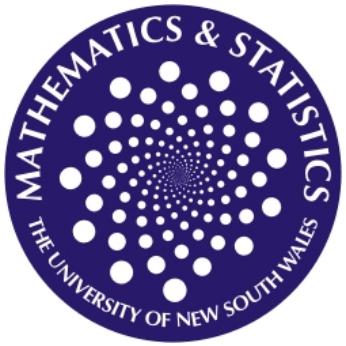


Lecture 3: More on Priors, Models & Monte Carlo

sem 2, 2018

August 13, 2018

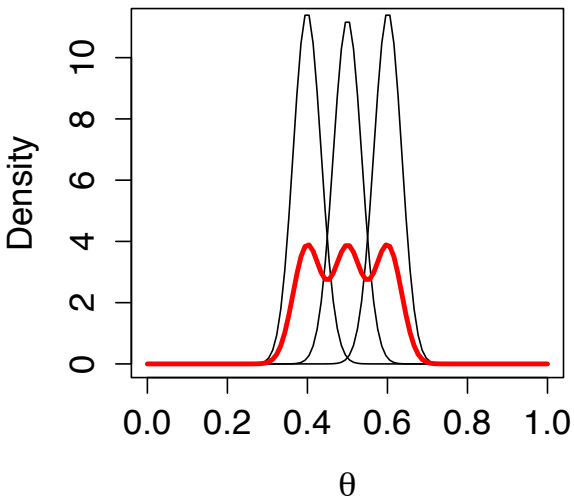
Outline

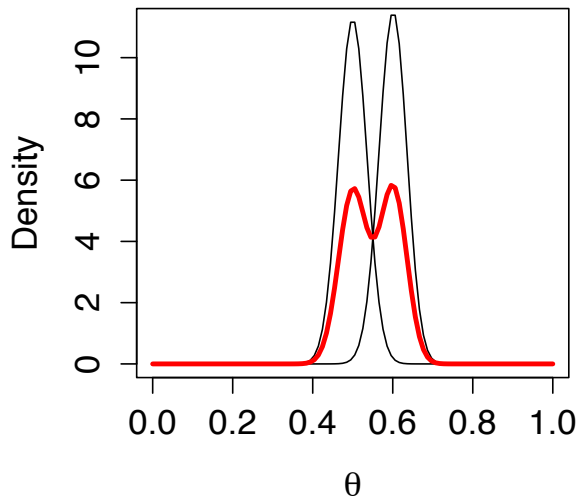
1. Mixture priors, prior predictive checking.
2. Multivariate models
3. Monte Carlo methods
    - Monte Carlo integration, rejection sampling.

Outline

1. Mixture priors, prior predictive checking.
2. Multivariate models
3. Monte Carlo methods
    - Monte Carlo integration, rejection sampling.

Mixtures of Priors





Conjugate priors:

- Mathematically convenient
- May not be sufficiently flexible to capture prior beliefs

Example:

Construct a prior for probability of H when spinning a coin on a table.

- Tossing coins give good chance of 0.5 prob
- When spinning, coin edge imperfections cause preference for H or T
- $\Rightarrow$ reasonable prior may favour prob ≈ 0.4 or 0.6
- May also include prob $= 0.5$.
- i.e. Prior belief may be bi-modal or tri-modal

Model is: $X \mid \theta \sim \mathsf{Bin}(n, \theta)$ so conjugate prior is Beta.
But Beta does not permit this kind of prior.

Mixtures of Priors

One solution is to use mixtures of conjugate distributions.

Suppose $\pi_1(\theta), \ldots, \pi_k(\theta)$ are all conjugate priors for θ, leading to posterior distributions $\pi_1(\theta \mid x), \ldots, \pi_k(\theta \mid x)$.

Consider the family of mixture distributions:

$$\pi(\theta) = \sum_{i=1}^{k} w_i \pi_i(\theta)$$

as a prior distribution. Then

$$\begin{aligned} \pi(\theta \mid x) &\propto \pi(\theta) L(x \mid \theta) \\ &= \textstyle\sum_{i=1}^{k} w_i \pi_i(\theta) L(x \mid \theta) \\ &\propto \textstyle\sum_{i=1}^{k} w_i^* \pi_i(\theta \mid x) \end{aligned}$$

so the posterior is in the same conjugate mixture family.

Allows more flexibility for conjugate priors.

Prior predictive checking

You have specified a prior $\pi(\theta)$. Is there a way to check that this represents your prior beliefs about the data?

Posterior predictive distribution:

$$\pi(y \mid x) = \int \pi(y \mid \theta, x)\pi(\theta \mid x)\mathrm{d}\theta$$

= distribution of future data y given observed data x.

Prior predictive distribution:

$$\pi(y) = \int \pi(y \mid \theta)\pi(\theta)\mathrm{d}\theta$$

= distribution of future data y.

Prior predictive procedure

- Generate samples from $\pi(y)$ given your choice of $\pi(\theta)$.
- Check if $\pi(y)$ represents your beliefs about the possible data you might observe under the model.
- If not, modify $\pi(\theta)$ and repeat.

Prior predictive checking

Example: Suppose $\theta \sim \mathsf{Beta}(\alpha, \beta)$ is the prob of an individual having a rare genetic disorder. $X \sim \mathsf{Bin}(n, \theta)$ is the number of people (out of n independently chosen individuals) that have the disorder.

Prior predictive distribution

$$\pi(y) \propto \int \mathsf{Ber}(\theta)\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta, \quad y = 0, 1.$$

So (algebraically)

$$\begin{aligned}
\mathbb{P}(Y=0) &\propto \int \theta \times \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta = \mathrm{B}(\alpha+1, \beta) \\
\mathbb{P}(Y=1) &\propto \int (1-\theta) \times \theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta = \mathrm{B}(\alpha, \beta+1)
\end{aligned}$$

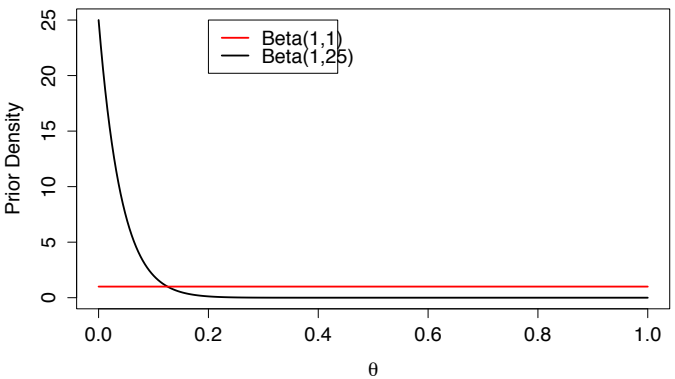
where $\mathrm{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

$\mathbb{P}(Y=0) = \mathrm{B}(\alpha+1, \beta)/[\mathrm{B}(\alpha+1, \beta) + \mathrm{B}(\alpha, \beta+1)] = 1 - \mathbb{P}(Y=1)$.

Prior predictive checking

Example (cont.):

	True		Monte Carlo	
	$\mathbb{P}(Y=0)$	$\mathbb{P}(Y=1)$	$\mathbb{P}(Y=0)$	$\mathbb{P}(Y=1)$
Beta(1,1)	0.5	0.5	0.496	0.504
Beta(1,25)	0.038	0.962	0.039	0.961



- Beta(1,1) prior gives 50-50 chance to have the disorder a priori.
- This may not represent prior beliefs for a rare disorder.
- Modify Beta prior parameters until prior probabilities are more realistic.
- Do not use data from the upcoming analysis to inform this decision!

Outline

1. Mixture priors, prior predictive checking.
2. Multivariate models
3. Monte Carlo methods
    - Monte Carlo integration, rejection sampling.

Multivariate models

All examples so far have been 1 parameter e.g., binomial, normal with known variance.

However, virtually every practical problem in statistics involves more than one unknown or unobservable quantity (parameter)

This is where Bayesian methods are much more straightforward than their classical counterpart.

Time permitting, the only other Bayesian theory we will discuss is Bayes factors and model selection. The rest of the course is devoted to the computational challenges in applying Bayesian updating.

Multivariate Models

We have d-dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$.

Specify a multivariate prior distribution, $\pi(\boldsymbol{\theta})$, and compute the posterior in the same manner:

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto \frac{L(\boldsymbol{x} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{x} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}.$$

If we are interested in only 1 or a few parameters, we can obtain the relevant marginal distribution from the joint posterior.

E.g. the marginal distribution of $\theta_1 \,|\, \boldsymbol{x}$ is obtained by integrating out the other components of $\boldsymbol{\theta}$

$$\pi(\theta_1 \,|\, \boldsymbol{x}) = \int_{\Theta_2} \ldots \int_{\Theta_d} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_d.$$

Can do this either a) Algebraically, or b) via Monte Carlo simulation

Q: Why are they called *marginal* distributions?

Multivariate Models

Bayesian inference for multivariate models requires overcoming the following challenges:

1. Prior specification:
    - Priors are now multivariate distributions.
    - Involve complex parameter dependencies.
    - More complicated to elicit.
2. Computation:
    - Required integrals are now more complicated.
    - Needs more sophisticated algorithms in absence of conjugates.
    - e.g. inversion sampling won't work.
3. Interpretation:
    - Structure of the posterior is more complicated.
    - Posterior inference and interpretation is more challenging.

Multivariate Models

Example:

- A machine-produced item is either:
  satisfactory ($x = 1$) or unsatisfactory ($x = 2$).
- The probability of the item being satisfactory depends on
  a) room temperature ($\theta_1 = 0$: cool, $\theta_1 = 1$: hot) and
  b) humidity ($\theta_2 = 0$: dry, $\theta_2 = 1$: humid).
- The probabilities of $x = 1$ under these conditions are given as:

$\mathbb{P}(x = 1 \mid \theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.6	0.8
$\theta_2 = 1$	0.7	0.6

Conditional probabilities of item being satisfactory

- The joint prior distribution of (θ_1, θ_2) is given as:

$\mathbb{P}(\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.3	0.2
$\theta_2 = 1$	0.2	0.3

Prior probabilities of room conditions

Multivariate Models

The joint posterior distribution can be calculated as:

		$\theta_1 = 0$	$\theta_1 = 1$	
$\mathbb{P}(x = 1 \mid \theta_1, \theta_2) \times \mathbb{P}(\theta_1, \theta_2)$	$\theta_2 = 0$	0.18	0.16	0.34
	$\theta_2 = 1$	0.14	0.18	0.32
		0.32	0.34	0.66

And normalising all probabilities gives:

		$\theta_1 = 0$	$\theta_1 = 1$	
$\mathbb{P}(\theta_1, \theta_2 \mid x = 1)$	$\theta_2 = 0$	18/66	16/66	34/66
	$\theta_2 = 1$	14/66	18/66	32/66
		32/66	34/66	1

From which we can obtain the marginal distributions:

$$\mathbb{P}(\theta_1 = 0 \mid x = 1) = 32/66, \quad \mathbb{P}(\theta_1 = 1 \mid x = 1) = 34/66$$

and

$$\mathbb{P}(\theta_2 = 0 \mid x = 1) = 34/66, \quad \mathbb{P}(\theta_2 = 1 \mid x = 1) = 32/66.$$

Multivariate Models

Example: Suppose that $Y_1 \sim \mathsf{Poi}(\alpha\beta)$ and $Y_2 \sim \mathsf{Poi}((1-\alpha)\beta)$ with Y_1 and Y_2 independent given α and β.

Now, suppose our prior information for α and β can be expressed as $\alpha \sim \mathsf{Beta}(p, q)$ and $\beta \sim \mathsf{Gam}(p+q, 1)$ with α and β independent, for specified hyperparameters p and q.

Then we have the following likelihood

$$\pi(y_1, y_2 \,|\, \alpha, \beta) = \frac{\exp(-\alpha\beta)(\alpha\beta)^{y_1}}{y_1!} \times \frac{\exp(-(1-\alpha)\beta)[(1-\alpha)\beta]^{y_2}}{y_2!}$$

and the joint prior is

$$\pi(\alpha, \beta) = \frac{\alpha^{p-1}(1-\alpha)^{q-1}}{\mathrm{B}(p, q)} \times \frac{\exp(-\beta)\beta^{p+q-1}}{\Gamma(p+q)}$$

Multivariate Models

Hence by Bayes' theorem:

$$\begin{aligned}\pi(\alpha, \beta \,|\, y_1, y_2) &\propto \exp(-\beta)\beta^{y_1+y_2}\alpha^{y_1}(1-\alpha)^{y_2}\alpha^{p-1}(1-\alpha)^{q-1}\exp(-\beta)\beta^{p+q-1} \\ &= \beta^{y_1+y_2+p+q-1}\exp(-2\beta)\alpha^{y_1+p-1}(1-\alpha)^{y_2+q-1}\end{aligned}$$

The marginal posterior distributions are given by

$$\pi(\alpha \,|\, y_1, y_2) = \int_0^\infty \pi(\alpha, \beta \,|\, y_1, y_2)\mathrm{d}\beta \propto \alpha^{y_1+p-1}(1-\alpha)^{y_2+q-1},$$

and

$$\pi(\beta \,|\, y_1, y_2) = \int_0^1 \pi(\alpha, \beta \,|\, y_1, y_2)\mathrm{d}\alpha \propto \beta^{y_1+y_2+p+q-1}\exp(-2\beta)$$

so

$$\alpha \,|\, y_1, y_2 \sim \mathsf{Beta}(y_1+p, y_2+q)$$

and

$$\beta \,|\, y_1, y_2 \sim \mathsf{Gam}(y_1+y_2+p+q, 2)$$

Multivariate Methods

Example:
Let $Y_1, \ldots Y_n$ be *iid* $\mathsf{N}(\mu, \sigma^2)$, mean μ and variance σ^2 are unknown (denote $\boldsymbol{y} := (y_1, \ldots, y_n)^\top$).

Take independent Jeffreys priors

$$\pi(\mu, \sigma^2) = \pi_J(\mu)\pi_J(\sigma) \propto 1/\sigma^2$$

This is a conventional improper prior for this model. Hence

$$\begin{aligned}
\pi(\mu, \sigma^2 \,|\, \boldsymbol{y}) &\propto \sigma^{-(n+2)} \exp\left(-\tfrac{1}{2\sigma^2} \textstyle\sum_{i=1}^n (y_i - \mu)^2\right) \\
&= \sigma^{-(n+2)} \exp\left(-\tfrac{1}{2\sigma^2} \left[\textstyle\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right]\right) \\
&= \sigma^{-(n+2)} \exp(-\tfrac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])
\end{aligned}$$

where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

is the sample variance of the y_i's.

Multivariate Methods

For the conditional distribution $\mu \,|\, \sigma^2, \boldsymbol{y}$ we have

$$\begin{aligned}
\pi(\mu \,|\, \sigma^2, \boldsymbol{y}) &\propto \sigma^{-(n+2)} \exp(-\tfrac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]) \\
&\propto \exp(-\tfrac{1}{2\sigma^2} n(\bar{y}-\mu)^2) \\
\mu \,|\, \sigma^2, \boldsymbol{y} &\sim \mathsf{N}(\bar{y}, \sigma^2/n).
\end{aligned}$$

For the marginal posterior distribution of σ^2

$$\begin{aligned}
\pi(\sigma^2 \,|\, \boldsymbol{y}) &\propto \int \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) \mathrm{d}\mu \\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \int \exp\left(-\frac{1}{2\sigma^2} n(\bar{y}-\mu)^2\right) \mathrm{d}\mu \\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{2\pi\sigma^2/n} \\
&\propto (\sigma^2)^{-(n+1)/2} \exp\left(-\tfrac{(n-1)s^2}{2\sigma^2}\right)
\end{aligned}$$

Scale-inverse χ^2 distribution:

$$\pi(x \,|\, \nu, \tau^2) = \frac{(\tau^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \frac{\exp\left(-\frac{\nu\tau^2}{2x}\right)}{x^{1+\nu/2}}$$

with $\nu > 0$, $\tau^2 > 0$.

$$\pi(\sigma^2 \,|\, \boldsymbol{y}) \quad \propto \quad (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)$$

which is a scaled inverse-χ^2 density

$$\sigma^2 \,|\, \boldsymbol{y} \sim \mathsf{Inv} - \chi^2(n-1, s^2).$$

Multivariate Methods

The marginal posterior of $\mu \,|\, \boldsymbol{y}$ is

$$\pi(\mu \,|\, \boldsymbol{y}) \quad \propto \quad \int \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) \mathrm{d}\sigma^2$$

Substitute $z = A/(2\sigma^2)$ where $A = (n-1)s^2 + n(\bar{y}-\mu)^2$.
$\mathrm{d}z = -\frac{A}{2\sigma^4}\mathrm{d}\sigma^2$, $\sigma^2 = A/(2z)$.
Hence

$$\begin{aligned}
\pi(\mu \,|\, y) \quad &\propto \quad \int \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) \mathrm{d}\sigma^2 \\
&= \quad \int (A/2z)^{-(n+2)/2} \exp(-z) \, \frac{2(A^2/(2z)^2)}{A} \mathrm{d}z \\
&\propto \quad A^{-n/2} \int z^{(n-2)/2} \exp(-z) \, \mathrm{d}z \\
&\propto \quad \left((n-1)s^2 + n(\bar{y}-\mu)^2\right)^{-n/2} \\
&\propto \quad \left(1 + \frac{n(\bar{y}-\mu)^2}{(n-1)s^2}\right)^{-n/2}
\end{aligned}$$

$$\pi(\mu \,|\, \boldsymbol{y}) \quad \propto \quad \left(1 + \frac{n(\bar{y} - \mu)^2}{(n-1)s^2}\right)^{-n/2}$$

t distribution:

$$\pi(x \,|\, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

with $\nu > 0$.

Hence

$$\frac{\mu - \bar{y}_n}{s/\sqrt{n}} \,\Big|\, \boldsymbol{y} \sim \mathsf{t}_{n-1}.$$

Multivariate Methods

Comment #1:
Sampling from the joint posterior distribution is easy, because of the closed form factorisation

$$\begin{aligned} \pi(\mu, \sigma^2 \,|\, \boldsymbol{y}) &= \pi(\mu \,|\, \sigma^2, \boldsymbol{y}) \pi(\sigma^2 \,|\, \boldsymbol{y}) \\ &\quad \text{Normal} \times \text{Inv-}\chi^2. \end{aligned}$$

Hence, first draw $\sigma^2 \,|\, \boldsymbol{y}$ (marginal) and then $\mu \,|\, \boldsymbol{y}, \sigma$ (conditional).

Comment #2:
A full conjugate analysis for this model is also possible. Specify

$$\begin{aligned} \mu \,|\, \sigma^2 &\sim \mathsf{N}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2). \end{aligned}$$

This gives the joint prior density

$$p(\mu, \sigma^2) \propto \sigma^{-1} (\sigma^2)^{-(\nu_0/2+1)} \exp(-\tfrac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2])$$

in the same form: the Normal-Inv-$\chi^2(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$ distribution.

Multivariate Methods

Example: Multivariate regression

Suppose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$.

Then $\boldsymbol{X}_i$ has a Multivariate Gaussian distribution $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$, if its probability density function is

$$\pi(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is a mean vector, and
Σ is a $d \times d$ positive semi-definite (covariance) matrix.

The precision matrix $\Lambda = \Sigma^{-1}$ is also positive semi-definite.

Univariate		Multivariate	
τ	Conjugate = Gamma	Λ	Conjugate = Wishart
σ^2	Conjugate = Inv-Gamma	Σ	Conjugate = Inv-Wishart

Multivariate Methods

Wishart distribution

The density function of a $\mathsf{Wishart}(\nu, \mathbf{V})$ is

$$\pi(\mathbf{X} \mid \nu, \mathbf{V}) = c^{-1} |\mathbf{X}|^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{X})\right), \ \mathbf{X} \succ \mathbf{0}\,.$$

- Degrees of freedom $\nu \in (d, \infty)$ (real)
- Scale parameter $\mathbf{V} \in \mathbb{R}^{d \times d}$ and $\mathbf{V} \succ \mathbf{0}$ (a positive semi-definite matrix).
- c is a normalizing constant (for $d = 1$ this is the Gamma distr.).

Inverse Wishart distribution

The density function of a $\mathsf{InvWish}(\nu, \mathbf{V})$

$$\pi(\mathbf{X} \mid \nu, \mathbf{V}) \propto |\mathbf{X}|^{-(\nu+d+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{V}\mathbf{X}^{-1})\right), \ \mathbf{X} \succ \mathbf{0}\,.$$

- If $\mathbf{X} \sim \mathsf{Wishart}(\nu, \mathbf{V})$, then $\mathbf{X}^{-1} \sim \mathsf{InvWish}(\nu, \mathbf{V})$.

Example: Multivariate regression

The conjugate prior for $\boldsymbol{\mu}$ when Σ is known, is $\mathsf{N}(\boldsymbol{\mu}_0, \Sigma_0)$.

The resulting posterior for $\boldsymbol{\mu} \,|\, \boldsymbol{x}$ is $\mathsf{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$, where

$$\hat{\boldsymbol{\mu}} = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{x}}_n)$$

and

$$\hat{\Sigma} = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}.$$

Clear analogy with univariate result.

Example: Multivariate regression

The conjugate prior for Σ when $\boldsymbol{\mu}$ is known, is $\mathsf{InvWish}(\nu, \mathbf{A})$.

The resulting posterior for Σ is $\mathsf{InvWish}(\hat{\nu}, \hat{\mathbf{A}})$, where

$$\hat{\nu} = \nu + n$$

and

$$\hat{\mathbf{A}} = \mathbf{A} + \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top.$$

Again, a clear analogy with univariate result.

Outline

1. Mixture priors, prior predictive checking.
2. Multivariate models
3. Monte Carlo methods
    - Monte Carlo integration, rejection sampling.

Monte Carlo Integration

When working with posterior distributions,
almost all inference can be expressed as an integral.

This is most obvious for expectations:

$$\mathbb{E}_f[X] = \int x f(x) \mathrm{d}x.$$

So given a sample $X_1, \ldots, X_N \sim f(x)$, we can approximate the theoretical mean by the sample mean to give

$$\mathbb{E}_f[X] \approx \frac{1}{N} \sum_{i=1}^{N} X_i.$$

Monte Carlo Integration

Generalising this argument (Law of Large Numbers):

Suppose we wish to calculate

$$\mathbb{E}[\phi(X)] = \int \phi(x) f(x) \mathrm{d}x$$

with respect to f.

Then given a sample $X_1, \ldots, X_N \sim f(x)$, we can approximate the expectation via

$$\mathbb{E}[\phi(X)] \approx \frac{1}{N} \sum_{i=1}^{N} \phi(X_i)$$

which is an unbiased estimate.

Monte Carlo Integration

Example: Approximate

$$\int_0^1 \exp(x^2)\mathrm{d}x.$$

Solution:
The integral can be expressed as $\mathbb{E}[\exp(U^2)]$ where $U \sim \mathsf{U}[0,1]$.

Hence:

```
> u=runif(5000)
> mean(exp(u^2))
[1] 1.466655
```

Monte Carlo Integration

Now consider

$$\int_a^b g(x)\mathrm{d}x$$

for finite a and $b > a$.

Create a transformation to map integration limits to $[0, 1]$:

Here $u = (x - a)/(b - a)$ and so $x = a + u(b - a)$. Hence

$$\int_a^b g(x)\mathrm{d}x = \int_0^1 (b - a)g(a + u(b - a))\mathrm{d}u$$

which is in the same form as before.

Note:

- Could alternatively simulate $X^{(1)}, \ldots, X^{(N)} \sim \mathsf{U}(a, b)$ and then compute $(b - a)\frac{1}{n}\sum_{i=1}^{N} g(X^{(i)})$.
- However, given samples from $\mathsf{U}(0, 1)$ are transformed to produce $\mathsf{U}(a, b)$, this is identical to the above.
- All Monte Carlo integration is based on samples from $\mathsf{U}(0, 1)$.

Monte Carlo Integration

Example: Calculate

$$\int_0^2 \exp(x^2)\mathrm{d}x.$$

Solution:

Make the transformation $u = x/2$:

$$\int_0^2 \exp(x^2)\mathrm{d}x = 2\int_0^1 \exp(4u^2)\mathrm{d}u.$$

Hence:

```
> u=runif(5000)
> mean(2*exp(4*u^2))
[1] 17.01787
```

```
> u=runif(5000,0,2)
> mean(2*exp(u^2))
[1] 16.65154
```

Monte Carlo Integration

Now suppose the range of integration is infinite:

$$\int_0^\infty g(x)\mathrm{d}x.$$

Make the transformation $u = 1/(x+1)$ so that $x = 1/u - 1$. Hence

$$\begin{aligned}\int_0^\infty g(x)\mathrm{d}x &= \int_1^0 -\frac{1}{u^2} g\left(\frac{1}{u} - 1\right)\mathrm{d}u \\ &= \int_0^1 \frac{1}{u^2} g\left(\frac{1}{u} - 1\right)\mathrm{d}u\end{aligned}$$

(as $\int_0^1 f(x)\mathrm{d}x = F(1) - F(0) = -(F(0) - F(1)) = -\int_1^0 f(x)\mathrm{d}x$.)

Can now estimate as before.

Monte Carlo Integration

Example: Calculate

$$\int_0^\infty \exp(-x^2)\mathrm{d}x.$$

Solution:
Make the transformation $u = 1/(x+1)$:

$$\int_0^\infty \exp(-x^2)\mathrm{d}x = \int_0^1 \frac{1}{u^2}\exp\left[-\left(\frac{1}{u}-1\right)^2\right]\mathrm{d}u.$$

Hence:

```
> u=runif(5000)
> z=1/u-1
> mean(1/u^2*exp(z^2))
[1] 0.8893568
```

Note: Exact answer $= \sqrt{\pi}/2 = 0.8862269$.

Monte Carlo Integration

Monte Carlo methods are most useful for approximating multidimensional integrals. E.g.

$$\int_0^1 \int_0^1 g(u_1, u_2) du_1 du_2.$$

Let U_1, $U_2 \sim \mathsf{U}(0,1)$. Their joint density is

$$f(u_1, u_2) = \begin{cases} 1 & \text{if } u_1 \in [0,1], u_2 \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

And so

$$\begin{aligned} \mathbb{E}_f[g(U_1, U_2)] &= \int_0^1 \int_0^1 g(u_1, u_2) f(u_1, u_2) \mathrm{d}u_1 \mathrm{d}u_2 \\ &= \int_0^1 \int_0^1 g(u_1, u_2) \mathrm{d}u_1 \mathrm{d}u_2. \end{aligned}$$

Monte Carlo Integration

Hence, given $(U_1^{(1)}, U_2^{(1)}), \ldots, (U_1^{(N)}, U_2^{(N)})$, with $U_1^{(i)}, U_2^{(i)} \sim \mathsf{U}(0,1)$, we have

$$\int_0^1 \int_0^1 g(u_1, u_2) \mathrm{d}u_1 \mathrm{d}u_2 \approx \frac{1}{N} \sum_{i=1}^{N} g(U_1^{(i)}, U_2^{(i)}),$$

with the approximation being exact as $N \to \infty$ under weak conditions.

Final comment

- This approach is remarkably easy to use, even in high dimensions.
- The price for this simplicity is that the variance is high.
- Techniques are available to improve the precision of this.
  e.g. Importance Sampling.
- We look at importance sampling next week.

Outline

1. Mixture priors, prior predictive checking.
2. Multivariate models
3. Monte Carlo methods
    - Monte Carlo integration, rejection sampling.

Rejection Sampling

We want to sample from a distribution $f(x)$, but this is difficult to do.

Algorithm (Rejection Sampling from $f(x)$)

Require: Proposal pdf g, which is easy to simulate from

1: Find constant K such that $f(x) \leq Kg(x)$ for all x
2: **while** Failing to accept **do**
3: Simulate $X \sim g(x)$
4: Accept X with probability $f(X)/(Kg(X))$.
5: **return** X

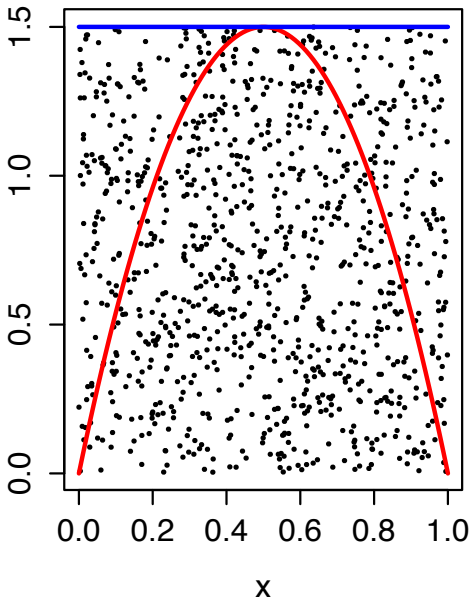
Example:

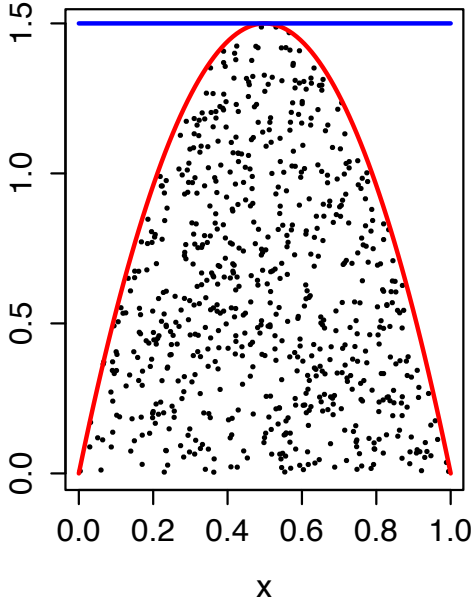
Obtain N samples from $\mathsf{Beta}(2, 2)$.

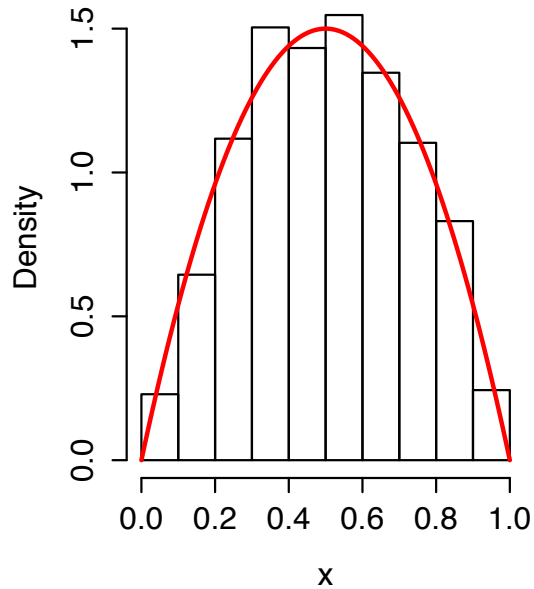
Solution:

Inversion sampling is inefficient as requires numerical inversion of $F(x)$.
Instead, try rejection sampling.

Rejection Sampling







red = f(x), blue = Kg(x)

- Bound the density $f(x)$ by a rectangle ($Kg(x)$)
- Simulate points (X_i, Y_i) uniformly over the rectangle
- Reject those points which do not lie under $f(x)$
- X_i co-ordinates of accepted points are a sample from $f(x)$.

Rejection Sampling

Comments:

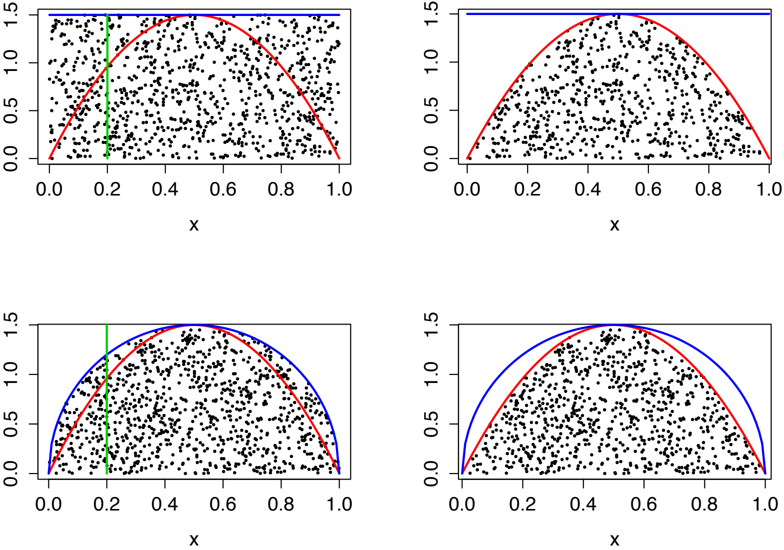
1. Y_i plays role of accepting X_i with probability $f(x)/(Kg(x))$ (green line on next slide)

2. Efficiency of algorithm depends on number of rejections. Bounding rectangle may be inefficient

3. To improve efficiency (and to allow for unbounded $f(x)$) modify bounding function to take any form $Kg(x)$, where
   - $g(x)$ is any density from which it is easy to simulate
   - $K > 0$ is some constant to be chosen

4. Algorithm acceptance probability is $= \frac{f(x)}{Kg(x)}$ (see previous pic)

Rejection Sampling



red = f(x), blue = Kg(x)

Top: g(x)=Beta(1,1), efficiency = 681/1000 = 0.681

Bottom: g(x)=Beta(1.5,1.5), efficiency = 848/1000 = 0.848

Rejection Sampling

Algorithm (Rejection Sampling)

Require: Proposal pdf g, enveloping constant K
  repeat
    Simulate $X \sim g(x)$
    Simulate $U \sim \mathsf{U}(0,1)$, independently.
  until $U < f(X)/(Kg(X))$
  return X

Conditional on X, the indicator of the statement $U < f(X)/(Kg(X))$ is a Bernoulli variable with success probability $f(X)/(Kg(X))$. Thus, the uniformly distributed U variable is only needed to ensure that we accept X with probability $f(X)/(Kg(X))$.

How to choose K for maximum efficiency?
We require $f(x) \leq Kg(x)$ (compulsory) $\forall x$.

$K = \max_x f(x)/g(x) \qquad \Rightarrow \qquad$ max acceptance probability $= 1$.

Rejection Sampling

Why does this work?

Let $X \sim g(x)$ such that $f(x) \leq Kg(x), \forall x$ for some K.

$f(x)$ is not required to be normalised! (i.e. $\int f(y)\mathrm{d}y = c < \infty$)

Let $h(x) = f(x)/Kg(x)$ be the probability that x is accepted.

Then

$$\mathbb{P}(X \leq x \text{ and } X \text{ accepted}) = \int_{-\infty}^{x} h(y)g(y)\mathrm{d}y$$

and

$$\mathbb{P}(X \text{ accepted}) = \int_{-\infty}^{\infty} h(y)g(y)\mathrm{d}y$$

so

$$\begin{aligned} \mathbb{P}(X \leq x \,|\, X \text{ accepted}) &= \frac{\int_{-\infty}^{x} h(y)g(y)\mathrm{d}y}{\int_{-\infty}^{\infty} h(y)g(y)\mathrm{d}y} \\ &= \frac{\int_{-\infty}^{x} f(y)\mathrm{d}y}{\int_{-\infty}^{\infty} f(y)\mathrm{d}y} \end{aligned}$$

so that the accepted values have pdf f.

Rejection Sampling

Comments:

- Overall acceptance rate:

$$\mathbb{P}(X \text{ accepted}) = \int h(y)g(y)\mathrm{d}y = \frac{1}{K}\int f(y)\mathrm{d}y \propto \frac{1}{K}.$$

  Hence want K as small as possible subject to $f(x) \leq Kg(x)$.
  i.e. $K = \max_x f(x)/g(x)$

- The normalising constant in the denominator means that f needs only be known up to proportionality in order for algorithm to work.

- Algorithm efficiency depends on the difference between f and the bounding envelope Kg. (Recall Beta(2,2) example).

  $\Rightarrow$ If $g(x)$ gets closer to $f(x)$ then K can be smaller via
  $K = \max_x f(x)/g(x)$

  Smaller K = greater efficiency.

Rejection Sampling

Example:
Simulate from the density

$$f(x) \propto \begin{cases} x(1-x)^3 & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Use rejection sampling with

$$g(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(that is, use the uniform density on $[0, 1]$).

Rejection Sampling

To find the optimal K we want

$$K = \max_x \frac{f(x)}{g(x)} = \max_x f(x).$$

The derivative of $f(x)$ is

$$\begin{aligned} f'(x) &= x(-3(1-x)^2) + (1-x)^3 \\ &= (1-x)^2(1-4x) \end{aligned}$$

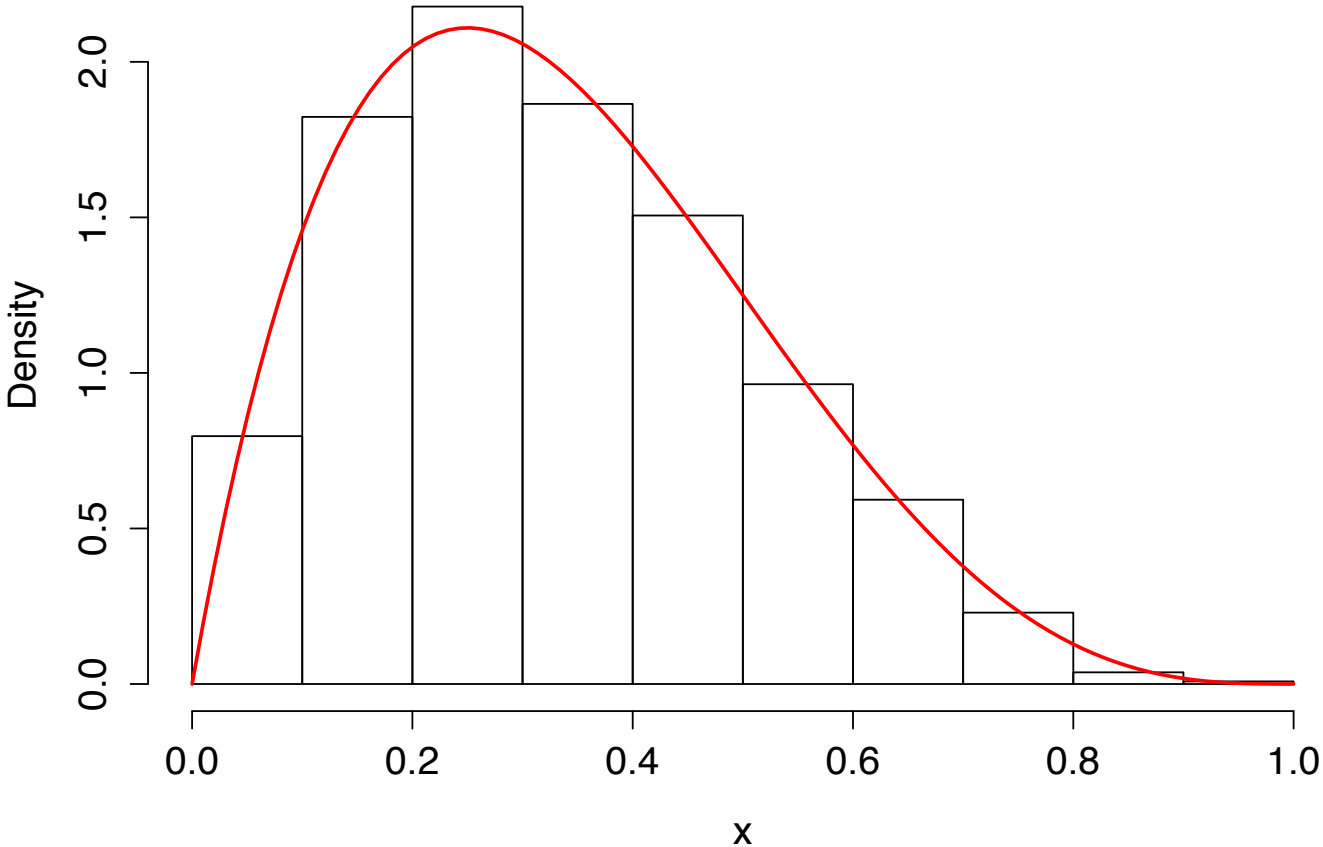
Hence $f'(x) = 0$ when $x = 1$ or $x = 1/4$.

Now $f(0) = f(1) = 0$ (local minimum) and
$f(1/4) = (1/4).(3/4)^3 = 27/256 = K$ (local max as $f''(x) < 0$).

Hence

$$K = \max_x \frac{f(x)}{g(x)} = \frac{27}{256}.$$

Rejection Sampling




```
L=5000
K=27/256
xstar=runif(L)
ind=(runif(L)<(xstar*(1-xstar)^3/K))
hist(xstar[ind],probability=T,xlab="x",ylab="Density",main="")
xx=seq(0,1,length=100)

lines(xx,20*xx*(1-xx)^3,lwd=2,col=2)
```