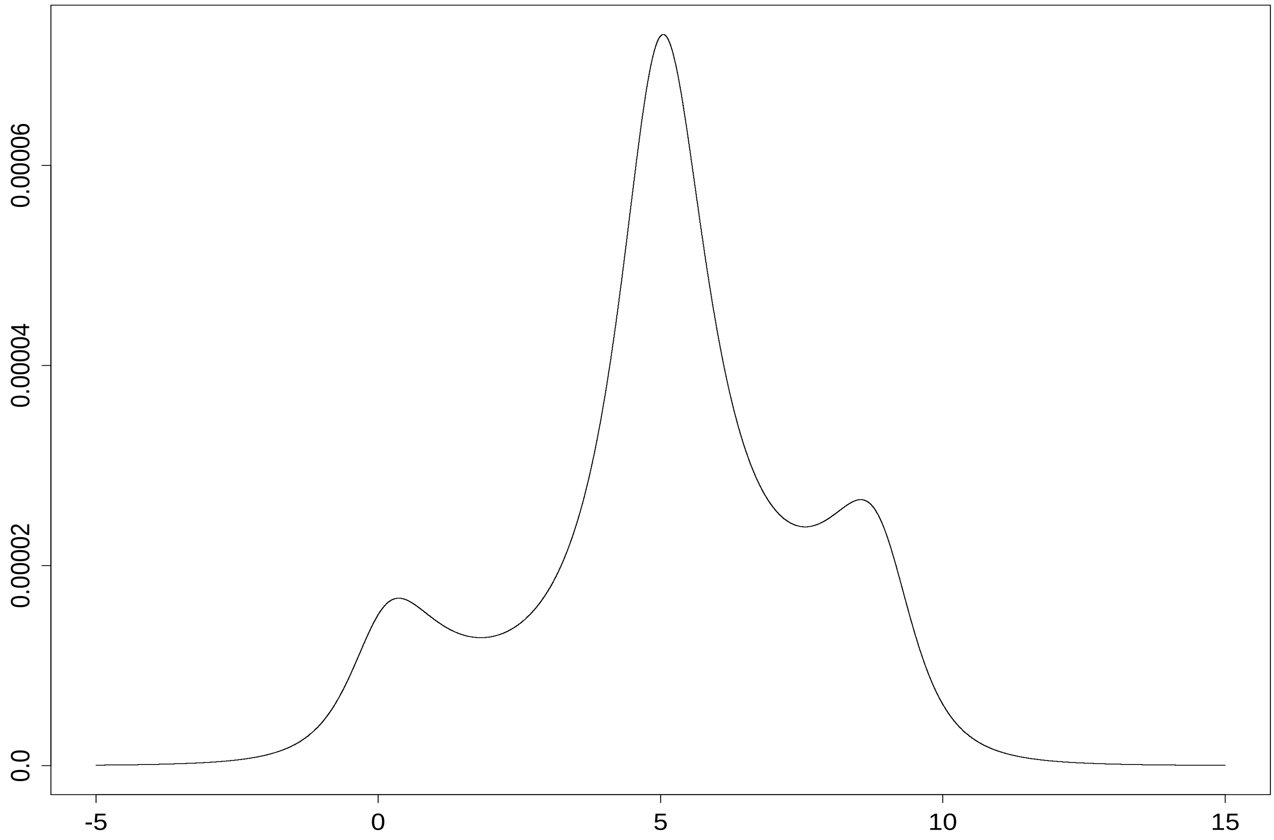
Example 1.9: Student's t distribution

- When p known and θ and σ 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.



- Illustration of the multiplicity of modes of the likelihood from a Cauchy distribution $\mathcal{C}(\theta, 1)$ $(p = 1)$ when $n = 3$ and $X_1 = 0$, $X_2 = 5$, and $X_3 = 9$.

Section 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 $\pi(\theta)$

Bayesian Methods

- 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 1.11: Binomial Bayes Estimator

- For an observation X from the binomial distribution Binomial(n, p) the (so-called) conjugate prior is the family of beta distributions Beta(a, b)
- The classical Bayes estimator δ^π is the posterior mean

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

- A Biased estimator of p

The Variance/Bias Trade-off

- Bayes Estimators are biased
- Mean Squared Error (MSE) = Variance + Bias^2
  - MSE $= \mathrm{E}(\delta^\pi - p)^2$
  - Measures average closeness to parameter
- Small Bias $\uparrow$ can yield large Variance $\downarrow$.

$$\delta^\pi = \frac{n}{a+b+n}\left(\frac{x}{n}\right) + \frac{a+b}{a+b+n}\left(\frac{a}{a+b}\right)$$

$$\mathrm{Var}\delta^\pi = \left(\frac{n}{a+b+n}\right)^2 \mathrm{Var}\left(\frac{x}{n}\right)$$

Conjugate Priors

- A prior is conjugate if

$$\pi(\theta)(\text{the prior}) \text{ and } \pi(\theta|x)(\text{the posterior})$$

  are in the same family of distributions.

- Examples
  - $\pi(\theta)$ normal , $\pi(\theta|x)$ normal
  - $\pi(\theta)$ beta , $\pi(\theta|x)$ beta
- Restricts the choice of prior
- Typically non-robust
- Originally used for computational ease

Example 1.13: Logistic Regression

- Standard regression model for binary $(0-1)$ responses: the *logit model* where distribution of Y modelled by

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

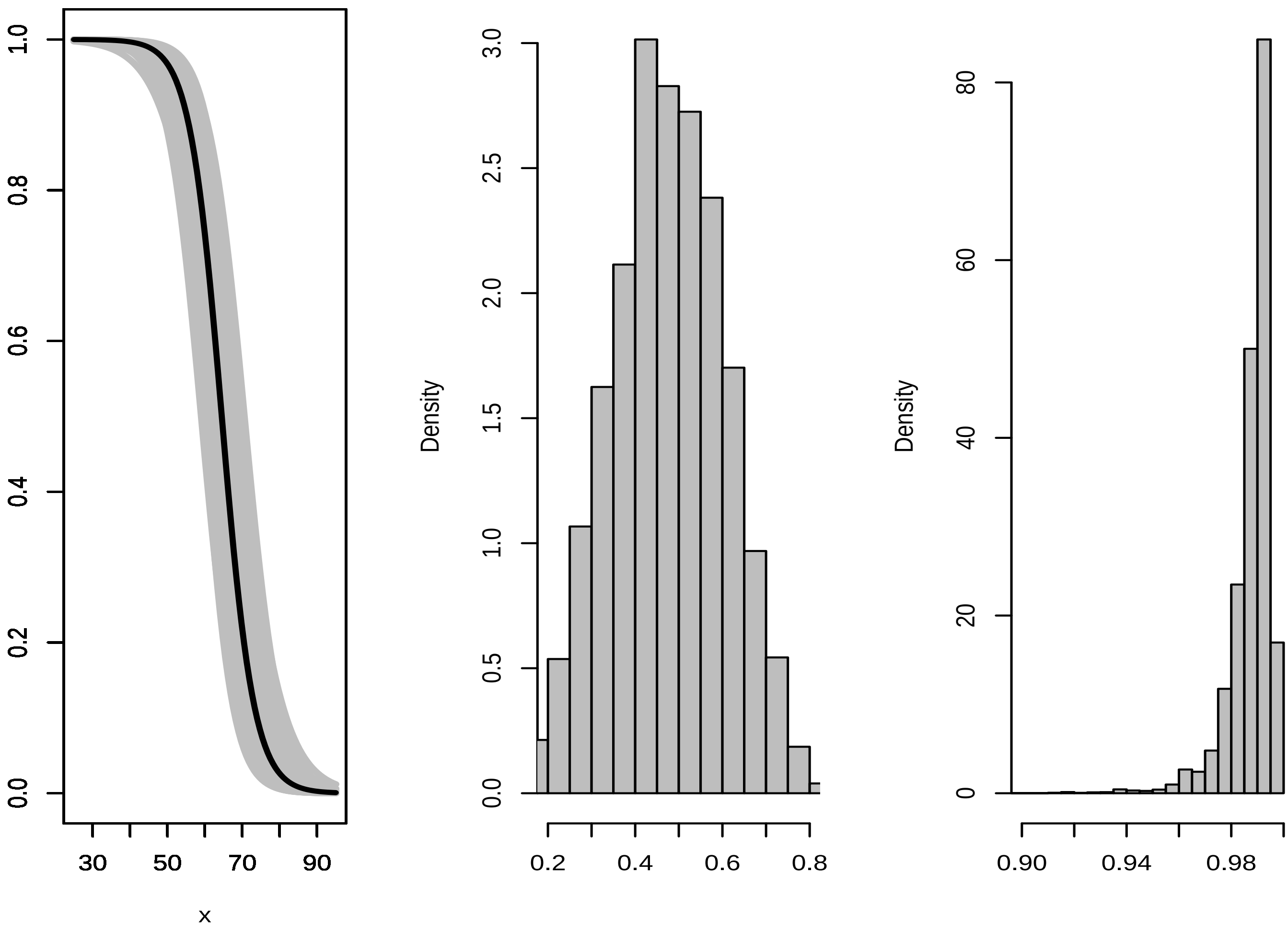
- Equivalently, the *logit* transform of p, $\text{logit}(p)=\log[p/(1-p)]$, satisfies $\text{logit}(p)=x^t\beta$.
- Computation of a confidence region on β quite delicate when $\pi(\beta|x)$ not explicit.
- In particular, when the confidence region involves only one component of a vector parameter, calculation of $\pi(\beta|x)$ requires the integration of the joint distribution over all the other parameters.

Challenger Data

- In 1986, the space shuttle Challenger exploded during take off, killing the seven astronauts aboard.
- The explosion was the result of an *O-ring* failure.

Flight No.	14	9	23	10	1	5	13	15	4	3	8	17
Failure	1	1	1	1	0	0	0	0	0	0	0	0
Temp.	53	57	58	63	66	67	67	67	68	69	70	70
Flight No.	2	11	6	7	16	21	19	22	12	20	18	
Failure	1	1	0	0	0	1	0	0	0	0	0	
Temp.	70	70	72	73	75	75	76	76	78	79	81	

- It is reasonable to fit a logistic regression, with p = probability of an O-ring failure and x = temperature.



- The left panel shows the average logistic function and variation
- The middle panel shows predictions of failure probabilities at 65^o Fahrenheit
- The right panel shows predictions of failure probabilities at 45^o Fahrenheit.